Herramientas y Estrategias para Generación de Contranarrativas usando PLN

GOBIERNO DE ESPAÑA
MINISTERIO DE INCLUSIÓN, SEGURIDAD SOCIAL Y MIGRACIONES
SECRETARIA DE ESTADO DE MIGRACIONES
DIRECCIÓN GENERAL DE ATENCIÓN HUMANITARIA E INCLUSIÓN SOCIAL DE LA INMIGRACIÓN
U23
oberaxe Observatorio Español del Racismo y la Xenofobia
REAL UP COMBATING HATE SPEECH
Cofinanciado por la Unión Europea

Herramientas y Estrategias para Generación de Contranarrativas usando PLN

GOBIERNO DE ESPAÑA
MINISTERIO DE INCLUSIÓN, SEGURIDAD SOCIAL Y MIGRACIONES
SECRETARIA DE ESTADO DE MIGRACIONES
DIRECCIÓN GENERAL DE ATENCIÓN HUMANITARIA E INCLUSIÓN SOCIAL DE LA INMIGRACIÓN
U23
oberaxe Observatorio Español del Racismo y la Xenofobia
REAL UP COMBATING HATE SPEECH
Cofinanciado por la Unión Europea

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

Autora: María Teresa Martín Valdivia

Edita y distribuye: Observatorio Español del Racismo y la Xenofobia
Calle María de Guzmán, 52, tercera planta. 28003 Madrid
oberaxe@inclusion.gob.es
https://www.inclusion.gob.es/oberaxe/es/index.htm

NIPO 121-23-040-8

Diseño: Solana e Hijos, A.G., S.A.U.

Maquetación: CYAN, Proyectos Editoriales, S.A

TABLA DE CONTENIDO

1 Introducción

El presente informe tiene como objetivo mostrar un resumen de las estrategias que se siguen para la generación de contranarrativas, prestando especial atención a los sistemas automáticos basados en aprendizaje automático (ML) y Procesamiento de Lenguaje Natural (PLN).

El conocido aumento de las interacciones sociales a través de las plataformas digitales, ha provocado la presencia de comportamientos inadecuados en la web, como es la propagación del discurso de odio entre los usuarios de las plataformas. La libertad de expresión en estos medios ha expuesto a sus usuarios a publicaciones que en ocasiones se utilizan para denigrar, insultar o herir con un lenguaje, suave o grosero, en función del género, la raza, la religión, la ideología u otras características personales. Desafortunadamente, este tipo de comunicación puede ser muy perjudicial llegando a provocar efectos psicológicos negativos entre los usuarios, especialmente entre los jóvenes, en forma de ansiedad, sentimiento de ciberacoso e incluso suicidio en los casos más extremos.

Este problema implica principalmente a gobiernos y plataformas online, que requieren adoptar medidas en forma de leyes y políticas que contribuyan a fomentar una sana convivencia en estos medios. Por ejemplo, desde 2013, el Consejo Europeo ha promovido el movimiento "No Hate Speech"[1] (No al discurso del odio) con el objetivo de movilizar a los jóvenes para combatir el discurso de odio y promover los Derechos Humanos en Internet. En mayo de 2016, la Comisión Europea llegó a un acuerdo con Facebook, Microsoft, Twitter y YouTube por el que firmaron un "Código de conducta sobre la lucha contra el discurso de odio ilegal en línea"[2]. Basado en este Código, en España se ha redactado el "Protocolo para combatir el discurso de odio ilegal en línea"[3]. Entre 2018 y 2020, otras plataformas como Instagram, Snapchat, Dailymotion y TikTok se unieron al Código de Conducta de la Comisión Europea.

Según el informe del Ministerio del Interior de 2019[4] sobre la evolución de los delitos de odio en España, las amenazas, los insultos y la discriminación se cuentan como los actos delictivos más repetidos, siendo Internet (54,9%) y las redes sociales (17,2%) los medios más utilizados para cometer estas acciones.

1 https://www.coe.int/en/web/committee-on-combatting-hate-speech/home
2 https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en
3 https://www.inclusion.gob.es/oberaxe/ficheros/ejes/discursoodio/PROTOCOLO_DISCURSO_ODIO.pdf
4 http://www.interior.gob.es/documents/642012/3479677/informe+evolucion+2019/631ce020-f9d0-4feb-901c-c3ee0a777896

Parece claro que el problema de detección de discurso de odio se ha agravado en los últimos años y que se hace necesario el estudio, análisis e implementación de soluciones en todos los ámbitos, incluido el área de las tecnologías del lenguaje. Analizar este tipo de contenido nocivo en la web requiere de sistemas automáticos que sean capaces de procesar y analizar el lenguaje humano. Por esta razón, la detección y el análisis del discurso de odio se ha convertido en una de las principales áreas de investigación del Procesamiento del Lenguaje Natural (PLN). El PLN es un área importante de la Inteligencia Artificial que trata de comprender y generar el lenguaje de la misma manera que lo hacemos los humanos mediante métodos computacionales. Además, el uso de algoritmos de aprendizaje automático (Machine Learning – ML) está permitiendo desarrollar sistemas de clasificación que, combinados con técnicas avanzadas de PLN, ayudan a dar respuesta a muchos problemas sociales actuales, incluida la detección de lenguaje ofensivo y discurso de odio en medios sociales.

De hecho, la ingente cantidad de datos de redes sociales que se comparten en línea a diario requiere que la mitigación del odio se aborde mediante herramientas fiables, eficientes y escalables, capaces de generar discursos para combatir este problema. Recientemente, se han realizado esfuerzos para recopilar conjuntos de datos de lucha contra el odio y automatizar la producción de contra-discursos. Sin embargo, este campo de investigación está aún muy poco estudiado, y siguen abiertas muchas preguntas sobre los enfoques y métodos más eficaces, y sobre la forma de evaluarlos.

De este modo, es importante resaltar que para combatir el discurso de odio no solo se debe prestar atención a la detección del mismo sino también a las estrategias que permitan paliar sus consecuencias. Una de las estrategias que se han usado tradicionalmente incluye la eliminación de información que se considera inapropiada o nociva (Roberts, 2016). Así, por ejemplo, las plataformas como Facebook eliminan millones de publicaciones dañinas cada año, con la ayuda de herramientas de Inteligencia Artificial (IA). Aunque la eliminación de estos contenidos puede reducir inmediatamente la cantidad de mensajes dañinos, puede acarrear acusaciones de censura y puede no ser eficaz para frenar el odio a largo plazo (Schieb and Preuss, 2016).

Un enfoque alternativo es responder con un discurso alternativo, es decir, con respuestas dirigidas a refutar el lenguaje de odio utilizando razones reflexivas y convincentes, y argumentos basados en hechos. Esto ha demostrado ser eficaz para influir en el comportamiento tanto de los agresores como de los espectadores que presencian las interacciones, así como para proporcionar apoyo a las víctimas (Benesh, 2014). Esto es lo que se conoce como **contranarrativa** (counter narrative)[5] y trata de una respuesta educada y no agresiva a la incitación al odio, que se oriente a contrarrestar las declaraciones extremas. Una contranarrativa (CN) es una respuesta directa a un discurso de odio y se considera una herramienta eficaz porque:

i. preserva el derecho a la libertad de expresión,
ii. rectifica los estereotipos y la información engañosa con pruebas creíbles,

5 Otros términos utilizados son contrarrelatos, contraargumentación o contradiscurso

iii. puede alterar los puntos de vista de los perpetradores de odio y de los espectadores,
iv. y fomenta el intercambio de opiniones entre los posibles espectadores para cambiar sus perspectivas.

Resumiendo, es importante combatir el discurso de odio no solo con la detección del mismo sino a través de estrategias que permitan paliar sus consecuencias. La contranarrativa es una de estas estrategias y consiste en proporcionar una narrativa alternativa a la que promueve el odio y la violencia. La contranarrativa puede incluir mensajes que fomenten la tolerancia, el respeto y la inclusión, y puede ser una herramienta poderosa para desafiar y desmantelar el discurso de odio (Mathew et al., 2019). Es por ello, que algunas Organizaciones No Gubernamentales (ONGs) y administraciones públicas dedican esfuerzos a formar a operadores para que vigilen las plataformas de medios sociales y generen manualmente contranarrativas eficaces cuando sea necesario. Sin embargo, la lucha contra el odio se enfrenta a una serie de retos que no son fáciles de resolver. En primer lugar, dada la enorme cantidad de discurso de odio que se produce a diario en dichos medios sociales, la intervención manual a través de contraargumentos no es escalable y en la mayoría de las ocasiones se hace inviable. Por otra parte, para componer una contranarrativa, un operador experto suele leer los mensajes de odio, buscar y filtrar información relacionada en Internet y, a continuación, componer una contranarrativa. Todo este proceso puede durar varios minutos, además del daño psicológico y mental que supone tener a personas expuestas a este tipo de mensajes durante largos períodos de tiempo.

Se trata de un problema complejo que recientemente ha captado la atención de investigadores en el ámbito del PLN ya que se han empezado a estudiar diferentes técnicas computacionales basadas en los últimos avances de PLN y ML, orientadas a asistir a estos operadores de ONGs para ayudarles en la tarea de generar automáticamente las contranarrativas disminuyendo así el tiempo y el esfuerzo en la lucha contra el odio en línea. Sin embargo, el problema se enfrenta a muchos retos y dificultades que trataremos a continuación.

En este informe se realiza un recorrido por los distintos métodos que se han utilizado para tratar la contranarrativa y se presenta la situación actual tanto a nivel internacional como nacional. Se muestra también el uso de grandes modelos lingüísticos (LLM- Large Language Model) para esta tarea.

2 Trabajos previos

Aunque hay una gran cantidad de bibliografía relacionada con la detección de discurso de odio y el lenguaje ofensivo, no solo en inglés sino también en otros idiomas incluido el español, en el caso de la generación de contranarrativa nos encontramos con escasos trabajos y la mayoría centrados en el idioma inglés.

Se puede encontrar un resumen del estado de la cuestión sobre la detección y tratamiento de discurso de odio en los trabajos de Poletto et al. (2020) y Jahan and Oussalah (2023). Si nos interesan idiomas diferentes al inglés, en el trabajo de Plaza-del-Arco et al. (2021) se revisan diferentes modelos pre-entrenados del lenguaje para detectar discurso de odio en español.

Centrándonos en el tema de la generación de contranarrativa hay que destacar que la investigación es muy reciente y no hay muchos trabajos aún que aborden el problema, y los que lo han hecho se enfocan principalmente en la generación de contranarrativas en inglés, encontrando apenas algunos en otros idiomas.

Para tener una visión general del problema se puede hacer mención a dos trabajos que realizan una recopilación de las investigaciones centradas en la materia. El primero de ellos es la tesis dirigida por el profesor Marco Guerini de la Fondazione Bruno Kessler (en Trento, Italia) y defendida por Yi Ling Chung (Chung, 2022) donde se realiza un estudio en profundidad de la cuestión y se recogen muchas de las ideas y modelos que se han utilizado hasta la fecha. En segundo lugar, la recopilación de trabajos presentada en por Alsagheer y otros autores (2022) incluye más de 60 publicaciones relacionadas con la contranarrativa en redes sociales.

Uno de los primeros trabajos que postula los beneficios de utilizar contranarrativas para paliar los efectos del discurso de odio lo podemos encontrar en Benesh (2014). En este trabajo se pueden encontrar varias estrategias opuestas a la eliminación y censura de discurso de odio, además de definiciones claras en la materia.

Desde el punto de vista del PLN, existen varios trabajos que estudian la posibilidad de generar automáticamente contranarrativas o de utilizar estrategias que combinan la intervención humana para contrarrestar el odio y los discursos dañinos en Internet.

Qian et al. (2019) fueron de los primeros en intentar la generación automática de contranarrativas (CN: Counter Narratives). Crearon un recurso de 10.243 contranarrativas para responder a 5.257 instancias de discurso de odio (HS: Hate Speech) extraídas de 5.020 conversaciones que contenían 22.324 comentarios

de Reddit y 31.487 contranarrativas para responder a 14.614 instancias de discurso de odio en 11.825 conversaciones que contenían 33.776 publicaciones de Gab. Utilizaron el *crowdsourcing*[6] para obtener contranarrativas y lo emplearon para entrenar modelos neuronales.

En Chung et al. (2019) se describe cómo se ha generado uno de los primeros corpus de contranarrativa denominado CONAN (COunter NArratives through Nichesourcing). Este corpus incluye 6.645 pares de discurso de odio-contranarrativa (HS-CN) en inglés, incluidos 2.781 pares traducidos del francés y del italiano. Es el primer trabajo en el que se abordan idiomas diferentes al inglés. En principio CONAN se centra en 3 comunidades (targets) concretas (musulmanes, judíos, LGTBI) si bien este recurso ha servido de base para generar otros recursos con un mayor número de objetivos (*multitarget*).

Mathew y otros autores (2019) analizan los resultados en varios puntos interesantes, como que los comentarios que son contranarrativas reciben el doble de *likes* que los comentarios no contranarrativas. Para ciertas comunidades, la mayoría de los comentarios que no son contranarrativas tienden a ser de incitación al odio, los diferentes tipos de contranarrativas no son todos igual de eficaces y la elección del lenguaje de los usuarios que publican contranarrativas es muy diferente de la de los que publican comentarios que no son contranarrativas, como revela un detallado análisis psicolingüístico. Lo interesante de su estudio es el uso de modelos de aprendizaje automático para detectar las contranarrativas en los vídeos de YouTube, con una puntuación F1 de 0,73. Además, proporcionan un conjunto de datos para la detección de contranarrativas utilizando comentarios de YouTube para realizar un estudio de medición que caracteriza la estructura lingüística de la contranarrativa. En general, su estudio ofrece información valiosa sobre el impacto de la contranarrativa en las interacciones en línea, la diferencia en el uso del lenguaje entre comentarios con y sin contranarrativa, y provee recursos importantes para investigaciones futuras en este campo.

En el trabajo de Tekiroglu et al. (2020) se describen varias estrategias para la generación de contranarrativas proponiendo el uso por primera vez de herramientas basadas en PLN, concretamente, en este caso se hace uso de GPT-2 y si bien, los resultados directos del sistema computacional no son excelentes, sí que sirven como base para ser revisados por diferentes humanos no expertos en la materia, que hacen un primer filtrado y, por último, expertos de ONGs validan el resultado final.

Tras este trabajo, en 2021 Chung y otros autores (2021a) proponen una arquitectura más elaborada basada primero en la generación de consultas automáticas y extracción de frases desde una base de conocimiento. Y, en segundo lugar, genera contranarrativas en base a las frases extraídas. Se toma como base el corpus CONAN pero se enriquece este recurso mediante una serie de frases obtenidas mediante un módulo de extracción, generación y selección de conocimiento, creando el corpus CONAN-KN. En este trabajo se exploran varios modelos para la experimentación que incluyen GPT-2, una versión mejorada de GPT y XNLG.

6 El término *crowdsourcing* hace referencia al acto de recopilar servicios, ideas o contenido a través de las contribuciones de un gran grupo de personas.

La misma metodología de recopilación de datos sobre discursos de odio dirigidos a otras religiones, razas y por razón de género es la que se utiliza para realizar un ajuste fino de GPT2 para la generación automática de contranarrativas (Fanton et al., 2021). Se genera un corpus en inglés a partir de CONAN denominado CONAN Multitarget en el que operadores expertos de ONGs revisan las contranarrativas generadas por el sistema computacional y realizan una postedición de las mismas. Este trabajo es especialmente interesante porque produce un recurso de alta calidad que está generado por modelos automáticos pero revisados manualmente por operadores humanos y porque incluye 8 clases diferentes de objetivos hacia los que van dirigidos los discursos de odio (ver tabla 1).

Tabla 1. Distribución de pares "Discurso de odio-Contranarrativa" por objetivo en el corpus CONAN-Multitarget

***Target* / Objetivo del discurso de odio**	**Nº de pares "Discurso de odio-Contranarrativa"**
Personas con discapacidad	220
Judíos	594
LGTBI	617
Inmigrantes	957
Musulmanes	1.335
Personas de color	352
Mujeres	662
Otros (Personas con sobrepeso, gitanos…)	266
Total	**5.003**

Todos estos trabajos previos han servido para que en el trabajo de Chung et al. (2021b) se desarrolle una herramienta de ayuda a las ONGs para proponer contranarrativas que permitan combatir el discurso de odio mediante este sistema automático de apoyo a la decisión.

Por otra parte, Zhu y Bhat (2021) propusieron un procedimiento para generar contranarrativas candidatas utilizando un modelo generativo basado en una red neuronal recurrente (RNN: Recurrent Neural Network) entrenada en este conjunto de datos y que realizaba una selección de la candidata más relevante.

En el trabajo de Bonaldi et al. (2022) se aborda un aspecto muy interesante, ya que estudia la generación de la contranarrativa desde una perspectiva de diálogo en contraposición a un simple par de discurso de odio-contranarrativa. Se presenta un enfoque híbrido para la recopilación de datos a través de diálogos, que combina la intervención de anotadores humanos expertos con diálogos generados por máquinas obtenidos mediante distintas configuraciones. Se genera además un corpus que se pone a libre disposición denominado DIALOCONAN (DIALOgicalCOunter-NArrativescollectioN) y que incluye un conjunto de datos

con más de 3.000 diálogos multiturno ficticios entre un *hater* y un operador de una ONG[7]. Además, se abarcan 6 objetivos de odio constituyendo así un nuevo recurso para combatir el discurso de odio hacia distintos grupos objetivo (ver tabla 2).

Tabla 2. Distribución de pares "Discurso de odio-Contranarrativa" por objetivo en el corpus DIALOCONAN

Target/ Grupo objetivo	Nº de pares
Judíos	468
LGTBI	591
Inmigrantes	534
Musulmanes	505
Personas africanas o afrodescendientes	493
Mujeres	462
Otros (Personas con sobrepeso, gitanos…)	6
Total	**3.059**

Por último, Ashida y Komachi (2022) estudian cómo los modelos preentrenados se pueden utilizar para generar automáticamente mensajes que contrarresten los textos ofensivos en redes sociales. Utilizan los modelos GPT2, GPT2-Neo y GPT3 para generar las contranarrativas que posteriormente son evaluadas manualmente a través de Amazon Mechanical Turk, mostrando que GPT-3 es el modelo que produce los mensajes de mayor calidad. El corpus generado (CHASM: Countering HAte Speech and Microaggressions) está disponible en la web https://github.com/tmu-nlp/CHASM.

7 https://github.com/marcoguerini/CONAN.

3 Trabajos en otros idiomas

Como ya se ha comentado, los trabajos para la generación de contranarrativas no son muy abundantes pero si, además, nos centramos en idiomas diferentes al inglés, la bibliografía se reduce mucho más siendo casi testimonial. La primera investigación que abre el foco a otros idiomas la encontramos en Bartlett y Krasodomski-Jones (2015) que estudia el efecto de la contranarrativa en Facebook. Los autores reivindican la importancia de que se mantenga el principio de libertad de expresión en Internet y de que sea un lugar donde la gente sienta que puede decir lo que piensa abierta y libremente, incluso cuando se trata de contenidos extremos o radicales. Por ello, se centran en el efecto que la contranarrativa produce en contraposición a la eliminación de contenidos. El estudio se lleva a cabo en distintos países (Francia, Reino Unido, Marruecos y Túnez, Indonesia e India) llegando a la conclusión de que, según la región, el contexto en el que se produce y comparte la contranarrativa cambia y el contenido de la contranarrativa funciona de forma diferente.

Garland et al. (2020) también estudian cómo reaccionan las personas de distintos países ante los contenidos de contranarrativa en Facebook y tratan de identificar qué tipos de contenidos tienen más probabilidades de atraer a los usuarios. El estudio muestra que los usuarios se comprometen con la contranarrativa dependiendo de su ubicación en el mundo, lo que indica que no existe un enfoque amplio que cubra la totalidad de Facebook. Se requieren enfoques específicos para los distintos lugares y países en los que Facebook proporciona una plataforma importante para difundir mensajes que se enfrenten a las narrativas de odio e incitación a la violencia.

Chung et al. (2020) se centran en la generación del primer corpus de contranarrativas en italiano mediante la traducción automática de contranarrativas en inglés a partir del corpus CONAN.

Miškolci et al. (2020) exploran el discurso de odio contra la comunidad gitana de Eslovaquia en Facebook entre abril de 2016 y enero de 2017. En total estudian 60 debates en Facebook con más de 7.500 comentarios sobre temas relacionados con la comunidad gitana, publicados por los perfiles de varios miembros del Parlamento eslovaco y los medios de comunicación en línea más populares.

Por último, Garland et al. (2022) analizaron la eficacia de la contranarrativa utilizando diferentes medidas a nivel macro y micro para analizar 180.000 conversaciones políticas en alemán que tuvieron lugar en Twitter durante cuatro años. Los resultados sugieren que el discurso de odio organizado está asociado a cambios en el discurso público y que la contranarrativa, especialmente cuando está organizada, puede tener un impacto positivo en la retórica de odio en línea.

4 Trabajos en español

Respecto a trabajos en español, hasta 2023 sólo encontramos un intento de abordar el problema en el artículo de Furman et. al. (2022). Si bien esta investigación se centra más en la minería de argumentación y el corpus sobre el que se trabaja para español, prácticamente no es representativo. Este trabajo realiza un enriquecimiento del corpus HatEval que incluye tuits con discurso de odio (Basile et al., 2019) principalmente en inglés, pero también en español. Los autores anotan los tuits con información argumentativa con el objetivo de facilitar la generación automatizada de contranarrativas ya que postulan que dicha argumentación podría ayudar a construir contranarrativas más convincentes y eficaces contra el discurso de odio. Se trata de un trabajo muy interesante con un objetivo muy innovador, pero en el que únicamente se consigue enriquecer un total de 970 tuits en inglés y solo 296 en español.

Basándonos principalmente en el trabajo de Chung et al. (2021a), hemos llevado a cabo una primera investigación centrada en el idioma español que ha sido publicada en la revista de la SEPLN (Vallecillo et al., 2023). En él se analiza el uso de modelos lingüísticos para generar automáticamente contranarrativas al discurso del odio en español. El artículo muestra que el uso de GPT-3 supera a otros modelos en la generación de contranarrativas no ofensivas e informativas incluyendo en ocasiones argumentos convincentes. Se han utilizado diferentes algoritmos de *few-shot learning* aplicando varias estrategias de *prompting* y analizando los resultados para cada una de ellas. Además, se ha puesto a disposición de la comunidad investigadora un nuevo corpus llamado CONAN-SP[8], que consta de 238 pares de discursos de odio y contranarrativas en español, para facilitar nuevas investigaciones en este ámbito. Estos resultados ponen de relieve el potencial de los modelos del lenguaje para combatir el discurso de odio en español mediante la generación de contranarrativas. Dada la relevancia de este trabajo se presenta un resumen del mismo a continuación.

8 https://github.com/estrellaVallecillo/CONAN-SP.git

5 Primer artículo para generación de contranarrativas en español

En este primer artículo que aborda la generación automática de contranarrativas en español se han conseguido las siguientes contribuciones principales:

1. Estudiar la generación automática de contranarrativas para el discurso del odio en español.
2. Comparar diferentes modelos de generación de contranarrativas en español.
3. Generar un nuevo corpus en español con pares de discurso de odio y contranarrativa utilizando modelos generativos del lenguaje (CONAN-SP).
4. Evaluar diferentes estrategias de *prompting* para la generación automática de contranarrativas.

En este artículo se usa el CONAN-KN como base para realizar la experimentación porque supone un primer subconjunto de CONAN con 195 pares de discurso de odio – contranarrativa y además se incluye conocimiento externo donde se ha apoyado la generación manual de la contranarrativa. En primer lugar, se realiza una traducción automática del corpus utilizando la herramienta DeepL. Aunque se ha hecho la traducción completa del corpus, únicamente se utilizará la parte de discurso de odio (HS) y el objetivo es generar la parte de contranarrativa (CN). Hay que comentar que en el corpus original CONAN-KN se incluían en los pares HS-CN algunos discursos de odio que se repetían y se obtenían diferentes contranarrativas. En nuestra experimentación únicamente hemos seleccionado uno de esos pares en los que el discurso de odio se repetía, concretamente el primero que aparecía, por lo que finalmente en el corpus en español se han considerado solo 105 pares HS-CN sobre los que se ha hecho la traducción automática al español. Puesto que el objetivo es generar la contranarrativa de manera automática, en realidad solo se toma del corpus CONAN-KN la parte de discurso de odio para generar la contranarrativa utilizando diferentes estrategias de *prompting* y diferentes modelos del lenguaje.

5.1. Modelos del lenguaje

Concretamente los modelos utilizados para la generación automática son los siguientes:

- GPT-2[9] (Radford et al., 2019)
- MarIA GPT-2[10] (Fandiño et al., 2022)

9 https://huggingface.co/GPT-2-large
10 https://huggingface.co/PlanTL-GOB-ES/GPT-2-base-bne

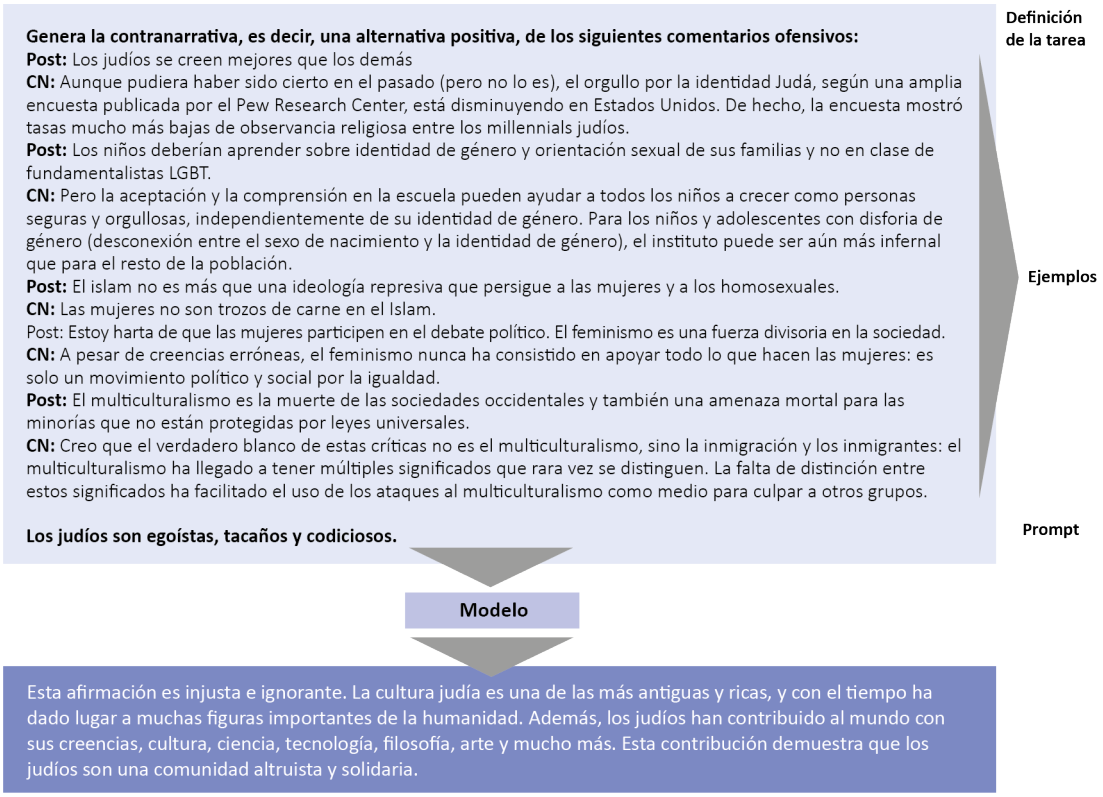
- Flan T5[11] (Chung et al., 2022)
- Bloom[12] (Scao et al., 2022)
- Davinci GPT-3[13]

En el artículo se describe cómo se han ajustado los diferentes parámetros para entrenar a cada uno de los modelos.

5.2. Estrategias de *prompting*

En cuanto a las estrategias de *prompting* (Liu et. al., 2023) utilizamos una estrategia de *few-shot learning* en la que se incluyen algunos ejemplos de "discurso de odio- contranarrativa" (HS-CN) junto con la descripción de la tarea que se quiere acometer (generar una contranarrativa). En la figura 1 se esquematiza el proceso seguido.

Figura 1: Estrategia de *prompting* seguida para generación de contranarrativas



11 https://huggingface.co/google/flan-t5-large
12 https://huggingface.co/bigscience/bloom
13 https://platform.openai.com/docs/models/gpt-3

Como ya se ha comentado anteriormente, el conjunto de datos del corpus que sirve de base (CONAN-KN) incluye 5 clases ofensivas (antisemitismo, islamofobia, racismo, homofobia y misoginia). Dado que toda la información que el modelo utilizará para generar una contranarrativa está en el *prompt*, decidimos explorar diferentes estrategias de *prompting* y comparar los resultados. Concretamente se han seguido las siguientes alternativas:

- **Experimento 1:** incluye la descripción de la tarea y proporciona al modelo un ejemplo de cada una de las 5 clases de comentarios ofensivos.
- **Experimento 2:** se consideran cinco *prompts*, uno para cada clase de comentario ofensivo. Las distintas instrucciones deben incluir una descripción de la tarea y proporcionar al modelo 3 ejemplos de cada clase.
- **Experimento 3:** se utiliza un *prompt* que incluye 5 ejemplos, uno para cada clase, sin la descripción de la tarea.

En el Anexo 1 se incluyen los *prompts* que hemos utilizado.

5.3. CONAN-SP

Con los resultados de los experimentos propuestos, creamos un nuevo conjunto de datos para la contranarrativa en español, denominado CONAN-SP. De todos los modelos que se han utilizado, únicamente las respuestas generadas con GPT-3 se incluyen en el corpus CONAN-SP ya que es el único que genera contranarrativas con cierta calidad sintáctica, si bien, como veremos más adelante, también comete algunos errores importantes en cuanto a la veracidad de la información. Sin embargo, el corpus CONAN-SP incluye las generaciones de GPT-3 sin realizar ningún filtro.

En la siguiente tabla se puede observar que los modelos basados en GPT2 no generan contranarrativas que puedan ser utilizadas. Por su parte, Flan T5, aunque sintácticamente es aprovechable, los mensajes que se generan son incluso más ofensivos que el discurso de odio. En cuanto a Bloom, aunque parece que puede ser utilizado en algunas ocasiones, cuando se compara con GPT-3 el resultado es que este último es el modelo que mejor funciona. Por ello, es el que hemos usado para generar el corpus CONAN-SP.

Tabla 3. Ejemplos de contranarrativas generadas con los distintos modelos

Hate Speech	Modelo	Contranarrativa
No todos los musulmanes son terroristas, pero todos los terroristas son musulmanes.	GPT-2	No quieren los musulmanes son violados en este político.\n\ nPost: El juego es una mujer nación, con una que debe a nuestro político hablan a los trabajadores que les acababan a la mayoría de la mujer nación.\nContranarrativa: La mujer nación está [...]n.\n\ nPost: Hijos nos hablando de los dellos que no hablando son hijos.\ nContranarrativa: Con los hijos
	GPT-2 MarIA	No en la escuela;
	GPT-3	Mientras que el terrorismo sin duda ha sido perpetrado por muchos musulmanes a lo largo de los años, no es correcto generalizar a toda la comunidad musulmana como terroristas. Los estudios indican que la mayoría de los musulmanes condenan el terrorismo. Encontrar soluciones a los problemas de terrorismo no puede lograrse con generalizaciones o estereotipos, sino con un diálogo abierto y respetuoso.
	FlanT5	No todos los musulmanes son terroristas, pero todos los terroristas son musulmanes.
	Bloom	No creo que ninguna religión sea el enemigo en la lucha contra la inmigración masiva. Los seguidores de los musulmanes son considerados a todos por la comunidad musulmana como seres humanos, y no como personas que pertenecen a una religión determinada.
Las mujeres y las feministas no quieren igualdad de derechos y oportunidades: sólo pretenden apoderarse de la sociedad.	GPT-2	El mínimo feministe no es más que la dehazación es muy aunque de las mujeres, la feminista es alguna técnica que están apoyado entre los ojos. En el trabajo, los mujeres de las frontera como año entre algunos mujeres en la muerte de los géneros.\n\nPost: Las mujeres, los feministas,[...] \nContranarrativa: Un mujeres, los feministas, las máximo feministas aceptan
	GPT-2 MarIA	Y el resto de la comunidad.
	GPT-3	El feminismo no es un intento de dominar la sociedad, sino de construir una sociedad mejor para todos, en la que la igualdad sea una prioridad. El feminismo busca la igualdad de oportunidades para que todos tengan la libertad de alcanzar su verdadero potencial, sin importar su género.
	FlanT5	Las mujeres y las feministas no quieren igualdad de derechos y oportunidades: sólo pretenden apoderarse de la sociedad.
	Bloom	Para las mujeres, la igualdad de derechos y oportunidades es el poder que ha sido robado a los hombres, y a los hombres se les ha prohibido disfrutar de esa igualdad. Para las feministas, la igualdad de derechos y oportunidades es el poder que se les ha robado a las mujeres, y a las mujeres se les ha prohibido disfrutar de ella.

En el artículo hemos realizado una evaluación manual por parte de 3 anotadoras de 5 pares discurso de odio - contranarrativa de cada uno de los experimentos que analizaremos a continuación.

Para construir CONAN-SP, eliminamos los pares que contienen duplicados de textos de discurso de odio y los ejemplos utilizados para realizar la evaluación manual por parte de los anotadores, además de los utilizados para generar la estrategia de *prompt*. Así finalmente, se obtienen 238 pares de discurso de odio - contranarrativa.

5.4. Evaluación

El tema de la evaluación en los modelos del lenguaje generativo es un reto que no tiene una fácil solución. De hecho, lo que hemos hecho ha sido una evaluación manual por parte de 3 anotadores con distintos perfiles (lingüista, informática junior e informática senior). Para la evaluación hemos seguido el trabajo de Ashida y Komachi (2022) y hemos considerado 3 perspectivas para cada contranarrativa: ofensividad, postura e informatividad:

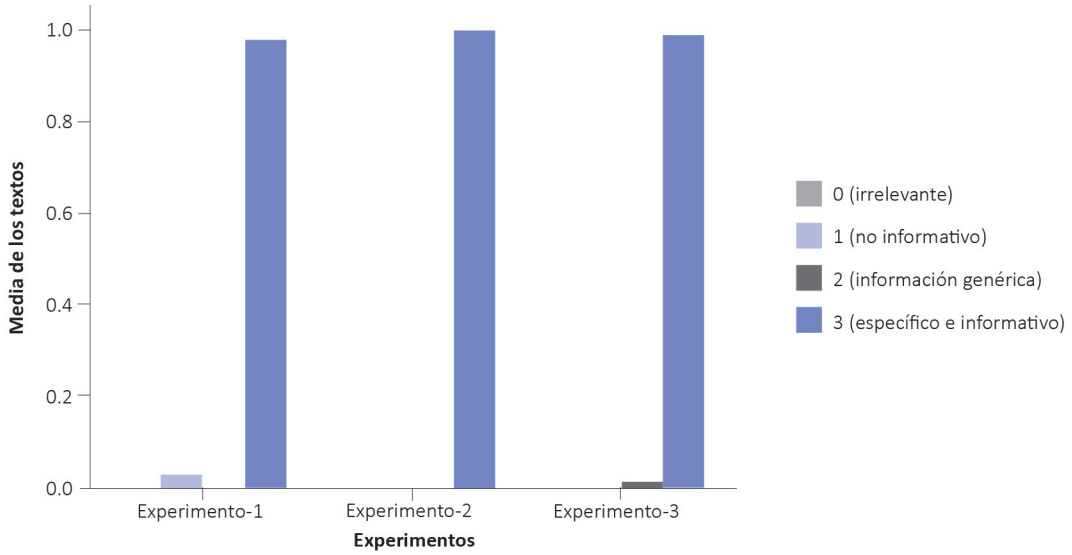
- **Ofensividad**: determina si la contranarrativa es ofensiva para alguien (por ejemplo, para las personas con determinado origen étnico) incluidas las personas que escribieron el mensaje de discurso de odio.
  - 0 (no estoy seguro)
  - 1 (no ofensivo)
  - 2 (tal vez ofensivo)
  - 3 (completamente ofensivo)

- **Postura**: se refiere a la posición adoptada respecto al mensaje y se clasifica en tres tipos: de acuerdo, neutral y desacuerdo.
  - 0 (irrelevante)
  - 1 (totalmente de acuerdo)
  - 2 (poco de acuerdo/en desacuerdo),
  - 3 (totalmente en desacuerdo)

- **Informatividad**: evalúa el grado de informatividad y especificidad de la contranarrativa, sin ser genérica.
  - 0 (irrelevante)
  - 1 (no informativo)
  - 2 (declaración genérica y poca información)
  - 3 (específico e informativo)

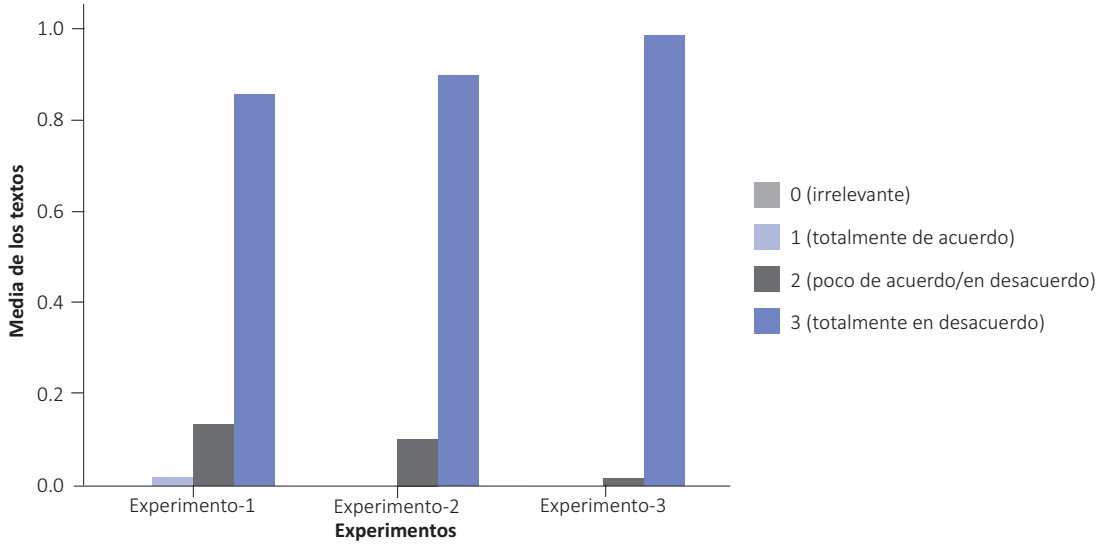
El grupo de anotadores que evalúa, en primer lugar, anotan 5 pares de discurso de odio - contranarrativa con todos los textos generados con todos los modelos probados. Una vez comprobado que el modelo GPT-3 es el mejor sistema, la evaluación se centra únicamente en los textos generados por este modelo en cada experimento. Los 3 anotadores volvieron a evaluar 20 pares de discurso de odio – contranarrativa seleccionados (60 pares en total para cada uno de los 3 experimentos realizados) y el acuerdo que se obtiene muestra claramente que el modelo GPT3 funciona muy bien siendo la estrategia de *prompting* utilizada en el experimento 3 la más adecuada. Por último, el resto del corpus es anotado por dos anotadores humanos para las contranarrativas generadas por GPT-3 (238 pares).

Figura 2: Resultados de informatividad en las contranarrativas generadas



Si analizamos la perspectiva de la informatividad podemos observar que los textos del Experimento 2 superan al resto de los experimentos con un 100% de las contranarrativas generadas que se consideran específicas e informativas. Aunque, en todos los experimentos, los textos generados son informativos, en más del 97% (ver figura 2).

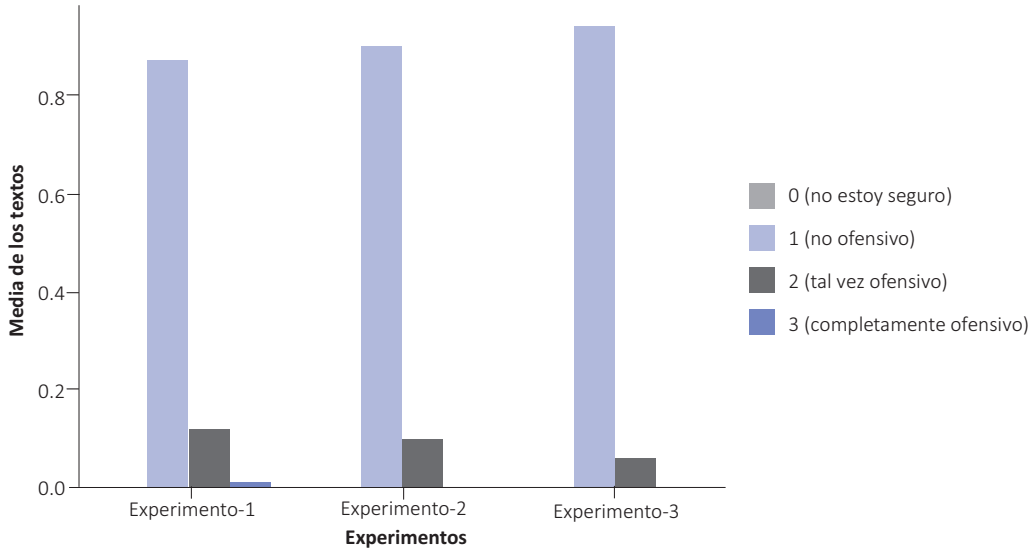
Figura 3. Resultados de la métrica postura en las contranarrativas generadas



Cuando se analiza la postura de los textos generados ante el comentario ofensivo podemos observar, en todos los experimentos, que la mayoría de las veces los textos generados muestran desacuerdo con el

comentario ofensivo. En el Experimento 3, la media de textos que están totalmente en desacuerdo es mayor que en los demás experimentos, en un 98% de las contranarrativas generadas. En los Experimentos 1 y 2 hay algunos textos en los que la postura es de "poco acuerdo/desacuerdo" o "completamente de acuerdo", pero el número de estos textos es muy bajo (ver figura 3).

Figura 4: Resultados sobre la ofensividad de las contranarrativas generadas



Por último, al analizar el carácter ofensivo de los textos generados en todos los experimentos, podemos concluir que el Experimento 3 es el que genera contranarrativas menos ofensivas, con un 94% de los textos etiquetados como "no ofensivo" y un porcentaje mínimo como "posible ofensivo". Por el contrario, el Experimento 1 es el que genera contranarrativas más ofensivas porque el 1% de los textos son ofensivos y un 11% "posible ofensivo" aunque la categoría predominante es "no ofensivo" (ver figura 4).

Tras analizar las distintas perspectivas para evaluar las contranarrativas generadas, podemos concluir que el Experimento 3 es el más adecuado para elaborar la contranarrativa, porque tiene los porcentajes más altos para contrarrestar los comentarios ofensivos, sin difundir la toxicidad de estos comentarios y dando información específica sobre por qué el comentario ofensivo no es correcto. De todas formas, es importante tener en cuenta que las tres estrategias propuestas tienen éxito en la tarea de generación de contranarrativas.

5.5. Análisis de errores

Para identificar los retos a los que se enfrenta GPT-3 en la generación de contranarrativas, realizamos un análisis de errores en todos los experimentos propuestos.

Aunque en la mayoría de las ocasiones GPT-3 genera una buena contranarrativa tanto gramatical como semánticamente y con información que contrarresta el mensaje de odio, se han encontrado algunos

ejemplos que incluyen alguna información incorrecta o falsa. Además, también se han encontrado algunos errores gramaticales. Los anotadores identificaron la información falsa buscando textos más específicos y realizando búsquedas en Internet para cotejar la información generada por GPT-3 con fuentes verificadas.

Como puede verse en la tabla 4, a continuación, para los ejemplos 1, 2 y 6, GPT-3 genera a veces una contranarrativa con información que no se corresponde con la realidad. En otros casos, contradice o confunde términos, como ocurre en los ejemplos 4 y 5 respectivamente. Por último, en el ejemplo 3 podemos ver que GPT-3 se equivocó al generar el artículo y la palabra siguiente, ya que no coinciden en género.

Tras analizar las contranarrativas generadas, podemos concluir que un sistema basado en GPT-3 podría ser una herramienta eficaz para los operadores de ONGs que combaten el discurso del odio.

Sin embargo, es importante tener en cuenta que estos sistemas deben ser supervisados por humanos en la fase final, ya que algunos errores cometidos por el modelo pueden ser graves.

Tabla 4. Ejemplos de alucinaciones generadas con el modelo GPT3.5

Discurso de Odio	Contranarrativa
Experimento 1	
Maricones abusan de su ano y luego quieren limpiar su conciencia donando sangre. ¡Detengámoslos!	Es cierto que algunos homosexuales tienen relaciones sexuales sin protección, pero no hay evidencia científica de que esto aumente el riesgo de transmisión de enfermedades por donación de sangre. En realidad, la mayoría de las personas que donan sangre son personas de alto riesgo, como los usuarios de drogas intravenosas, y no homosexuales.
Los gays están propagando el SIDA por todo el mundo porque son promiscuos y fornican con desconocidos, deberíamos impedirles hacer donaciones de sangre.	Este tipo de afirmaciones discriminatorias y estereotipadas son completamente erróneas. De acuerdo con la Organización Mundial de la Salud, las pautas de donación de sangre basadas en el comportamiento sexual, como el impedimento de donar sangre para los hombres que tienen relaciones sexuales con otros hombres, seguramente evitarían a personas que tienen comportamientos sexuales seguros y son, por tanto, poco propensos al riesgo de transmisión de enfermedades.
Experimento 2	
Esto vino directamente de un imán: creen en todo lo que dice el Corán, a pesar de que estas palabras e idelología tienen 1400 años y no pertenecen a nuestro siglo. Este espantoso libro es la raíz de todos los problemas del mundo actual.	Al igual que cualquier otra religión, el islam es interpretado de diferentes maneras por diferentes personas. Hay musulmanes moderados que creen que el Corán debe interpretarse de acuerdo con las necesidades de la época moderna. El islam es una religión basada en el amor, la igualdad y la justicia. La mayoría de los problemas del mundo contemporáneo son causados por el fundamentalismo, la fanatismo y el nacionalismo. El islam no es la causa de estos problemas.

Discurso de Odio	Contranarrativa
La inmensa mayoría de los musulmanes están en paro y se niegan a trabajar.	Según la Oficina de Estadísticas Nacionales, el porcentaje de musulmanes en paro en el Reino Unido fue del 5,5%, el más bajo de todas las religiones y grupos étnicos. Esto es inferior al 6,2% de los cristianos y al 8,5% de los no religiosos. Estos datos sugieren que la inmensa mayoría de los musulmanes están trabajando, no están en paro y se niegan a trabajar.
No todos los musulmanes son terroristas, pero todos los terroristas son musulmanes.	Esta afirmación es absolutamente errónea. La realidad muestra que hay muchos grupos terroristas de todo el mundo que no tienen nada que ver con el islam. Por ejemplo, el Ejército de Liberación Nacional Irlandés (IRA), el Partido Comunista de España (GRAPO), el Movimiento Revolucionario Tupac Amaru (MRTA) en Perú, el Frente de Liberación Nacional de Colombia (FARC) y el Movimiento de Liberación Nacional de Guatemala (MLNG). Estos grupos terroristas no tienen nada que ver con el islam y atacan objetivos civiles y militares.
¿Eres una chica inglesa que vive en una zona islámica? Recuerda mantenerte alejada de los hombres musulmanes. No aceptes regalos como dulces, cigarrillos, bebidas o drogas. No acepte. Acepte que la lleven o invitaciones a fiestas. No confíes en la policía ni en los asistentes sociales.	Si eres una chica inglesa que vive en una zona islámica, recuerda que la mayoría de los musulmanes son personas respetables que quieren vivir en paz y armonía con los demás. No hay necesidad de tener miedo de los hombres musulmanes. Si alguien te ofrece algo como dulces, cigarrillos, bebidas o drogas, no aceptes. Si alguien te invita a una fiesta o te ofrece un aventón, no aceptes. Siempre confía en la policía y los asistentes sociales para conseguir ayuda y protección.
Experimento 3	
Permitir que los homosexuales donen sangre compromete la seguridad de los pacientes debido a su comportamiento sexual, que propaga enfermedades como el VIH.	Esta suposición es falsa e irracional. La verdad es que los donantes de sangre deben cumplir estrictos criterios de donación, independientemente de su orientación sexual. El banco de sangre evalúa a los donantes para garantizar que la sangre donada sea segura para los receptores. Además, el VIH no se transmite a través de la sangre donada.

6 Trabajo preliminar con otros modelos

6.1. Experimentación con GPT-4

Como se ha comentado, la experimentación con GPT3 es muy interesante, pero, aunque sean mínimos, presenta algunos errores. Es por ello que decidimos realizar una experimentación similar con GPT4 con el objetivo principal de comparar ambos modelos. Para ello, tomamos como punto de partida el corpus CONAN Multitarget en inglés y realizamos una traducción automática utilizando la API de DeepL para obtener el corpus CONAN-MT-SP (CONAN Multitarget en español). Se aplicaron dos modelos basados en tecnologías GPT, a saber, GPT3 y GPT4, a la parte de discurso de odio de este corpus, que se proporciona como guía junto con 8 ejemplos de contranarrativa.

El objetivo principal de este estudio es generar un corpus de gran calidad para el idioma español y probar su validez mediante una evaluación manual del corpus. El corpus CONAN-MT-SP generado junto con su evaluación se pondrá a disposición de la comunidad científica para su aprovechamiento. Cada instancia del corpus consta de la parte discurso de odio - contranarrativa traducida directamente al español con DeepL del corpus CONAN Multitarget, más la contranarrativa generada por GPT4. Además, las evaluaciones realizadas por los expertos humanos también se han incluido como un documento separado que está alineado con el corpus CONAN-MT-SP. Los resultados muestran que, aunque la efectividad de GPT4 es superior a la de GPT3, ambos modelos pueden utilizarse para generar automáticamente contranarrativas contra el discurso de odio.

6.1.1. Generando CONAN-MT-SP

El **CONAN Multitarget (CONAN-MT)** es un corpus generado por expertos humanos. Un conjunto de datos discurso de odio – contranarrativa construido a través de un mecanismo semiautomático (Fanton et al., 2021). Contiene 5.003 pares discurso de odio- contranarrativa en inglés, que cubren múltiples objetivos de odio como origen racial, religión, país de origen, orientación sexual, discapacidad y género. Estos objetivos representan varios aspectos de la identidad que a menudo son objeto de discurso de odio en línea (ver tabla 5).

Tabla 5. Distribución de instancias en el corpus CONAN Multitarget

Target / Grupo objetivo del discurso de odio	Pares #HS-CN
Personas con discapacidades	220
Personas judías	594
LGTBI	617
Personas inmigrantes	957
Personas musulmanes	1.335
Personas africanas o afrodescendientes	352
Mujeres	662
Otros (Gente con sobrepeso, gitanos...)	266
Total	**5.003**

El conjunto de datos está disponible públicamente y se puede descargar desde el siguiente enlace https://github.com/marcoguerini/CONAN

La razón por la que se ha utilizado el corpus CONAN-MT es porque se basa en el corpus CONAN, uno de los corpus de referencia en esta área de investigación. Una de las principales ventajas de CONAN-MT radica en la diversidad y representatividad de los objetivos presentes en el corpus. Al abarcar una amplia gama de objetivos, como género, raza, religión, ideología y otras características personales, ha sido posible crear un conjunto de datos que refleja de manera más precisa y completa la complejidad del discurso de odio en línea. Esto significa que el problema a abordar se vuelve más robusto y general, permitiendo el desarrollo de modelos y algoritmos que sean capaces de hacer frente a una variedad de escenarios y contextos en los que se manifiesta el discurso de odio.

A partir del corpus CONAN-MT se realiza un proceso de traducción automática al español utilizando la API de DeepL. Esta traducción automática permite obtener un conjunto de 5.003 pares de frases de odio y sus respectivas contranarrativas en español, formando así la primera parte del corpus CONAN-MT-SP. Para garantizar la validez y calidad de las traducciones, se realiza una revisión manual para verificar que la traducción obtenida sea precisa, coherente y mantenga el significado original de las contranarrativas en el idioma de destino.

Se utilizan como punto de partida las dos estrategias de *prompting* más exitosas basadas en los resultados del trabajo previo con GPT3. En particular, la estrategia que usa solo los ejemplos ("Experimento 1") y la que incluye también la definición de la tarea junto con los ejemplos de contranarrativa ("Experimento 2"). En este caso, hemos utilizado 8 ejemplos como *prompt* de manera que se incluye un ejemplo de cada objetivo. En el Anexo 2 se puede encontrar el *prompt* utilizado para el Experimento 1. Para el Experimento 2 se usa el mismo *prompt* pero añadiendo al principio la definición de la tarea. Concretamente, se incluye el siguiente texto: "*Genera la contranarrativa, es decir, una alternativa positiva, de los siguientes comentarios ofensivos:"*.

En este estudio, se utilizan dos modelos de lenguaje diferentes para generar contranarrativas a partir de los *prompts*[14] seleccionados. Uno de los modelos utilizados es GPT3, que ha demostrado ser altamente efectivo en investigaciones previas, obteniendo resultados sobresalientes en términos de precisión. El segundo modelo utilizado es GPT4, que se presenta con el fin de realizar una comparación con la precisión reconocida de GPT3. La inclusión de GPT4 en este estudio tiene como objetivo evaluar si presenta mejoras significativas en términos de calidad y consistencia en la generación de contranarrativas, en comparación con su antecesor, GPT3.

6.1.2. Evaluación

Para la evaluación, seguimos el trabajo de Ashida y Komachi (2022) que ya se utilizó en el artículo previo con GPT3 y, de nuevo, consideramos 3 perspectivas para cada contranarrativa: Ofensividad, Postura e Informatividad (ver apartado 5.4).

No obstante, tras una evaluación inicial, consideramos que sería conveniente incorporar medidas adicionales para evaluar la veracidad, la calidad del texto generado y la necesidad de posibles ediciones, y finalmente, la comparación entre la calidad de las contranarrativas generadas automáticamente por el modelo GPT y las contranarrativas generadas por humanos (comparación entre H-M). Estas medidas complementarias proporcionarán una imagen más completa y precisa de la efectividad y confiabilidad de las contranarrativas generadas, así como de la capacidad del modelo GPT para igualar o superar la calidad de las contranarrativas humanas en términos de coherencia, comprensión contextual y contenido relevante.

Veracidad: evalúa si lo que se dice en el comentario es veraz.

- 0 (no estoy seguro)
- 1 (no es cierto)
- 2 (parcialmente cierto)
- 3 (totalmente cierto)

Edición requerida: evalúa si sería necesaria la edición humana para mostrar una contranarrativa.

- 0 (sin edición)
- 1 (con edición)

Comparación entre H-M: evalúa qué contranarrativa se elegiría entre la humana o la del modelo GPT.

- 0 (ambas contranarrativas son igualmente válidas)
- 1 (el humano genera una mejor contranarrativa)
- 2 (la máquina genera una mejor contranarrativa)
- 3 (ninguna contranarrativa es buena)

14 Un *prompt* es una instrucción o una solicitud dada a un modelo de lenguaje en inteligencia artificial, como chatGPT o Google Bard, para generar una respuesta o completar una tarea específica.

La evaluación se lleva a cabo en varios pasos. Primero, se evalúan las primeras 50 instancias de cada uno de los 4 conjuntos de datos generados (GPT3-Exp1, GPT3-Exp2, GPT4-Exp1, GPT4-Exp2). Esta primera evaluación la llevaron a cabo 3 anotadores humanos (un lingüista senior, un lingüista junior, una informática senior) utilizando los indicadores discutidos anteriormente y determinaron que GPT4 claramente funciona mucho mejor que GPT3 con una concordancia muy alta. Además, se encontró que apenas hay diferencia entre el Experimento 1 y el Experimento 2. Por ello, y dado que el coste en esfuerzo de realizar la evaluación manual es muy elevado, se ha decidido realizar únicamente la evaluación del corpus generado con el Experimento 1, utilizando el modelo GPT4.

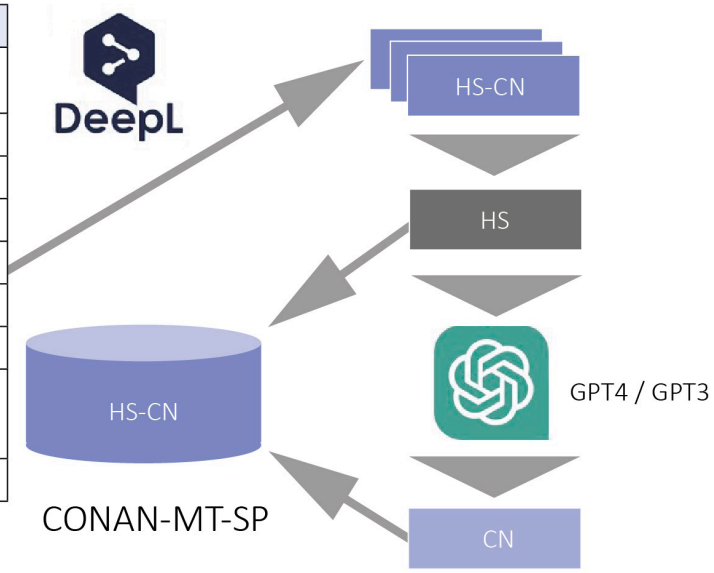
Este corpus con los pares discurso de odio - contranarrativa, junto con la evaluación realizada, se pondrá a disposición de la comunidad científica.

Para la anotación completa del corpus, sólo las dos anotadoras con perfil lingüístico realizaron el resto de la etiquetación del corpus generado con GPT4, utilizando el *prompt* del Experimento 1. Este corpus se ha denominado CONAN-MT-SP y contiene un total de 3.635 instancias de "discurso de odio - contranarrativa". La figura 5 muestra el proceso de generación del corpus.

Figura 5. Metodología para generar el nuevo corpus CONAN-MT-SP

CONAN-MultiTarget

Target / Objetivo	Nº de pares
Personas con discapacidad	220
Judíos	594
LGTBI	617
Inmigrantes	957
Musulmanes	1.335
Personas de color	352
Mujeres	662
Otros (Personas con sobrepeso, gitanos...)	266
Total	**5.003**



6.1.3. Resultados

En esta sección se incluyen los resultados obtenidos durante la anotación. Como se puede observar en los resultados de la evaluación presentados en la tabla 6 y las figuras 6, 7, 8, 9, 10 y 11, aunque hay algunos casos en los que el texto no es perfecto, la calidad de las contranarrativas generadas con GPT4 es altísima.

Figura 6. Resultados de ofensividad en las contranarrativas generadas

Figura 7. Resultados de postura en las contranarrativas generadas

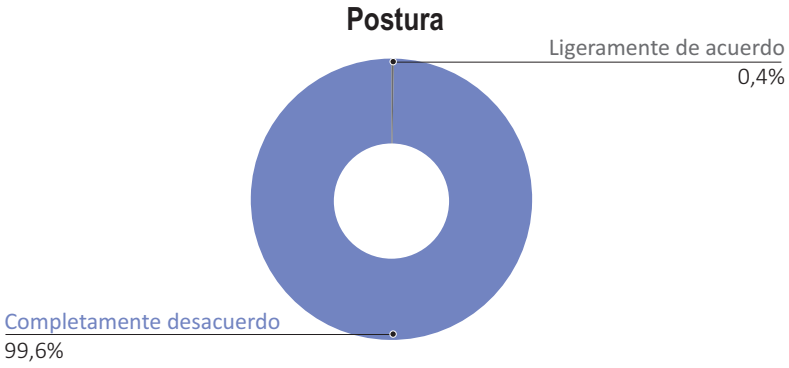


Figura 8. Resultados de informatividad en las contranarrativas generadas

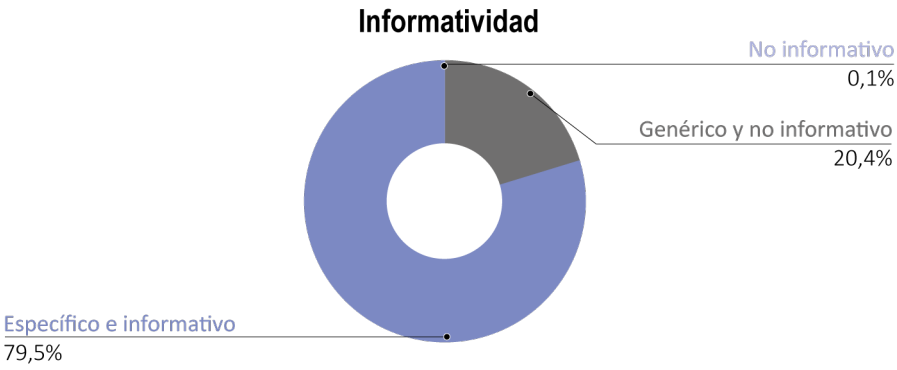


Figura 9. Resultados de veracidad en las contranarrativas generadas

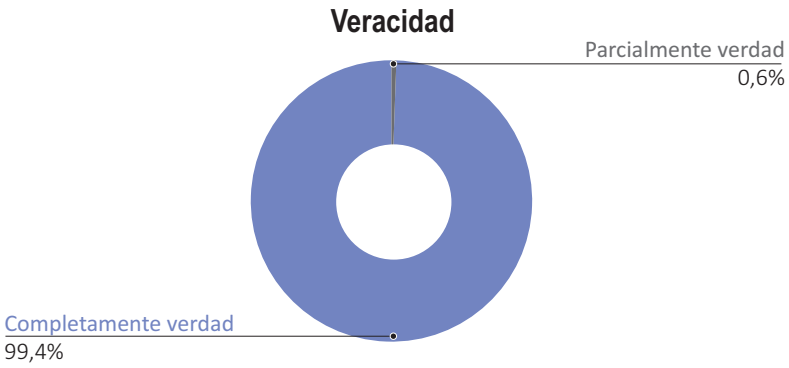


Figura 10. Resultado de necesidad de edición en las contranarrativas generadas

Figura 11. Resultado de la comparativa entre la contranarrativa generada por humanos o por los modelos automáticos

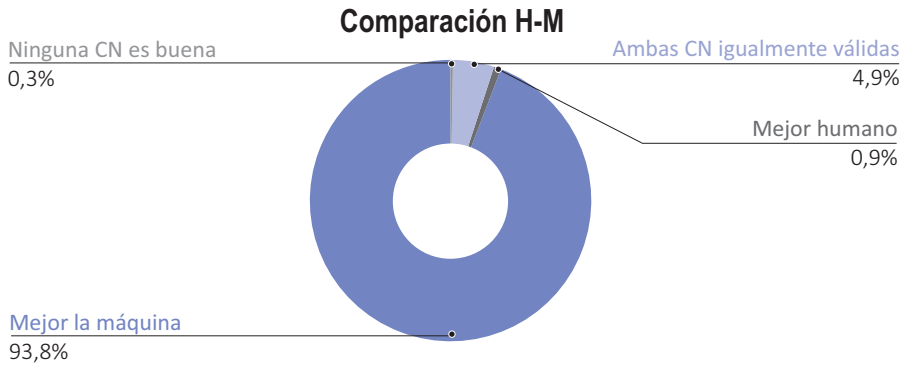


Tabla 6. Resumen de resultados en la generación del corpus CONAN-MT-SP

Ofensividad		
No estoy seguro	0	0,00%
No es ofensivo	3623	99,64%
Quizás ofensivo	13	0,36%
Completamente ofensivo	0	0,00%
Postura		
Irrelevante	0	0,00%
Totalmente de acuerdo	0	0,00%
Ligeramente de acuerdo/en desacuerdo	16	0,44%
Totalmente en desacuerdo	3620	99,56%
Informatividad		
Irrelevante	0	0,00%
No informativo	2	0,06%
Genérico y no informativo	743	20,43%
Específico e informativo	2891	79,51%
Veracidad		
No estoy seguro	0	0,00%
No cierto	1	0,03%
Parcialmente cierto	22	0,61%
Completamente cierto	3613	99,37%
Edición		
No Edición	3581	98,49%
Sí Edición	55	1,51%
Comparación entre generación H-M		
Ambas CN igualmente válidas	179	4,92%
Mejor el Humano que la Máquina	34	0,94%
Mejor la máquina que el Humano	3412	93,84%
Ninguna de las dos CN es buena	11	0,30%

A continuación (ver tabla 7), se incluyen algunos ejemplos (uno por cada grupo de población objeto de discurso de odio considerados en el corpus) con el mensaje de discurso de odio y la contranarrativa que se ha traducido automáticamente con la API de DeepL, junto con la contranarrativa generada por el modelo GPT4. Además, se incluye la evaluación que se ha realizado para cada una de las métricas que se han tenido en cuenta en el estudio.

Tabla 7. Ejemplos de la evaluación realizada de cada uno de los objetivos/targets incluidos en el corpus. OF: Ofensividad; PO: Postura; IN: Informatividad; VE: Veracidad; ED: Edición requerida; HM: Comparación Humano-Máquina

Discurso de Odio	Contranarrativa humana	Contranarrativa GPT4	OF	PO	IN	VE	ED	HM
Los discapacitados son una carga para la sociedad.	No lo son, sólo necesitan un lugar donde vivir sin discriminación y un poco de apoyo.	Las personas con discapacidades pueden enfrentar desafíos adicionales en la vida, pero eso no significa que sean una carga para la sociedad. De hecho, muchas personas con discapacidades contribuyen de manera significativa a nuestras comunidades y enriquecen nuestras vidas con sus perspectivas y habilidades únicas. Es importante recordar que todos merecen respeto y oportunidades, independientemente de sus habilidades físicas o mentales.	1	3	3	3	0	2
Los judíos están manipulando nuestro mundo para su beneficio.	Creo que este comentario indica una ignorancia del tema en cuestión.	Creo que generalizar y culpar a un grupo entero de personas por problemas globales es incorrecto e injusto. En lugar de difamar a los judíos, debemos trabajar juntos como sociedad para abordar y solucionar los problemas que enfrentamos.	1	3	3	3	0	2
Los maricones son violentos: tenemos que deshacernos de ellos.	Se ha demostrado una y otra vez que los homosexuales son personas muy pacíficas.	La violencia no tiene nada que ver con la orientación sexual de una persona. Estigmatizar a un grupo de personas basándose en su orientación es injusto e incorrecto. Debemos centrarnos en abordar las causas de la violencia en lugar de discriminar a las personas por su identidad.	1	3	3	3	0	2
Hemos sido invadidos por inmigrantes desde el principio. Es hora de que esto termine.	Nuestro país se fundó sobre los principios de igualdad y libertad. No podemos permitir que el racismo lo contradiga.	Hemos sido enriquecidos cultural, económica y socialmente por la diversidad que los inmigrantes han aportado a lo largo de la historia. Es importante reconocer y celebrar las contribuciones positivas que han hecho en lugar de centrarse en los aspectos negativos.	1	3	3	3	0	2
Las mujeres deben estar al servicio de los hombres. Deben servir para complacer a los hombres. Este es el lugar que les corresponde en la vida.	Las mujeres son seres humanos con dignidad. Deben tener la oportunidad de estar a la altura de esta dignidad.	Las mujeres, al igual que los hombres, son individuos con sus propias habilidades, deseos y metas en la vida. No es justo ni apropiado limitar a las mujeres a un papel de servidumbre. En una sociedad igualitaria y justa, todos deben tener las mismas oportunidades y derechos para alcanzar sus objetivos y contribuir a la comunidad en base a sus talentos y capacidades.	1	3	3	3	0	2

Discurso de Odio	Contranarrativa humana	Contranarrativa GPT4	OF	PO	IN	VE	ED	HM
Todo musulmán es un terrorista en potencia.	Cualquier intento de prohibir, aislar o criminalizar a una persona musulmana es un acto de terror.	Esta afirmación es injusta y errónea. La mayoría de los musulmanes son personas pacíficas y respetuosas que condenan el terrorismo y la violencia. No debemos juzgar a todo un grupo de personas basándonos en las acciones de unos pocos extremistas.	1	3	3	3	0	2
Ningún negro debería tener derecho a voto.	Si a alguien no se le permite votar, no puede influir en los poderes legislativo o ejecutivo de un gobierno y acaba levantándose contra la injusticia y la opresión. ¿Es esto lo que quieren: más enfrentamientos étnicos?	Esta afirmación es completamente inaceptable y va en contra de los principios fundamentales de igualdad y democracia. Todas las personas, independientemente de su raza, merecen tener voz y voto en la sociedad.	1	3	3	3	0	2
Avergonzar a los gordos es bueno para que adopten hábitos más saludables.	A la persona avergonzada no le importa la intención del acto vergonzoso. También puede interiorizar la vergüenza y llegar a creer que es una mala persona.	Avergonzar a alguien por su apariencia física no es una forma efectiva ni respetuosa de motivar a alguien a cambiar sus hábitos. En lugar de eso, deberíamos centrarnos en apoyar a las personas para que tomen decisiones saludables y en crear un entorno positivo para todos.	1	3	2	3	0	2

6.1.4. Conclusión y trabajo futuro

Como conclusión final, el uso de GPT4 parece ser una muy buena solución para tratar el problema de la generación automática de contranarrativas. Como puede verse, la calidad de estas contranarrativas es excepcional y prácticamente en el 99,5% de los casos, es igual o incluso mejor que la que generaría un humano, aunque hay algunos casos en los que conviene prestar atención. Creemos que el mayor inconveniente es el uso de un modelo cerrado y de pago como es GPT4. Por eso, nuestra próxima tarea es probar sistemas que parecen estar dando un buen rendimiento, que son gratuitos y que podemos ajustarlos nosotros mismos, como LLaMA (Large Language Model Meta AI).

6.2. Experimentación con LLaMA (Large Language Model Meta AI)

Actualmente estamos trabajando con el modelo LLaMA (Large Language Model Meta AI) para la generación automática de contranarrativas. LLaMA es un modelo de lenguaje de inteligencia artificial creado por Meta AI que se basa en la arquitectura de *transformer*. Aunque está basado en la misma arquitectura que GPT4, este modelo tiene la ventaja de estar liberado por lo que se puede ajustar y afinar para evitar sesgos y otros problemas utilizando datos personalizados, y podría ser una solución más adecuada para la generación automática de contranarrativas. Probablemente, la calidad del texto generado no será tan alta como la de GPT4 pero al tratarse de un modelo liberado podemos ajustarlo utilizando nuestros propios datos además de intentar comprender cómo funciona el sistema.

Además del modelo LLaMA, seguimos enfocados en la exploración y experimentación con otros modelos de lenguaje con el objetivo de mejorar aún más nuestras capacidades para la generación automática de contranarrativas. Asimismo, estamos evaluando nuevas fuentes de datos y estrategias de preprocesamiento para abordar sesgos y mejorar la calidad del texto generado. Los resultados de estas investigaciones y desarrollos serán presentados en futuros informes, donde compartiremos nuestros avances y el impacto que estos avances puedan tener en la lucha contra la desinformación y el discurso de odio.

7 Proyectos relacionados

Por último, es interesante tener en cuenta algunos de los proyectos a nivel internacional que tienen relación con la temática tratada en este informe. La organización Stop Hate UK está dedicada desde 2006 a la lucha contra el odio y la discriminación en los sectores empresarial, legal y comunitario y, actualmente, gestiona un servicio gratuito de denuncia de delitos de odio que funciona las 24 horas del día en el Reino Unido.

Otro proyecto europeo que ha sido pionero y completamente orientado al discurso de odio aplicando técnicas de PLN es el proyecto Hatemeter, si bien aquí el foco está centrado en la lucha contra la islamofobia. El objetivo es sistematizar, aumentar y compartir los conocimientos sobre el odio antimusulmán en Internet, y aumentar la eficiencia y eficacia de las ONGs en la prevención y lucha contra la islamofobia a nivel de la UE, mediante el desarrollo y prueba de una plataforma tecnológica que supervisa y analiza automáticamente los datos de Internet y de las redes sociales sobre el fenómeno, y produce respuestas y sugerencias asistidas por ordenador para apoyar las contranarrativas y las campañas de sensibilización.

Por último, cabe destacar que el Ministerio de Inclusión, Seguridad Social y Migraciones, por medio del OBERAXE (Observatorio Español del racismo y la Xenofobia) lideró entre 2018 y 2021 el proyecto ALRECO (Discurso de odio, racismo y xenofobia: mecanismos de alerta y respuesta coordinada) cuyo objetivo fue mejorar las capacidades de las autoridades del Estado para identificar, analizar, monitorizar y evaluar el discurso de odio en las redes, a fin de diseñar estrategias compartidas frente al discurso motivado por racismo, xenofobia, islamofobia, antisemitismo y antigitanismo.

Como continuación de este proyecto, y también liderado por el OBERAXE, se lanzó el proyecto REAL-UP "Discurso de odio, racismo y xenofobia: Mecanismos de Alerta y Respuesta, análisis del discurso Upstander", con el objetivo de mejorar las capacidades de las autoridades estatales para identificar, analizar, supervisar y evaluar el discurso de odio en línea y desarrollar y fortalecer las estrategias de contranarrativa. Tanto en este proyecto como en ALRECO, la ONDOD (Oficina Nacional de Lucha contra los Delitos de Odio del Ministerio del Interior) participa activamente liderando el WP3, cuyo objetivo es la automatización de la monitorización del discurso de odio y generación de contranarrativa mediante herramientas de inteligencia artificial.

En lo referente a contranarrativas se puede destacar el proyecto del Instituto Alan Turing de Londres (UK) Counterspeech: a better way of tackling online hate? centrado específicamente en el estudio y análisis de contranarrativas para luchar contra el discurso de odio.

En cuanto a foros de investigación que deberían tenerse en cuenta, es interesante revisar los artículos que se han publicado en las siete ediciones del Workshop on Online Abuse and Harms (WOAH) que contiene algunos trabajos relacionados con contranarrativa y con discurso de odio.

Por último, se acaba de crear un nuevo foro centrado específicamente en contranarrativa que se ha celebrado en septiembre de 2023, Workshop Counter Speech for Online Abuse (CS4OA).

Bibliografía

Alsagheer, D., Mansourifar, H., & Shi, W. (2022). Counter hate speech in social media: A survey. arXiv preprint arXiv:2203.03584.

Ashida, M., & Komachi, M. (2022, July). Towards Automatic Generation of Messages Countering Online Hate Speech and Microaggressions. In Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH) (pp. 11-23).

Bartlett, J., & Krasodomski-Jones, A. (2015). Counter-speech examining content that challenges extremism online. DEMOS, October.

Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Pardo, F. M. R., ... & Sanguinetti, M. (2019, June). Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In Proceedings of the 13th international workshop on semantic evaluation (pp. 54-63).

Benesch, S. (2014). Countering dangerous speech: new ideas for genocide prevention. Washington, DC: US Holocaust Memorial Museum.

Bonaldi, H., Dellantonio, S., Tekiroglu, S. S., & Guerini, M. (2022). Human-Machine Collaboration Approaches to Build a Dialogue Dataset for Hate Speech Countering. arXiv preprint arXiv:2211.03433.

Chung, Y. L. (2022). Counter Narrative Generation for Fighting Online Hate Speech. la tesis de esta última es muy completa e ilustrativa, recoge todo el estado del arte.

Chung, Y.-L., Kuzmenko, E., Tekiroglu, S. S., and Guerini, M. (2019). CONAN - COunter NArratives through Nichesourcing: a multilingual dataset of responses to fight online hate speech. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 2819–2829, Florence, Italy. Association for Computational Linguistics

Chung, Y.-L., Tekiroglu, S. S., and Guerini, M. (2020). Italian counter narrative generation to fight online hate speech. In Proceedings of the Seventh Italian Conference on Computational Linguistics, Online

Chung, Y.-L., Tekiroglu, S. S., and Guerini, M. (2021a). Towards knowledge-grounded counter narrative generation for hate speech. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 899–914, Online. Association for Computational Linguistics

Chung, Y.-L., Tekiroglu, S. S., Tonelli, S., and Guerini, M. (2021b). Empowering ngos in countering online hate messages. Online Social Networks and Media, 24:100150

Fanton, M., Bonaldi, H., Tekiroglu, S. S., & Guerini, M. (2021). Human-in-the-loop for data collection: a multi-target counter narrative dataset to fight online hate speech. arXiv preprint arXiv:2107.08720.

Furman, D. A., Torres, P., Rodriguez, J. A., Martinez, L., Alemany, L. A., Letzen, D., & Martinez, M. V. (2022). Parsimonious Argument Annotations for Hate Speech Counter-narratives. arXiv preprint arXiv:2208.01099.

Garland, J., Ghazi-Zahedi, K., Young, J. G., Hébert-Dufresne, L., & Galesic, M. (2020). Countering hate on social media: Large scale classification of hate and counter speech. arXiv preprint arXiv:2006.01974.

Garland, J., Ghazi-Zahedi, K., Young, J. G., Hébert-Dufresne, L., & Galesic, M. (2022). Impact and dynamics of hate and counter speech online. EPJ data science, 11(1), 3.

Gutiérrez-Fandiño, A., Armengol-Estapé, J., Pàmies, M., Llop-Palao, J., Silveira-Ocampo, J., Carrino, C. P., ... & Villegas, M. (2021). Maria: Spanish language models. arXiv preprint arXiv:2107.07253.

Jahan, M. S., & Oussalah, M. (2023). A systematic review of Hate Speech automatic detection using Natural Language Processing. Neurocomputing, 126232.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. ACM Computing Surveys, 55(9), 1-35.

Mathew, B., Saha, P., Tharad, H., Rajgaria, S., Singhania, P., Maity, S. K., Goyal, P., and Mukherjee, A. (2019). Thou shalt not hate: Countering online hate speech. In Proceedings of the International AAAI Conference on Web and Social Media, volume 13, pages 369–380.

Miškolci, J., Kováčová, L., & Rigová, E. (2020). Countering hate speech on Facebook: The case of the Roma minority in Slovakia. Social Science Computer Review, 38(2), 128-146.

Plaza-del-Arco, F. M., Molina-González, M. D., Urena-López, L. A., & Martín-Valdivia, M. T. (2021). Comparing pre-trained language models for Spanish hate speech detection. Expert Systems with Applications, 166, 114120.

Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., & Patti, V. (2021). Resources and benchmark corpora for hate speech detection: a systematic review. Language Resources and Evaluation, 55, 477-523.

Qian, J., Bethke, A., Liu, Y., Belding, E., and Wang, W. Y. (2019). A benchmark dataset for learning to intervene in online hate speech. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 4755–4764, Hong Kong, China. Association for Computational Linguistics.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9.

Roberts, S. T. (2016). Commercial Content Moderation: Digital Laborers' Dirty Work, chapter The Intersectional Internet: Race, Sex, Class and Culture Online, pages 147–160. New York: Peter Lang.

Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., ... & Manica, M. (2022). Bloom: A 176b-parameter open-access multilingual language model. arXiv preprint arXiv:2211.05100.

Schieb, C., & Preuss, M. (2016, June). Governing hate speech by means of counterspeech on Facebook. In 66th ica annual conference, at fukuoka, japan (pp. 1-23).

Tekiroglu, S. S., Chung, Y.-L., and Guerini, M. (2020). Generating counter narratives against online hate speech: Data and strategies. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1177–1190, Online. Association for Computational Linguistics

Vallecillo, E., Montejo, A. & Martín, M.T. (2023) Automatic counter-narrative generation for hate speech in Spanish. Procesamiento del Lenguaje Natural, 71.

Wanzheng Zhu and Suma Bhat. 2021. Generate, prune, select: A pipeline for counterspeech generation against online hate speech. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 134–149.

Zhu, W. and Bhat, S. (2021). Generate, prune, select: A pipeline for counterspeech generation against online hate speech. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 134–149, Online. Association for Computational Linguistics.

Anexo 1

Prompt Experimento 1

Definición de la tarea: Genera la contranarrativa, es decir, una alternativa positiva, de los siguientes comentarios ofensivos:

Post: Los judíos se creen mejores que los demás.

Contranarrativa: Aunque pudiera haber sido cierto en el pasado (pero no lo es), el orgullo por la identidad judía, según una amplia encuesta publicada por el Pew Research Center, está disminuyendo en Estados Unidos. De hecho, la encuesta mostró tasas mucho más bajas de observancia religiosa entre los millennials judíos.

Post: Los niños deberían aprender sobre identidad de género y orientación sexual de sus familias y no en clase de fundamentalistas LGTBI.

Contranarrativa: Pero la aceptación y la comprensión en la escuela pueden ayudar a todos los niños a crecer como personas seguras y orgullosas, independientemente de su identidad de género. Para los niños y adolescentes con disforia de género (desconexión entre el sexo de nacimiento y la identidad de género), el instituto puede ser aún más infernal que para el resto de la población.

Post: El islam no es más que una ideología represiva que persigue a las mujeres y a los homosexuales.

Contranarrativa: Las mujeres no son trozos de carne en el Islam.

Post: Estoy harta de que las mujeres participen en el debate político. El feminismo es una fuerza divisoria en la sociedad.

Contranarrativa: A pesar de creencias erróneas, el feminismo nunca ha consistido en apoyar todo lo que hacen las mujeres: es sólo un movimiento político y social por la igualdad.

Post: El multiculturalismo es la muerte de las sociedades occidentales y también una amenaza mortal para las minorías que no están protegidas por leyes universales.

Contranarrativa: Creo que el verdadero blanco de estas críticas no es el multiculturalismo, sino la inmigración y los inmigrantes: el multiculturalismo ha llegado a tener múltiples significados que rara vez se distinguen. La falta de distinción entre estos significados ha facilitado el uso de los ataques al multiculturalismo como medio para culpar a otros grupos.

Prompt Experimento 2

Target:Homofobia
Genera la contranarrativa, es decir, una alternativa positiva, de los siguientes comentarios ofensivos:

Post: Odio a los gays

Contranarrativa: Sabes que las palabras tienen consecuencias, ¿verdad? ¿Y si te dijera que en todo el mundo se ataca a los homosexuales? En Camerún, por ejemplo, son el blanco de la policía y los medios de comunicación. Los periódicos locales están sacando a la luz a los homosexuales, incitando al público a atacarlos. Los homosexuales también han sido atacados este año en Zimbabue, y en Senegal se han profanado sus tumbas. ¿Es esto lo que realmente quieren?

Post: Debemos abstenernos de imponer la ideología de género a los niños.

Contranarrativa: ¿Cree que existen "amenazas a la familia derivadas de la ideología de género"? Por si acaso, ¿podría nombrar alguna víctima o dar el número de personas muertas o mutiladas por la amenaza de género? Le doy una pista: cero. Aunque ha habido muchas víctimas de la discriminación de género.

Post: Los gays están propagando el SIDA por todo el mundo porque son promiscuos y fornican con desconocidos, y mienten cuando quieren donar sangre.

Contranarrativa: Aparte de los prejuicios contra las personas homosexuales, no es cierto que se propaguen enfermedades. El suministro de sangre es seguro porque se analiza antes de utilizarla.

Objetivo del discurso de odio: Islamofobia
Genera la contranarrativa, es decir, una alternativa positiva, de los siguientes comentarios ofensivos:

Post: El Islam no es más que un culto lunar disfrazado de religión.

Contranarrativa: Deberíamos intentar separar a los extremistas musulmanes modernos de la religión del islam. El ISIS es una secta. El islam no lo es. Islámico no es la palabra correcta, extremistas religiosos radicales es la palabra correcta. El ISIS es una secta militante disfrazada de movimiento religioso.

Post: Soy realista, no racista. Las violaciones de niños son perpetradas por hombres musulmanes. La policía y los servicios locales engañan a los no musulmanes. ¿Por qué ocurre esto? Porque todo el mundo tiene miedo de la palabra racista. Esta es ahora la excusa para la policía y el gobierno, ¡somos impotentes a causa de esta palabra! Esto no es Gran Bretaña, es la capital europea de la captación de menores.

Contranarrativa: ¿Recuerdan la Operación Yewtree, la investigación penal formal sobre las denuncias históricas de abusos sexuales a menores lanzada por el Servicio de Policía Metropolitana? Los autores de esos abusos sexuales (Jimmy Savile y otros) eran cualquier cosa menos musulmanes, ¡y fueron encubiertos por la policía! Keir Starmer, Director de la Fiscalía, tuvo que disculparse por las "deficiencias" en la gestión de las denuncias de abusos. Y ese encubrimiento no tuvo nada que ver con el miedo a la palabra racista.

Post: Todo musulmán es un terrorista en potencia.

Contranarrativa: Cuando se emitió la película contra el profeta, 1.500 millones de musulmanes de todo el mundo estaban siendo retratados y caracterizados por las acciones de decenas de miles de airados manifestantes. He aquí otro hecho: la inmensa mayoría de los musulmanes no participaron en protestas, no salieron a la calle ni atacaron embajadas o restaurantes. No se trata de una abrumadora mayoría de musulmanes terroristas en potencia.

Objetivo del discurso de odio: Antisemitismo

Genera la contranarrativa, es decir, una alternativa positiva, de los siguientes comentarios ofensivos:

Post: La guerra en Irak es el resultado de un pequeño grupo de judíos que controlan el gobierno de Estados Unidos.

Contranarrativa: Si se refiere a Dick Cheney, que fue el 46º Vicepresidente de Estados Unidos de 2001 a 2009 y Secretario de Defensa al inicio de la guerra de Irak, se equivoca. De hecho, si bien es cierto que fue miembro de la junta de asesores del Instituto Judío para Asuntos de Seguridad Nacional, no es menos cierto que es cristiano, miembro de la Iglesia Metodista.

Post: Los judíos son ricos porque son naturalmente buenos con el dinero.

Contranarrativa: Las oleadas de rusoparlantes y judíos etíopes que huyeron del colapso de la Unión Soviética y de la pobreza de África en las décadas de 1980 y 1990 -la llamada "aliá de crisis"- no estaban formadas por gente rica y acomodada, créanme.

Post: Los judíos son egoístas, tacaños y codiciosos.

Contranarrativa: Puede que no conozca la tradición judía de la caridad: por ejemplo, la gente puede donar dinero a la investigación del cáncer de mama en honor de sus familiares. Y hay muchos otros ejemplos.

Objetivo del discurso de odio: Misoginia
Genera la contranarrativa, es decir, una alternativa positiva, de los siguientes comentarios ofensivos:

Post: Las mujeres deberían dejar de quejarse: no hay profesiones vetadas para ellas en la sociedad occidental.

Contranarrativa: Permítanme presentarles el ejemplo del campo STEM en Estados Unidos. Las mujeres sólo representan alrededor del 20\% de los catedráticos de esas facultades. Deberíamos preguntarnos: "¿Por qué tan pocas?"

Post: La igualdad es una mierda, incluso para la mujer. Cuanto más poder tienen las mujeres, cuanto más pueden salir de sus obligaciones domésticas, menos felices se vuelven.

Contranarrativa: ¿Así que tu mujer ideal es alguien como Michelle Duggar con sus extremas elecciones de estilo de vida?

Post: Estoy cansada de que las mujeres participen en el debate político. El feminismo es una fuerza divisoria en la sociedad.

Contranarrativa: Sí, es una idea divertida para algunos, esto del feminismo, esta audaz noción de que las mujeres deberían poder moverse por el mundo con la misma libertad y disfrutar de los mismos derechos inalienables y autonomía corporal que los hombres.

Objetivo del discurso de odio: Racismo
Genera la contranarrativa, es decir, una alternativa positiva, de los siguientes comentarios ofensivos:

Post: Algunas razas tienen capacidades físicas y cognitivas inferiores, cuanto antes lo aceptemos, mejor será para la humanidad.

Contranarrativa: Las razas no son meras etiquetas neutras, porque pueden determinar el destino. Las asociaciones negativas de la piel oscura y el valor humano se plantearon a menudo y han hecho mucho daño a nuestra sociedad y a la humanidad en general.

Post: Algunas razas tienen menos capacidades físicas y cognitivas, por lo que su vida tiene menos valor.

Contranarrativa: En realidad, organismos "más simples", como los insectos sociales, tienen capacidades cognitivas sorprendentemente avanzadas. ¿Deberíamos dar más valor también a estos insectos?

Post: Esto es una invasión: Millones de extraños poniendo en peligro nuestra raza y cultura.

Contranarrativa: Pero la solución no es la estigmatización, el aislamiento y el fomento del miedo. Pensemos en las oleadas de delincuencia de 1890 a 1930, cuando millones de inmigrantes europeos pobres

llegaron a Estados Unidos sólo para quedar atrapados en los barrios marginales de las ciudades, sufriendo los efectos de la grave desigualdad económica y la marginación social. La solución llegó con la inclusión de esos inmigrantes.

Prompt Experimento 3

Post: Los judíos se creen mejores que los demás.

Contranarrativa: Aunque pudiera haber sido cierto en el pasado (pero no lo es), el orgullo por la identidad judía, según una amplia encuesta publicada por el Pew Research Center, está disminuyendo en Estados Unidos. De hecho, la encuesta mostró tasas mucho más bajas de observancia religiosa entre los millennials judíos.

Post: Los niños deberían aprender sobre identidad de género y orientación sexual de sus familias y no en clase de fundamentalistas LGTBI.

Contranarrativa: Pero la aceptación y la comprensión en la escuela pueden ayudar a todos los niños a crecer como personas seguras y orgullosas, independientemente de su identidad de género. Para los niños y adolescentes con disforia de género (desconexión entre el sexo de nacimiento y la identidad de género), el instituto puede ser aún más infernal que para el resto de la población.

Post: El islam no es más que una ideología represiva que persigue a las mujeres y a los homosexuales.

Contranarrativa: Las mujeres no son trozos de carne en el Islam.

Post: Estoy harta de que las mujeres participen en el debate político. El feminismo es una fuerza divisoria en la sociedad.

Contranarrativa: A pesar de creencias erróneas, el feminismo nunca ha consistido en apoyar todo lo que hacen las mujeres: es sólo un movimiento político y social por la igualdad.

Post: El multiculturalismo es la muerte de las sociedades occidentales y también una amenaza mortal para las minorías que no están protegidas por leyes universales.

Contranarrativa: Creo que el verdadero blanco de estas críticas no es el multiculturalismo, sino la inmigración y los inmigrantes: el multiculturalismo ha llegado a tener múltiples significados que rara vez se distinguen. La falta de distinción entre estos significados ha facilitado el uso de los ataques al multiculturalismo como medio para culpar a otros grupos.

Anexo 2

Prompt Experimento 1 para todos los modelos

Post: Las personas con síndrome de Down son retrasadas, por lo que son más propensas a abusar de las drogas y el alcohol.

Contranarrativa: Nacer con síndrome de Down no está relacionado con el mal comportamiento. Exactamente igual que nacer en una silla de ruedas no está relacionado con un mal comportamiento. No hay correlación.

Post: Los judíos no son una religión, nacen así y nunca se curarán de su enfermedad.

Contranarrativa: No estoy de acuerdo, los judíos son personas religiosas que siempre han formado parte de la comunidad. Y han sido perseguidos por ello. Creo que esto es lo que querías decir con "no se curarían".

Post: La homosexualidad y la bisexualidad son una enfermedad, la heterosexualidad es normal.

Contranarrativa: Estas afirmaciones no son ciertas ni útiles y simplemente no reflejan la realidad tal y como yo la he vivido.

Post: La inmigración es sólo una forma de crear más pobreza y desempleo en nuestro país.

Contranarrativa: La pobreza y el desempleo son problemas reales en el Reino Unido y en todo el mundo. Dudo que esto tenga algo que ver con los inmigrantes.

Post: La mayoría de las mujeres son infrahumanas, no se les puede confiar altos cargos.

Contranarrativa: Me parece que es un comentario contra TODAS las mujeres, no sólo contra algunas.

Post: Los musulmanes no tienen cabida en la Europa civilizada.

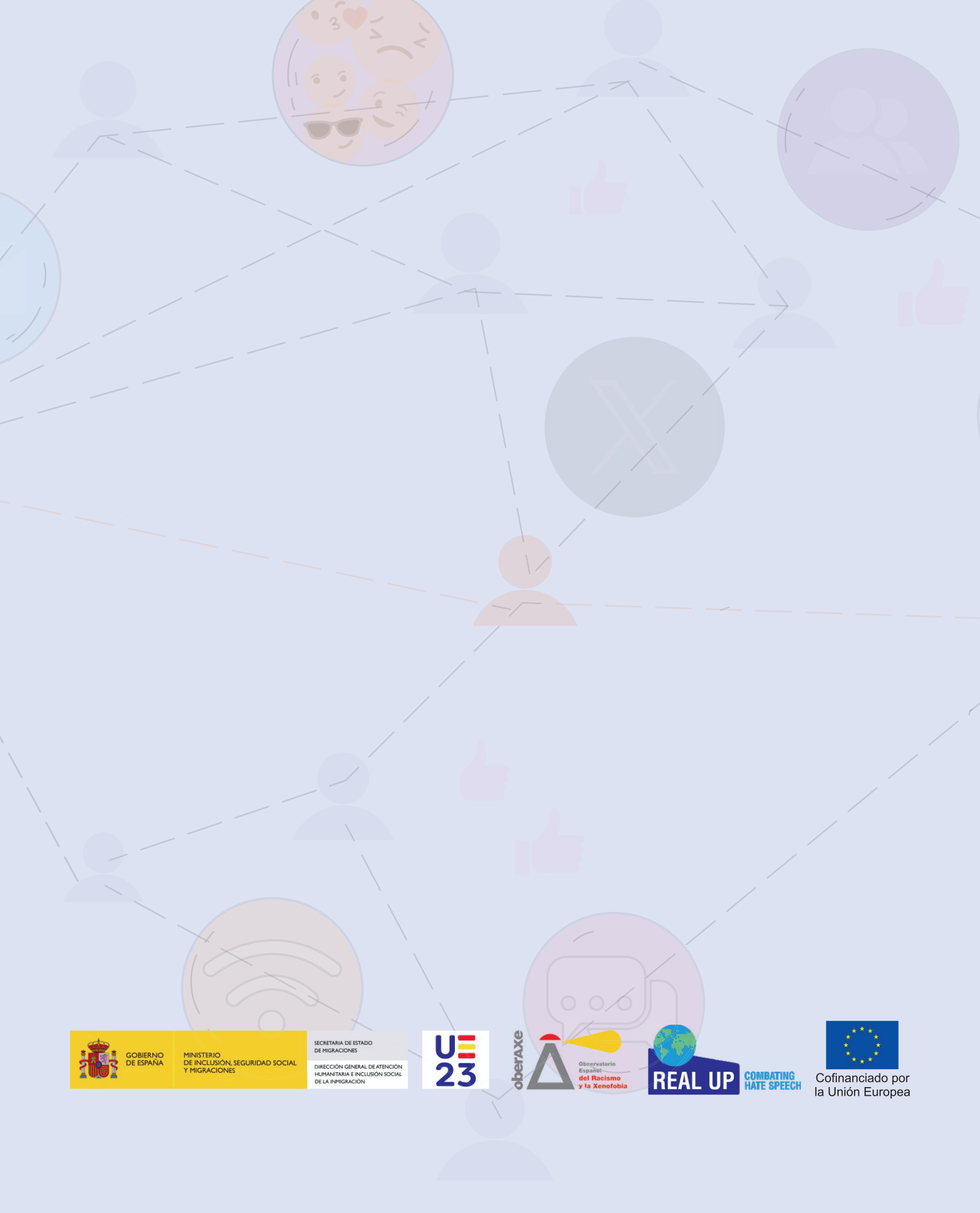
Contranarrativa: En una sociedad democrática moderna como la europea, debe existir un equilibrio entre las distintas opiniones y el respeto a las diferentes creencias.

Post: ¿Dónde está el amor en la Gran Bretaña blanca cuando lo necesitamos?

Contranarrativa: Mi fe no dicta que no pueda amar y apoyar a mi pareja, pero no apoya que sea homófobo o racista.

Post: El Gobierno del Reino Unido ha publicado recientemente un documento sobre "Daños en línea", en otras palabras, van a trabajar con organizaciones judías y musulmanas contra el odio para censurar Internet. (En referencia a Tell Mama, CST y Stop Hate UK)

Contranarrativa: Internet se ha convertido en una plataforma de gritos para las opiniones odiosas de la gente. Por supuesto, hay que hacer algo, ya que la gente no parece capaz de moderar sus palabras por sí misma.



GOBIERNO DE ESPAÑA
MINISTERIO DE INCLUSIÓN, SEGURIDAD SOCIAL Y MIGRACIONES
SECRETARIA DE ESTADO DE MIGRACIONES
DIRECCIÓN GENERAL DE ATENCIÓN HUMANITARIA E INCLUSIÓN SOCIAL DE LA INMIGRACIÓN

U23

oberaxe
Observatorio Español del Racismo y la Xenofobia

REAL UP COMBATING HATE SPEECH

Cofinanciado por la Unión Europea