LGTBfobia en redes sociales: Revisión sistemática de la detección y clasificación de discurso de odio a gran escala

LGBTphobia on social media: Systematic review of large-scale hate speech detection and classification

Marcos Barbosa

Universidad de Salamanca.
https://orcid.org/0000-0001-9684-3528
marcosbarbosa@usal.es (ESPAÑA)

Carlos Arcila

Universidad de Salamanca
http://orcid.org/0000-0002-2636-2849

Patricia Sánchez-Holgado

Universidad de Salamanca
https://orcid.org/0000-0002-6253-7087

Recibido: 12.04.2024
Aceptado: 04.02.2025

RESUMEN

El discurso de odio es una amenaza significativa para la diversidad social y los derechos humanos. La detección del discurso de odio contra la comunidad LGTB en las redes sociales es un área de investigación compleja que requiere un enfoque polifacético y adaptable. Este trabajo tiene como objetivo examinar las publicaciones científicas de enero de 2019 a febrero de 2024 para identificar los métodos computacionales más utilizados en la detección y clasificación de discurso de odio LGTBfóbico en redes sociales. Se examinan las metodologías de detección del discurso de odio LGTBfóbico en las redes sociales, abarcando aspectos teóricos y prácticos, analizando recursos recientes, discutiendo sobre las técnicas y métodos de clasificación e identificando los retos. El marco temporal de esta revisión de literatura se centra entre 1 de enero de 2019 al 29 de febrero

de 2024, incluyendo trabajos con un enfoque específico en el discurso de odio dirigido a la comunidad LGTB a través de las redes digitales, enfatizando los estudios que aplican métodos computacionales de detección. En el proceso de búsqueda hemos revisado documentos en diferentes bases de datos, como Scopus y WoS. Identificamos 28 trabajos relevantes en bases de datos académicas, observando que es esencial tener en cuenta la diversidad lingüística y cultural, así como la sensibilidad del tema a la hora de recopilar datos y desarrollar modelos de detección. Muchas de las investigaciones previas se han enfocado en el uso de métodos computacionales para detectar grandes cantidades de discurso de odio en redes y en crear estrategias para combatirlo. Además, se observa una tendencia hacia el uso de modelos de aprendizaje profundo, como BERT, y se enfatiza la necesidad de adaptar estos modelos a contextos específicos para mejorar su eficacia. Se mencionan también algunas tendencias emergentes centradas en el discurso esperanzado, las contranarrativas y la evaluación del nivel de toxicidad. Como líneas futuras, destaca la exploración de enfoques multilingües y la colaboración interdisciplinaria para abordar los desafíos en la detección y prevención del discurso de odio en las redes sociales.

PALABRAS CLAVE

Discurso de odio, LGTB; LGTBfobia, análisis computacional, redes sociales.

ABSTRACT

Hate speech is a significant threat to social diversity and human rights. Detecting anti-LGBT hate speech on social media is a complex area of research that requires a multifaceted and adaptable approach. This work aims to present a review of recent efforts in the investigation of hate speech towards the LGBT population on social networks and explore the use of computational techniques for the detection and classification of these speeches. Methodologies for detecting LGBTphobic hate speech on social networks are examined, covering theoretical and practical aspects, analyzing recent resources, discussing classification techniques and methods, and identifying challenges. The time frame of this literature review focuses between January 1, 2019 to February 29, 2024, including works with a specific focus on hate speech directed at the LGBT community through digital networks, emphasizing studies that apply computational detection methods. In the search process we have reviewed documents in different databases, such as Scopus and WoS. We identified 28 relevant works in academic databases, noting that it is essential to take into account linguistic and cultural diversity, as well as the sensitivity of the topic when collecting data and developing detection models. Much of the previous research has focused on using computational methods to detect large amounts of hate speech on networks

and creating strategies to combat it. Furthermore, there is a trend towards the use of deep learning models, such as BERT, and the need to adapt these models to specific contexts to improve their effectiveness is emphasized. Some emerging trends centered on hopeful discourse, counternarratives, and assessment of the level of toxicity are also mentioned. As future lines, the exploration of multilingual approaches and interdisciplinary collaboration to address the challenges in the detection and prevention of hate speech on social networks stands out.

KEY WORDS

Hate speech, LGBT, LGTBphobia, computational analysis, social networks.

1. INTRODUCCIÓN

En las últimas décadas, hemos sido testigos de un aumento significativo en la visibilidad y reconocimiento de la comunidad LGTB en la sociedad. Sin embargo, la igualdad de derechos y el fin del odio y la violencia hacia esos colectivos continúan siendo un desafío importante. Un ejemplo de eso es que actualmente sólo 48 de los 193 Estados Miembros de las Naciones Unidas poseen leyes que agravan la responsabilidad penal por delitos cometidos contra personas basados en su orientación sexual (ILGA World, 2020).

La discriminación persiste incluso en los países jurídicamente avanzados. Según la European Commission (2019), mientras que el 76% de los europeos apoya la igualdad de derechos para las personas LGTB y el 69% respalda el matrimonio entre personas del mismo sexo, más de la mitad (53%) reconoce una discriminación frecuente basada en la orientación sexual, la identidad de género o el hecho de ser intersexual. Las muestras públicas de afecto son más aceptadas para las parejas heterosexuales (78%) que para las parejas de lesbianas (53%) o gays (49%).

El camino hacia la igualdad implica procesos y marcos de derechos complejos (Browne, 2014). Esto subraya la urgente necesidad de hacer frente a la discriminación persistente que sufre la comunidad LGTB, que a menudo se traduce en una oleada de discursos de odio dirigidos a este grupo demográfico y que se manifiesta de diversas formas en Internet. Aunque el odio no se limita a los ámbitos virtuales, la esfera en línea facilita la conexión entre individuos que comparten prejuicios similares (Windisch et al,. 2022). La exposición frecuente a este tipo de contenidos normaliza el odio en línea (Çetınkaya et al., 2021; Soral et al., 2018).

Por esta razón, el estudio del odio en línea es una tendencia en ascenso en la academia (Paz et al., 2020; Nascimento et al., 2023). En este sentido, nuestro trabajo tiene como objetivo examinar las publicaciones científicas de enero de 2019 a febrero de 2024 para identificar los métodos computacionales más utilizados en la detección y clasificación de discurso de odio LGTBfóbico en redes sociales.

Nuestra contribución examina las metodologías de detección del discurso de odio LGTBfóbico en las redes sociales, abarcando aspectos teóricos y prácticos, analizamos recursos recientes, discutimos sobre las técnicas y métodos de clasificación e identificamos retos y tendencias emergentes. Mientras que estudios anteriores (Nascimento et al., 2023), han revisado la literatura sobre métodos computacionales para la detección del discurso de odio genérico, nuestro enfoque es específico sobre el discurso de odio relacionado con el colectivo LGTB. Además, mientras que Sánchez-Sánchez et al., (2024) han realizado un mapeo exhaustivo de la LGTBfobia en las redes sociales basándose en estudios científicos existentes, nuestro trabajo se distingue por profundizar en los aspectos computacionales de estos estudios.

2. DISCURSO DE ODIO Y LGTBFOBIA EN REDES SOCIALES

2.1. Definición y desafíos en el estudio del discurso de odio en redes sociales

Podemos considerar que el discurso de odio se distingue de otros tipos de ofensas por su principal factor motivador: el prejuicio. Este prejuicio puede dirigirse hacia un grupo en su conjunto o hacia individuos que pertenecen a ese grupo. Allport (1954) sugiere que las actitudes rechazantes pueden empezar con el antagonismo verbal (etapa que incluye el discurso de odio), evolucionado a otras formas de violencia, incluso el exterminio de un colectivo.

Aunque el discurso de odio no es un fenómeno reciente y se extiende más allá de las plataformas en línea, la presencia de las nuevas tecnologías ha provocado cambios en la manera de distribuir la información (Castillo, 2020) y ha acaparado cada vez más atención tanto en el discurso público como en las esferas académicas. Paz et al. (2020) examinó 1.112 artículos científicos poniendo de relieve que ha habido un notable aumento de la producción académica desde 2016 y que el derecho y la comunicación son los campos en los que se concentra la investigación sobre el discurso de odio. En el ámbito de la comunicación, los estudios se centran predominantemente en Internet y redes sociales. En otros ámbitos, como la informática y la ingeniería, ha habido una trayectoria similar, puesto que antes de 2014 el discurso de odio recibía escasa atención (Fortuna & Nunes, 2019).

El aumento y la complejidad del problema de la difusión de discursos de odio en la red ha originado debates en torno al concepto y metodologías de detección (Rosemary Rodríguez et al., 2023), su criminalización (Teijón Alcalá, 2022), desafíos a la sanción penal (Gómez Bellvis & Castro Toledo, 2022; Miró Linares & Gómez Bellvis, 2021), indicadores (Papcunová et al., 2021), perspectivas lingüísticas (Guillen-Nieto, 2023), efectos sobre la opinión pública (Zhang & Trifiro, 2022), la libertad de expresión (Martínez Valerio & Mayagoitía Soria, 2021; Ishita, 2019), efectos psicológicos (Saha et al., 2019; Ștefăniță & Buf, 2021), y visibilidad pública (Sponholz & Christofoletti, 2019).

Investigaciones anteriores sugieren que la exposición al odio en línea puede estar vinculada tanto a experimentar (Wachs et al., 2021) como a perpetrar ciberodio (Wachs et al., 2019). Algunos experimentos han probado contramedidas para contener la propagación del discurso de odio en la redes, como la suspensión de cuentas de individuos o grupos de las principales plataformas para la reducción del "odio organizado" (Thomas & Wahedi, 2023), el combate al odio en las redes por medio de estrategias de "contra-discurso" (Donzelli, 2021; Alsagheer et al., 2022) y de normas de ciudadanía solidaria (Kunst et al. 2021), la moderación de contenido por las plataformas (Gonçalves et al., 2021), y el ciberactivismo (Müller & Lopez-Sanchez, 2021).

En este sentido, son amplios los retos para detectar el discurso de odio, como el uso de metáforas (Lemmens et al., 2021) o del discurso implícito (ElSherief et al., 2021), incluso la propia clasificación tiene sesgos a superar (Badjatiya et al., 2019; Mozafari et al., 2020), ya que los grupos históricamente marginados, como las personas LGBT, tienen más probabilidades de percibir los comentarios como tóxicos, especialmente después de haber sufrido acoso (Kumar et al., 2021).

2.2. LGTBfobia en las plataformas de redes sociales

La investigación del discurso LGTBfóbico en Internet ha recibido especial atención especialmente en 2021 y 2022, destacando las investigaciones sobre las redes sociales e identificando a X como la principal para su difusión (Sánchez-Sánchez et al., 2024). Además, se observan debates generados por redes coordinadas que buscan provocar la confrontación online (Arce-García & Menéndez-Menéndez, 2023). La polarización tiene consecuencias en la LGTBfobia en contextos específicos de eventos sociales, como los explorados por Colussi et al., (2024) y Sánchez-Holgado et al., (2023) que examinan la percepción pública de la Ley Trans española y el uso del discurso de odio para criticar la legislación y a la comunidad transexual. Del mismo modo, Valerio (2022) analiza los comentarios durante la Semana del Orgullo, encontrando una mezcla de apoyo y crítica, con cierta incitación al odio.

En resumen, la revisión de los contenidos LGTBfóbicos en las redes sociales subraya la urgente necesidad de abordar esta forma de expresión del discurso de odio. Además, aunque la investigación ha experimentado un notable aumento en los últimos años, persisten retos significativos, especialmente en el ámbito computacional, donde la detección precisa y la mitigación del discurso LGTBfóbico en línea siguen siendo difíciles de lograr debido a su naturaleza.

3. METODOLOGÍA

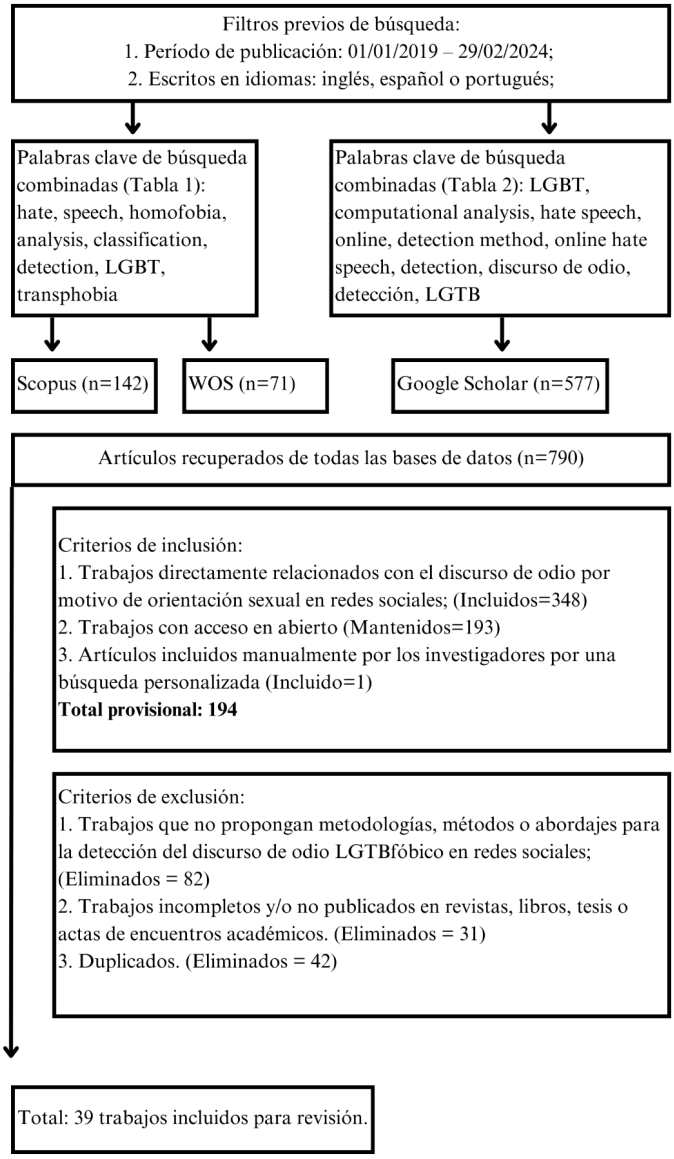
En esta sección, describimos la metodología de revisión sistemática empleada como base de nuestro estudio, que se delimita a partir de la siguiente pregunta

de investigación (PI): ¿Qué métodos computacionales son los más estudiados, en las publicaciones científicas, para detectar el discurso de odio LGTBfóbico en las plataformas sociales?

Nuestro objetivo es examinar los últimos avances en el análisis automatizado del discurso de odio dentro de la investigación académica. Hay un notable aumento en las publicaciones sobre métodos computacionales para detectar discurso de odio en 2019 (Nascimento et al., 2023). Basándonos en esta tendencia, nuestra investigación se centra en este marco temporal, examinando los trabajos publicados desde el 1 de enero de 2019 hasta el 29 de febrero de 2024, con un enfoque específico en el discurso de odio dirigido a la comunidad LGTB a través de las redes digitales, enfatizando los estudios que aplican métodos computacionales de detección. Dado que nuestro estudio se centra en España, priorizamos los idiomas predominantes en el contexto iberoamericano, como el español y el portugués, e incluimos el inglés debido a su relevancia global en la literatura científica, que también se aplica a esta región (Baladillo, 2020).

El esquema mostrado en la Figura 1 refleja el procedimiento seguido en este proceso de selección y búsqueda. Se ha utilizado una estrategia que ayuda a limitar sesgos y errores aleatorios, incluyendo la búsqueda exhaustiva de todos los artículos relevantes, criterios de selección reproducibles y explícitos, valoración del diseño y características de los estudios, y síntesis e interpretación de los resultados.

Figura 1. Procedimiento seguido en la revisión sistemática de la literatura



Fuente: elaboración propia

En el proceso de búsqueda hemos revisado documentos en diferentes bases de datos, como Scopus y WoS en primer lugar. Para ello, hemos utilizado combinaciones de palabras en inglés capaces de reunir los tres principales ejes de nuestro estudio: a) el discurso de odio hacia la población LGTB, b) las plataformas de redes sociales, c) los métodos de clasificación. En la Tabla 1 se presenta la estructura seguida en las combinaciones de palabras. Como resultado de esta búsqueda se ha obtenido un total de 213 documentos.

Tabla 1. Búsqueda en Scopus y WOS.

Palabras-clave	Base de datos	Número de resultados
hate AND speech AND LGBT AND classification	Scopus	8
	WOS	3
hate AND speech AND LGBT AND detection	Scopus	23
	WOS	6
hate AND speech AND LGBT AND analysis	Scopus	18
	WOS	14
hate AND speech AND homophobia AND classification	Scopus	9
	WOS	3
hate AND speech AND homophobia AND detection	Scopus	29
	WOS	10
hate AND speech AND homophobia AND analysis	Scopus	22
	WOS	21
hate AND speech AND transphobia AND classification	Scopus	0
	WOS	1
hate AND speech AND transphobia AND detection	Scopus	21
	WOS	5
hate AND speech AND transphobia AND analysis	Scopus	12
	WOS	8
Total		213

Aunque inicialmente se identificaron 213 documentos, se observó que muchos de ellos se repetían al cambiar las palabras clave. Tras eliminar los duplicados, se determinó que el número real de documentos únicos es 70.

En segundo lugar, se realizaron búsquedas complementarias en Google Scholar, que presenta resultados de búsqueda más amplios, por lo que hemos redefinido las palabras utilizadas para lograr registros filtrados y más específicos (ver Tabla 2).

Tabla 2. Búsqueda en Google Scholar.

Palabras-clave	Base de datos	Número de resultados
LGBT AND automated analysis AND hate speech AND online	Google Scholar	30
LGBT AND computational analysis AND hate speech AND online	Google Scholar	47
LGBT AND detection method AND hate speech AND online	Google Scholar	25
LGBT AND online hate speech AND detection	Google Scholar	344
LGBT AND discurso de odio AND detección	Google Scholar	75
LGTB AND discurso de odio AND detección	Google Scholar	56
Total		577

En base a la aplicación de los criterios de inclusión y de exclusión (Figura 1), hemos seleccionado un total de 39 artículos, divididos entre:

- Los que utilizan métodos basados en humanos para la detección y análisis de discurso de odio LGTBfóbico en redes sociales: Carvalho et al. (2023); Colussi et al. (2024); Sánchez-Holgado et al. (2023); Silva & Silva (2021); Valerio (2022).
- Los que utilizan métodos computacionales o híbridos de detección y análisis de discurso de ódio LGTBfóbico en redes sociales: Akhtar et al. (2019); Akhtar et al. (2020); Arcila et al. (2021); Babakov et al. (2022); Balaji & Chinmaya (2022); Banerjee & Nguyen (2023); Bel-Enguix et al. (2023); Chakravarthi (2023); Chakravarthi et al. (2021); Chakravarthi et al. (2022); Chakravarthi et al. (2023); ElSherief et al. (2021); Franza & Fišer (2019); Franza et al. (2022); Gevers et al. (2022); Kumaresan et al. (2023); Lemmens et al. (2021); Locatelli et al. (2023); Ljubesic et al. (2019); Molina-Villegas et al. (2023); Pelicon et al. (2021); Sharma et al. (2023)
- Los que indican líneas futuras para el estudio del discurso de odio LGTBfóbico en redes sociales: Alsagheer et al. (2022); Chakravarthi (2020); Chakravarthi (2022a); Chakravarthi (2022b); Dacon et al. (2022); Doganç & Markoy (2023); Fanton et al. (2021); García-Baena et al. (2023); Gupta et al. (2023); Khan & Hafiq (2023); Oliva et al. (2021); Tekiroglu et al. (2022).

4. RESULTADOS

4.1. Clasificación y detección de los discursos de odio en redes sociales: enfoque basado en humanos

Las técnicas de detección basadas en el ser humano se refieren a enfoques que implican el juicio, la aportación o la anotación realizados por humanos para el análisis de los datos de discurso de odio. Están normalmente basadas en la experiencia y la interpretación de los anotadores y desempeñan un papel fundamental a la hora de abordar los matices y la subjetividad dentro del discurso, permitiendo un enfoque más cualitativo del tema del discurso de odio y pueden lograr objetivos que todavía son complejos para los métodos computacionales en el análisis del lenguaje, como el discurso de odio encubierto (Carvalho et al., 2023) o implícito (ElSherief et al., 2021). Otra ventaja de la clasificación manual es la capacidad de profundizar en contextos específicos en los que el discurso puede tener características propias, como los matices geográficos y culturales (Molina-Villegas et al., 2023).

En la revisión realizada se han identificado trabajos que utilizan la clasificación manual del discurso de odio hacia la población LGTB en plataformas de redes sociales desde distintos contextos geográficos, como Brasil, Portugal o España. Aunque todos los estudios utilizan métodos manuales para el análisis del discurso de odio, las técnicas y enfoques específicos pueden variar. Los trabajos de Valerio (2022), Sánchez-Holgado et al., (2023) y Colussi et al., (2024), por ejemplo, buscan relacionar el análisis del odio con eventos sociales relevantes, mientras que Carvalho et al., (2023) y Silva & Silva (2021) no eligen un evento específico.

Si bien todos los estudios tienen como objetivo comprender el discurso de odio dirigido a la comunidad LGTB, pueden tener objetivos de investigación adicionales. Carvalho et al., (2023) categoriza el discurso de odio hacia varios colectivos, no limitándose al grupo LGTB, mientras que los trabajos de Sánchez-Holgado et al. (2023) y Colussi et al., (2024) examinan la percepción pública sobre un tema relacionado, como es la "Ley Trans" en España.

Aunque estos trabajos contribuyen a la comprensión de la dinámica del discurso de odio LGTBfóbico en plataformas de redes sociales, su enfoque presenta limitaciones. La más notable es el tamaño de la muestra que se puede analizar manualmente, puesto que trabajan con comentarios o publicaciones recopiladas de redes sociales, que no superan los 6.000 registros (Valerio, 2022), mientras que el análisis automático permite el entrenamiento para la clasificación de muestras considerables de datos por medio de la replicación. Además, la falta de generalización en la clasificación manual puede resultar difícil para analizar patrones y características del discurso de odio. De este modo, la clasificación manual puede ser viable como una fase inicial que precede al análisis automatizado.

4.2. Recopilación, anotación y caracterización de datos en el análisis automatizado

Los estudios revisados recolectan datos de redes sociales, siendo las más habituales X, Facebook, o YouTube, pero también de foros de discusión populares en contextos locales. Para recopilar la variedad de información disponible, los trabajos recurren a herramientas como Youtube Comment Scrapper Tool, que extrae comentarios de YouTube, y X Developer API, que proporciona acceso a datos específicos de X. Para la identificación de registros de discurso de odio los autores emplean estrategias como palabras ofensivas relacionadas con el tema o búsqueda de eventos desencadenantes de posibles olas de odio. Algunos trabajos también realizan un proceso de preprocesamiento de datos en el análisis de texto, con herramientas que se encargan de tareas como eliminar la puntuación, remover las URLs, detectar idiomas, anonimizar usuarios, y normalizar el texto, entre otras.

Una vez conseguido el conjunto de datos final, se suele separar una parte de esos datos para realizar una anotación manual, que consiste en asignar etiquetas a los textos. Esta tarea puede incluir métodos manuales con voluntarios, expertos o crowdsourcing. En la literatura analizada se ejecuta por colaboradores en plataformas como Amazon Mechanical Turk o similares (Babakov et al., 2022; Franza et al., 2022). Este enfoque proporciona una valiosa perspectiva colectiva sobre diferentes tipos de discurso de odio dirigidos a grupos sociales. Por otro lado, uno de los inconvenientes es que puede ser costoso (Nascimento et al., 2023) sobre todo si se desea mantener la calidad de la anotación, además de verse influenciado por aspectos culturales y prejuicios personales (Chakravarthi et al., 2021).

Diversos estudios han propuesto metodologías para garantizar la calidad de los procesos de anotación en la detección del discurso de odio, que suelen combinar la experiencia de especialistas en la materia con la de anotadores entrenados (Akhtar et al., 2019; Arcila et al., 2021; Molina-Villegas et al., 2023). En el contexto de la anotación manual centrada en el discurso de odio LGTB se han realizado experimentos con personas de la comunidad LGTB, aliados o con afiliaciones relevantes al tema (Akhtar et al., 2020; Banerjee & Nguyen, 2023; Kumersan et al., 2023; Locatelli & Damo, 2023). El objetivo es mejorar el acuerdo entre anotadores y aproximar las perspectivas del odio desde el punto de vista de las víctimas o de las personas empáticas. Sin embargo, Carvalho et al. (2023) observaron discrepancias en este enfoque. Su estudio, en el que participó un equipo de cinco anotadores de grupos marginados y no marginados, descubrió que los anotadores de comunidades marginadas mostraban un mayor desacuerdo entre ellos en comparación con los de grupos no marginados.

Hemos identificado una amplia variedad de conjuntos de datos que abordan el odio dirigido hacia la comunidad LGTB (Tabla 3). Algunos se centran exclusivamente en clasificar el discurso de odio LGTB en redes sociales, mientras que otros incluyen el odio hacia esta comunidad como una categoría entre otros tipos de discursos de odio. Además, los idiomas utilizados son diversos, como

el alemán, español, esloveno, francés, hindi, holandés, inglés, italiano, noruego, portugués, idiomas dravídicos (como kannada, tamil y malayo), además de contextos multilingües.

Tabla 3. Conjuntos de datos para el análisis del discurso de odio LGTB en redes sociales.

Tipo de Enfoque	Autores
Exclusivamente sobre discurso de odio LGTB	Arcila et al. (2021), Chakravarthi et al. (2021), Balaji & Chinmaya (2022), Banerjee & Nguyen (2023), Bel-Enguix et al. (2023), Locatelli & Damo (2023), Sharma et al. (2023)
Discurso de odio LGTB como una categoría	Akhtar et al. (2019), Ljubesic et al. (2019), Lemmens et al. (2021), Babakov et al. (2022), Carvalho et al. (2023)

Las categorías de etiquetas varían desde una clasificación binaria simple de presencia o ausencia de odio (Arcila et al., 2021; Banerjee & Nguyen, 2023), o de "positivo", "negativo" y "neutral" (Locatelli & Damo, 2023); hasta clasificaciones en árbol con varios niveles para cada texto (Chakravarthi et al., 2021), pasando por clasificaciones detalladas que distinguen el odio entre los distintos grupos que componen el colectivo LGTB (Bel-Enguix et al., 2023; Sharma et al., 2023).

4.3. Extracción y clasificación de características

Otra tarea importante en el proceso de clasificación automática es el uso de técnicas para la extracción de características (conocidas habitualmente como *features*), que son las propiedades, atributos o elementos extraídos de los datos que representan información relevante para el análisis. En el contexto de texto o contenido digital, las features sirven como entradas para algoritmos de aprendizaje automático o herramientas de procesamiento del lenguaje natural, por lo que estas técnicas se utilizan para convertir el texto en representaciones numéricas que los algoritmos de aprendizaje automático pueden entender y procesar.

Uno de los más comunes es la "frecuencia de término-frecuencia inversa de documento" (TF-IDF), técnica de ponderación que evalúa la importancia de una palabra en un documento. Otros métodos son por ejemplo el "Bag of words" («saco de palabras») que contiene todas las palabras del texto, y se cuenta cuántas veces aparece cada palabra; el análisis de sentimiento, que utiliza el procesamiento del lenguaje natural para determinar la actitud o el tono emocional de un texto, clasificándose generalmente como positivo, negativo o neutro; el "*word embedding*", que representa las palabras como vectores de números reales, donde las relaciones semánticas entre palabras se conservan en el espacio vectorial;

la "fastText", una biblioteca de aprendizaje automático para el procesamiento y clasificación de texto.

Tabla 4. Características (features) más comunes en la clasificación del discurso de odio LGTB»

Feature	Descripción	Referencias
TF-IDF	Importancia de las palabras en un texto basada en su frecuencia ponderada.	Akhtar et al. (2019); Chakravarthi et al. (2021); Banerjee & Nguyen (2023); Chakravarthi (2023); García-Baena et al. (2023); Kumaresan et al. (2023)
Bag of Words (BoW)	Representa el texto como un conteo de las palabras sin considerar el orden.	Akhtar et al. (2019); Arcila et al. (2021)
Análisis de Sentimiento	Clasifica el tono emocional del texto como positivo, negativo o neutro.	Locatelli & Damo (2023)
Word Embedding	Representa palabras como vectores numéricos, preservando relaciones semánticas.	García-Baena et al. (2023); Sharma et al. (2023)
fastText	Genera vectores de palabras eficientes para la clasificación de texto.	Chakravarthi et al. (2021); Babakov et al. (2022); Balaji & Chinmaya (2022); Chakravarthi (2023)

Algunos trabajos emplean varios métodos en un mismo estudio, como Sharma et al., (2023) que además de algunos de los mencionados, utiliza: "tokenización", que divide el texto en unidades más pequeñas; "convolutional layer", que actúa como una capa de una red neuronal convolucional (CNN) utilizada para extraer características de datos multidimensionales, como imágenes o secuencias; y "padding", que agrega relleno a los datos de entrada para que todos tengan la misma longitud.

Importante mencionar, además, el uso de BERT, un modelo de lenguaje preentrenado con una arquitectura de *Transformers,* desarrollado por Google que logra un rendimiento de vanguardia en una variedad de tareas de procesamiento de lenguaje natural, y algunas de sus variantes como RoBERTa (Chakravarthi et al., 2022; Chakravarthi et al., 2023; Kumaresan et al. 2023), y mBERT (Chakravarthi et al., 2022; Kumaresan et al., 2023; Sharma et al., 2023).

Existe un amplio espectro de enfoques para convertir el texto en representaciones numéricas comprensibles por los algoritmos de aprendizaje automático. La combinación de múltiples métodos en un mismo estudio puede proporcionar

una comprensión más completa y precisa de los datos y mejorar el rendimiento de los modelos de clasificación automática.

En la literatura revisada hemos identificado trabajos que utilizan una serie de clasificadores. Esto contempla los análisis basados en léxico (por ejemplo análisis de sentimiento), los métodos de aprendizaje automático (Naïve Bayes, Logistic Regression, Random Forests, Decision Trees, Support Vector Machines etc.), los métodos de aprendizaje profundo que implican el uso de redes neuronales (Convolutional Neural Networks, Recurrent Neural Networks, Long Short-Term Memory), las técnicas que se centran en la comprensión y el procesamiento del lenguaje natural humano (Count Vectorizer, BERT, TF-IDF Vectorizer), y los estudios que combinan múltiples enfoques de inteligencia artificial para mejorar la precisión y robustez del sistema.

4.4. Exploración comparativa de estudios con análisis automatizado del discurso de odio LGTB

Algunos trabajos en la literatura revisada abordan el análisis del discurso de odio LGTB de manera más amplia, al asociarlo con otros grupos marginados. El estudio de Akhtar et al., (2019) ha empleado dos conjuntos de datos (inglés e italiano) para categorizar el discurso de odio en X relacionado con racismo, sexismo y homofobia, utilizando SVM y técnicas adicionales, pero el modelo mostró un bajo rendimiento en la detección de homofobia (F1-Score 0,33). Ljubesic et al., (2019) extrae comentarios de Facebook sobre migrantes y LGTB. Empleando anotaciones manuales y clasificadores SVM, se forma el conjunto de datos FRENK, compuesto prevalentemente por contenido ofensivo y violento. Estudios como Pelicon et al. (2021) utilizan este corpus para explorar la detección en múltiples idiomas, destacando el rendimiento de los modelos BERT. Franza y Fišer (2019) analizan la dependencia contextual del discurso de odio hacia migrantes y personas LGTB y Franza et al. (2022) destacan una mayor intensidad emocional en comentarios dirigidos a grupos marginados también usando el FRENK.

Los estudios de Lemmens et al., (2021) y Gevers et al., (2022) amplían el análisis del discurso tóxico en el corpus LiLaH. Lemmens et al., (2021) utilizan modelos SVM, BERTje y RobBERT para clasificar el tipo y objetivo en comentarios sobre migrantes y LGTB en holandés. Gevers et al., (2022) analizan distinciones entre comentarios tóxicos y no tóxicos. Los comentarios tóxicos son más largos, menos diversos léxicamente y presentan más desviaciones lingüísticas. Los comentarios en esloveno, inglés y holandés muestran patrones en longitud y uso de emoji, sugiriendo que la longitud es una característica clave para la detección automática de toxicidad.

Babakov et al. (2022) utilizaron ruBERT y otros métodos computacionales para clasificar mensajes en foros rusos. Sin embargo, no desarrolló un modelo específico para detectar LGTBfobia. ElSherief et al. (2021) introduce una ta-

xonomía teóricamente justificada del discurso de odio implícito junto con un corpus etiquetado detallado, pero tampoco profundiza en la LGTBfobia.

En nuestro trabajo nos centramos en los retos de la de la investigación sobre la detección del discurso de odio LGTBfóbico. En la literatura ya existen algunos esfuerzos hacia la construcción de conjuntos de datos y el entrenamiento de algoritmos para ello. Aunque haya experimentos en diversos idiomas, la mayoría prioriza el inglés. Además, las principales redes sociales exploradas son X y YouTube, esta segunda sobre todo en India, donde es señalada como la más popular.

Algunos trabajos utilizan el análisis de sentimiento para clasificar los textos como LGTBfóbicos. Tratase de una técnica del procesamiento del lenguaje natural útil para determinar la actitud o el tono emocional (positivo, negativo o neutro) en enormes volúmenes de texto. Normalmente, se asocia el discurso de odio al sentimiento negativo (Schmidt & Wiegand 2017). Nascimento et al. (2023) destaca como principal limitación del método la necesidad de utilizar otras técnicas para mejorar los resultados. Locatelli & Damo (2023) analizan la homotransfobia en tuits sobre temas LGTB en siete idiomas (inglés, español, portugués, italiano, alemán, francés y noruego) utilizando Contextualized Topic Modeling (CTM) para extraer los temas principales y un clasificador pre-entrenado de análisis de sentimiento para evaluar las actitudes expresadas en los tuits.

Balaji & Chinmaya (2022) realizan un análisis de sentimiento e identificación de lenguaje ofensivo homofóbico y transfóbico para más de 60 mil comentarios en YouTube, en idiomas dravídicos mezclados con inglés. Los conjuntos de entrenamiento contienen datos etiquetados como contenido no anti-LGTB, homofóbico y transfóbico. El estudio utiliza dos métodos de clasificación: el Multimodal Pre-trained Network (MP-Net), que clasifica sentimientos y detecta homofobia/transfobia en inglés y tamil-inglés, y una combinación de fastText y LightGBM empleada para detectar homofobia/transfobia en tamil, kannada y malayo. La evaluación del análisis de sentimientos tiene puntajes F1 macro de 0,19, 0,3 y 0,2 para tamil, kannada y malayo respectivamente. En cuanto a la detección de homofobia-transfobia, los puntajes F1 macro son 0,234, 0,493, 0,942 y 0,316 para tamil, inglés, malayo y tamil-inglés respectivamente.

También hemos identificado estudios que han empleado métodos más avanzados de procesamiento del lenguaje natural y aprendizaje automático para abordar la detección del discurso de odio dirigido a la comunidad LGTB. En este campo de estudio, destacan trabajos en las lenguas de India (Chakravarthi 2023; Chakravarthi et al. 2021; Chakravarthi et al. 2022; Chakravarthi et al. 2023; Kumaresan et al. 2023; Sharma et al. 2023). Estudios preliminares centrados en la identificación del lenguaje ofensivo en tamil condujeron a la creación de conjuntos de datos de plataformas de medios sociales como X y YouTube.

El trabajo de Chakravarthi et al. (2021) presenta contribuciones al definir una taxonomía para la homofobia y la transfobia, explorando el discurso de la esperanza y el contra-discurso, e introduciendo conjuntos de datos para textos en inglés, tamil y textos mezclados en código tamil-inglés, abordando lagunas en los recursos para lenguas con pocos recursos como el tamil. Utilizándose datos de

YouTube, los autores obtuvieron un total de 4.946 comentarios en inglés, 4.161 comentarios en tamil y 6.034 comentarios con código mixto tamil-inglés etiquetados manualmente por anotadores voluntarios aliados del movimiento LGTB. El alfa de Krippendorf para evaluar el acuerdo entre los anotadores obtenido es de 0,67, 0,76 y 0,54 para el inglés, el tamil y el tamil-inglés, respectivamente. El primer nivel de clasificación incluye tres: contenido homofóbico, transfóbico o no LGTBfóbico. El segundo nivel de clasificación incluía categorías como contenido despectivo homofóbico/transfóbico, contenido amenazante homofóbico/transfóbico, y tres categorías de discurso no LGTBfóbico: contra-discurso, discurso esperanzado, o ninguno. Se aplicó una estrategia de muestreo estratificado para dividir los datos en grupos, asegurando que cada grupo tuviera el mismo porcentaje de etiquetas. Los datos se prepararon para conjuntos de etiquetas de 3, 5 y 7 clases. Se utilizaron varios clasificadores para el análisis, incluyendo LR, NB, RF, SVM, DT, BiLSTM y mBERT (Multilingual BERT). Además, se emplearon diversas características, como TF-IDF (tri-gram), CountVectorizer (tri-gram), FastText y BERT. Los mejores resultados se obtuvieron con el conjunto de datos de 3 clases, donde solo se incluyen etiquetas de contenido homofóbico, transfóbico y no anti-LGTB+, que se obtuvieron utilizando RF+BERT para inglés (F1= 0,926) y tamil-inglés (F1= 0,852), mientras que para tamil el mejor resultado se obtuvo con RF+FastText (F1=0,912). Basándose en los resultados de sus experimentos con los tres idiomas y las tres configuraciones diferentes de etiquetas de clase, se descubrió que una combinación de aprendizaje profundo y aprendizaje automático funcionaba significativamente mejor que el aprendizaje profundo o el aprendizaje automático por separado.

En estudio subsecuente, Chakravarthi et al. (2022) aborda la mejora del rendimiento de los modelos utilizando el mismo conjunto de datos de Chakravarthi et al. (2021), mediante la pseudo-etiquetación y la transliteración, utilizando modelos de lenguaje pre-entrenados como mBERT, MuRIL, IndicBERT y XLM-R. Los resultados muestran un mejor desempeño en el conjunto de datos ampliado con pseudo-etiquetas, obteniendo mejores métricas de precisión, exhaustividad y puntuación F1. Entre los modelos evaluados, mBERT muestra un rendimiento generalmente sólido en todas las tareas de clasificación, con puntajes consistentes en los conjuntos de etiquetas de 3, 5 y 7 clases: 0,940, 0,902 y 0,898, respectivamente. En otro estudio utilizando el mismo conjunto de datos de Chakravarthi et al. (2021), Chakravarthi (2023) discute sobre una tarea compartida llevada a cabo en el taller LTEDI-ACL 2022 para mejorar la investigación en la detección de homofobia y transfobia. Como resultado, se obtuvieron 10 sistemas para el idioma tamil, 13 sistemas para el idioma inglés y 11 sistemas para la combinación de los idiomas tamil e inglés. Los mejores resultados alcanzados para tamil y tamil-inglés utilizaron arquitecturas de aprendizaje automático y profundo (F1= 0,940 y 0,89 respectivamente). Para inglés se empleó el ajuste fino del modelo de lenguaje pre-entrenado Roberta-base (F1= 0,920).

Kumaresan et al. (2023) investiga la detección de homofobia y transfobia en el contexto de India agregando al conjunto de datos 5.193 comentarios en malayo y 3.203 comentarios en hindi también publicados en YouTube. El método de

anotación del nuevo conjunto de datos se basa en el crowdsourcing, utilizando anotadores entrenados, quienes poseen educación de posgrado y se identifican como miembros de la comunidad LGTB o aliados. Igual que en el estudio de Chakravarthi et al. (2021) la clasificación se realiza en varios niveles: el primero tiene tres etiquetas; el segundo, cinco; y el tercero, siete. El grado de acuerdo es de 0,72, que se considera relativamente alto y sugiere un grado razonable de acuerdo entre los anotadores. Las características empleadas para entrenar y evaluar los modelos de clasificación fueron TF-IDF, BERT *embeddings*, fastText, RoBERTa Base, RoBERTa Large, mBERT uncased, XLM-RoBERTa Small, XLM-RoBERTa Large. Además, se probaron los clasificadores LR, NB, DT, RF y SVM. El estudio tiene un enfoque interlingüístico y utilizó los mejores modelos de cada clase para evaluar si los modelos pueden predecir con precisión los comentarios cuando se aplican a distintas lenguas. En general, los resultados indican que, si bien algunos modelos presentan un buen rendimiento en sus idiomas, su eficacia disminuye cuando se aplican a otros idiomas. Esto puede atribuirse a las variaciones lingüísticas y a las dificultades para captar patrones y matices específicos de cada lengua.

En la tarea compartida llevada a cabo durante el LT-EDI 2023, Chakravarthi et al. (2023) presenta una extensión al trabajo, incluyendo textos en español producidos en X. En la primera tarea los conjuntos de etiquetas poseían tres clases, y en la segunda siete. Se obtuvieron 35 modelos para la primera tarea distribuidos entre inglés (10), tamil (7), español (4), malayo (7) e hindi (7) y 21 modelos en la segunda tarea, entre inglés (8), tamil (7) y malayo (6). Se realizaron diversas clasificaciones utilizando una variedad de técnicas, entre las cuales se destacan BERT, TF-IDF con SVM, *weight-space ensembling*, XLM-RoBERTa, afinado de XLM-RoBERTa, mBERT con re-muestreo y GPT2. Los mejores resultados fueron obtenidos por BERT para la clasificación en tres clases en inglés e hindi (F1= 0,969 y 0,979 respectivamente), mientras que, para malayo, tamil y español, el *weight-space ensembling*, técnica que combina modelos multilingües y afinados mezclando sus pesos mediante interpolación lineal, mostró un rendimiento superior (F1= 0,997, 0,888 y 0,949 respectivamente) en el análisis de textos en varias lenguas, especialmente en las de recursos limitados. Por otro lado, para la clasificación de siete clases, BERT fue el modelo más efectivo para el inglés (F1= 0,822), mientras que XLM-RoBERTa y afinado de XLM-RoBERTa destacaron para malayo (F1= 0,884) y tamil (F1= 0,865) respectivamente.

Todavía en el contexto de lenguas dravídicas, Sharma et al. (2023) utilizan el dataset DravidianLangTech, que contiene mensajes malayo y tamil. Para la tarea de clasificación, los investigadores utilizan cuatro métodos de aprendizaje profundo: CNN (+GloVE), LSTM (+GloVe), mBERT e IndicBERT. Los modelos IndicBERT presentaron mejores performances, tanto para el análisis en malayo como para tamil. El promedio del F1-score fue de 0,86 para los discursos en malayo y 0,77 para tamil. Dado que F1 para los mensajes no anti-LGTB+ es significativamente más alto que para los mensajes homofóbicos y transfóbicos, podemos concluir que el modelo tiene un mejor desempeño en la clasificación de mensajes neutrales en comparación con los mensajes de odio.

Por otro lado, en el contexto de análisis del discurso de odio LGTBfóbico únicamente en inglés, el estudio de Banerjee & Nguyen (2023) se centró en la clasificación de Queerfobia en comentarios publicados en YouTube. Se utilizaron datos descargados mediante Google AppScript para construir el conjunto de datos, de 10.000 mensajes, con la participación de 3 anotadores voluntarios LGTB. Para resolver los desacuerdos entre los anotadores, se utilizó un método de dos-tercios. Después de la clasificación, se probaron 16 modelos en estos datos, utilizando cuatro clasificadores (DT, RF, SVM y GB) con cuatro técnicas de extracción de características (GloVe, Word2Vec, TF-IDF y CountVectorizer), de las cuales TF-IDF y CountVectorizer obtuvieron la mejor evaluación. Una de las explicaciones puede ser porque GloVe y Word2Vec, que tuvieron un rendimiento peor, son algoritmos de *embedding* (incrustaciones) de palabras que representan palabras como vectores densos en un espacio de alta dimensionalidad, donde las palabras relacionadas están ubicadas más cerca. Por otro lado, CountVectorizer y TF-IDF son algoritmos más simples que utilizan la frecuencia de palabras para extraer características del texto. No tienen en cuenta la semántica y tratan cada palabra individualmente. El rendimiento superior de CountVectorizer y TF-IDF sobre GloVe y Word2Vec puede indicar que las relaciones semánticas entre las palabras no son muy importantes para la tarea de reconocimiento de la Queerfobia, o que el conjunto de datos no es lo suficientemente grande para que los *embeddings* de palabras se entrenen de manera efectiva. Además, los dos mejores modelos son los que usan el GB como clasificador, llegando ambos a una puntuación F1 de 0,854.

Finalmente, hemos identificado dos estudios que detectan el odio hacia poblaciones LGTB+ en español. El primero, desarrollado por Bel-Enguix et al. (2023), realiza un estudio con una muestra de 11.000 mensajes de X en México, con etiquetado manual. En un primer nivel, identifica si un mensaje es LGTB+fóbico, y en un segundo nivel a qué sujetos del colectivo LGTB se dirigen los mensajes (1.339 mensajes). Estos datos son usados en el Iberlef 2023 para la tarea colectiva denominada «HOMO-MEX», cuyo objetivo es fomentar el desarrollo de sistemas de NLP que puedan detectar y clasificar contenido fóbico hacia la comunidad LGTB. Ocho equipos participan del experimento que se ha dividido en dos tareas. La primera tarea, detección de LGTB-fobia, logró resultados moderadamente satisfactorios con una clasificación de tres clases. Sin embargo, la segunda tarea, que implicaba la identificación de múltiples etiquetas, obtuvo puntajes de rendimiento más bajos. En la primera tarea, el mejor resultado se logra con un enfoque basado en el modelo RoBERTuito, que obtuvo una puntuación de F1 macro de 0,843.

El segundo estudio fue conducido por Arcila et al. (2021) con un conjunto de 21.000 registros de X en español. Tras el preprocesamiento de los datos, el 33,7% fueron etiquetados como no odio, el 11,6% como odio motivado por género o sexualidad y el 54,5% fueron descartados. En el proceso de anotación participaron nueve anotadores manuales. Para la tarea de clasificación, se emplearon seis modelos individuales generados por algoritmos de aprendizaje superficial, cada uno basado en técnicas tradicionales de clasificación. Estos modelos utili-

zan representaciones de «Bag of Words» para el texto los algoritmos NB, MNB, Bernoulli NB, LR, clasificador SGD y SVC lineal. Posteriormente, se crea un séptimo clasificador que combina los votos de los seis modelos anteriores. Además, se desarrolla un octavo modelo utilizando *embeddings* y técnicas de modelado profundo, específicamente RNN con capas de unidades recurrentes GRU y una capa de salida densa con activación sigmoide. La evaluación mostró que el algoritmo de aprendizaje profundo funcionó significativamente mejor que los algoritmos de modelado superficial, a pesar de las métricas F-Score más bajas, lo que indica que hay margen de mejora en la precisión y la recuperación para la detección del discurso del odio en X. El modelo LR obtuvo el mejor rendimiento de los algoritmos superficiales (F1= 0,86).

La Tabla 5 resume los mejores modelos de cada estudio sobre la detección de discursos de odio en las redes sociales abarcando una variedad de enfoques, que van desde algoritmos tradicionales de aprendizaje automático hasta sofisticadas arquitecturas de aprendizaje profundo.

Tabla 5. Mejores modelos entre los trabajos revisados.

Referencia	Origen	Idioma	Método de clasificación	F-Score
Arcila et al. (2021)	X	Español	*Embeddings* + RNN	0,65
Chakravarthi et al. (2021)	YouTube	Inglés	RF + BERT	0,926
Balaji & Chinmaya (2022)	YouTube	Malayo	FastText+LightGBM	0,942
Chakravarthi et al. (2022)	YouTube	Múltiples	mBERT	0,94
Banerjee & Nguyen (2023)	YouTube	Inglés	GB+TFIDF/ CountVectorizer	0,854
Bel-Enguix et al. (2023)	X	Español	RoBERTuito	0,843
Chakravarthi (2023)	YouTube	Tamil	*Shallow + Deep learning*	0,94
Chakravarthi et al. (2023)	YouTube	Malayo	*weight-space ensembling*	0,997
Kumaresan et al. (2023)	YouTube	Malayo	RF+FastText	0,94
Sharma et al. (2023)	YouTube	Malayo	IndicBERT	0,86

4.5. Tendencias emergentes en la investigación del discurso de odio LGTB

Hemos identificado algunas tendencias en este campo de investigación. Una de ellas sería el *Hope Speech* (discurso esperanzado). Esta categoría engloba los trabajos de Chakravarthi (2020), Chakravarthi (2022a), Chakravarthi (2022b) y García-Baena et al. (2023). Estos trabajos utilizaron anotadores para clasificar comentarios en las categorías de «Hope» (esperanza) y «Not Hope» (no esperanza, que por veces contendría discurso de odio). A continuación, en la Tabla 6, se presenta un resumen de los estudios mencionados.

Tabla 6. Resumen de Estudios sobre Hope Speech.

Estudio	Conjunto de Datos	Modelos Utilizados	Resultados
Chakravarthi (2020)	Hope EDI (Inglés, Tamil, Malayo)	SVM, MNB, KNN, DT y LR	Bajo rendimiento de los modelos
Chakravarthi (2022a)	Hope EDI (Inglés, Tamil, Malayo)	SVM, MNB, KNN, DT, LR y CNN	Mejor modelo: CNN F1 macro: 0,75 (Inglés), 0,62 (Tamil), 0,67 (Malayo)
Chakravarthi (2022b)	Hope EDI (Inglés, Tamil, Malayo)	SVM, MNB, KNN, DT, LR y RoBERTa	Mejor modelo: RoBERTa F1 macro: 0,52 (Inglés), 0,61 (Tamil), 0,81 (Malayo)
García-Baena et al. (2023)	SpanishHopeEDI (Español)	TFIDF+SVM, TFIDF+MNB, TFIDF+LR, BTO+SVM, BTO+ MNB, BTO+LR, TF+ SVM, TF+MNB, TF+LR, NG+MLP, LF+MLP, SE+ MLP, WE+MLP, WE+CNN, WE+BiLSTM, BE+MLP, BF+MLP	Mejor resultado: BF+MLP: F1=0,85

La segunda tendencia es la experimentación de contranarrativas. Alsagheer et al. (2022) experimenta la generación de discursos que contrarresten el odio en redes sociales y destaca la importancia de entender su impacto y las dinámicas que anteceden y suceden la intervención. Gupta et al. (2023) desarrollan un conjunto de datos de contranarrativas etiquetadas según cinco intenciones: informativa, denuncia, pregunta, positiva y humorística.

Fanton et al. (2021) han desarrollado en conjunto de datos MultiCONAN, que incluye prototipos de discursos de odio contra la comunidad LGTB y otros grupos, empleando GPT-2 para generar contranarrativas. Mediante un enfoque de bucle autor-revisor, validan y editan las respuestas para mejorar la calidad del conjunto. Tekiroglu et al. (2022) y Doganç & Markoy (2023) se basan en el trabajo de Fanton et al. (2021). Los primeros usan modelos lingüísticos preentrenados para contranarrativas automáticas, mientras que los segundos destacan la eficacia de MultiCONAN y subrayan la importancia de GPT-3.5 para la elaboración de contranarrativas personalizadas.

Por último, la tercera tendencia interesante son los estudios de análisis del lenguaje tóxico para evaluar el nivel de comportamiento dañino en línea utilizando herramientas como la API Perspective[1], que utiliza el aprendizaje automático para detectar la toxicidad en el texto, proporcionando puntuaciones de probabilidad, desde 0 (no tóxico) a 1 (altamente tóxico). Oliva et al. (2021) señala que Perspective presenta limitaciones para interpretar el lenguaje de las *drag queens* de EE.UU., porque no comprende el contexto social y los matices del discurso, lo que lleva a errores al clasificar palabras que reciben un sentido distinto en este contexto.

Khan & Hafiq (2023) analizan más de 8 millones de tuits y 2 millones de perfiles LGTB en X, utilizando DistilBERT y Perspective para medir la toxicidad. Los resultados indican alta toxicidad en mensajes asociados con ira, miedo y tristeza. Dacon et al. (2022) recopilaron 9.930 mensajes sobre la comunidad LGTB en Reddit y utilizaron Detoxify2 (basado en BERT), para detectar varios tipos de toxicidad y compararon el rendimiento de modelos de aprendizaje automático (LM y SVM) y modelos pre-entrenados (BERT, RoBERTa y HateBERT) en tareas binarias y multi-etiqueta, donde los pre-entrenados demostraron un rendimiento superior.

5. DISCUSIÓN Y CONCLUSIONES

La detección del discurso de odio contra la comunidad LGTB en las redes sociales es un área de investigación compleja que requiere un enfoque polifacético y adaptable. Es esencial tener en cuenta la diversidad lingüística y cultural, así como la sensibilidad del tema a la hora de recopilar datos y desarrollar modelos de detección. Esto subraya la necesidad de investigar y abordar el discurso de odio LGTBfóbico en una variedad de lenguas y culturas para una comprensión más completa y global del problema.

Hemos observado que la anotación manual puede presentar ventajas y aportar valor con los temas de subjetividad, pero es un proceso laborioso y requiere una inversión significativa de tiempo y recursos humanos, lo que resulta poco práctica, especialmente para etiquetar grandes volúmenes de datos. Además, la capacidad de análisis humano para el procesamiento efectivo de datos en un

[1] https://perspectiveapi.com/how-it-works/

período de tiempo es limitada, lo que restringe la escalabilidad a medida que aumenta el volumen de datos.

Por lo tanto, es evidente que la clasificación manual del discurso de odio dirigido a la comunidad LGTB es más relevante cuando se dispone de un corpus restringido de datos, como en el contexto de determinados acontecimientos sociales (Valerio 2022; Sánchez-Holgado et al. 2023; Colussi et al. 2024). Igualmente, la clasificación manual es importante como etapa complementaria y los conjuntos de datos elaborados manualmente fomentan el desarrollo de modelos eficientes de detección automatizada (Carvalho et al. 2023) como hemos observado en muchos de los trabajos discutidos en esta revisión. También es un inconveniente la sensibilidad del tema y la falta de anotadores cualificados, especialmente en contextos culturales específicos. Además, la recopilación de conjuntos de datos diversificados y representativos sigue siendo un reto.

Con respecto a los estudios de clasificación automatizada revisados, aunque se han realizado esfuerzos en varios idiomas y contextos culturales, la mayoría contemplan el inglés e idiomas dravídicos. Este enfoque en los idiomas dravídicos está impulsado principalmente por Chakravarthi, quien participa en muchas de las investigaciones revisadas. Aunque el inglés domina la investigación en tecnología del lenguaje, la inclusión de idiomas de bajo recurso y el trabajo con código mixto son cruciales para captar la complejidad lingüística y cultural del discurso de odio en contextos multilingües. Incorporar estos idiomas no solo ayuda a superar sesgos, sino que también proporciona una visión más completa y precisa del discurso de odio, mejorando así la robustez y equidad de los modelos.

Es importante subrayar que hay un número considerable de intentos de análisis de la LGTBfobia en español, tanto usando los métodos manuales (Valerio 2022; Sánchez-Holgado et al. 2023; Colussi et al. 2024), como los de detección de discurso de odio LGTBfóbico computacionales (Arcila et al. 2021; Bel-Enguix et al. 2023; Chakravarthi et al. 2023; Locatelli & Damo 2023) y de discurso esperanzado (García-Baena et al. 2023), lo que puede indicar una tendencia emergente de los estudios en este campo.

Otro punto por destacar es que los estudios revisados emplean una amplia gama de métodos y técnicas, desde el análisis de sentimiento hasta modelos de aprendizaje profundo, así como clasificadores tradicionales. Esta diversidad demuestra la complejidad y la necesidad de adaptarse a diferentes contextos lingüísticos y culturales. Además, aunque los trabajos presentan una variedad de conjuntos de datos, métodos de extracción de características y clasificadores, los modelos de aprendizaje profundo como BERT y sus variantes tienden a superar a los modelos tradicionales en términos de rendimiento en diversos casos. Sin embargo, la eficacia de los modelos varía en función del contexto lingüístico y cultural, lo que subraya la importancia de adaptar y afinar los modelos para cada caso específico.

Como principales oportunidades, vemos la posibilidad de investigaciones multilingües, explorando más idiomas, además de contextos poco explorados. También creemos que los estudios existentes forman una base que se va solidificando, lo que permitiría más estudios comparativos, utilizando diferentes mode-

los, características y conjunto de datos, pero también experimentos que evalúen los ya existentes. Reforzamos, además, las tendencias emergentes de estudios de uso de métodos computacionales para el análisis del discurso LGTBfóbico asociado a contranarrativas, discurso esperanzado e indicadores de toxicidad.

En conclusión, aunque nuestra revisión arroja luz sobre los avances y retos en la clasificación de la incitación al odio, en particular en lo que respecta a la comunidad LGTB, algunas consideraciones merecen atención, como la colaboración interdisciplinar, así como la participación de la comunidad, además de los estudios longitudinales que pueden ofrecer información valiosa sobre la evolución de la naturaleza del discurso del odio relacionado con LGTB a lo largo del tiempo.

6. BIBLIOGRAFÍA

AKHTAR, S., BASILE, V., & PATTI, V. (2019): A new measure of polarization in the annotation of hate speech. In Advances in Artificial Intelligence: Proceedings of the XVIIIth International Conference of the Italian Association for Artificial Intelligence, Rende, Italy, 18 (pp. 588-603). Springer International Publishing.

AKHTAR, S., BASILE, V., & PATTI, V. (2020, October): Modeling annotator perspective and polarized opinions to improve hate speech detection. In Proceedings of the AAAI conference on human computation and crowdsourcing (Vol. 8, pp. 151-154).

ALLPORT, G. W. (1954): The nature of prejudice. Addison-Wesley.

ALSAGHEER, D., MANSOURIFAR, H., & SHI, W. (2022): Counter hate speech in social media: a survey. arXiv:2203.03584. https://doi.org/10.48550/arXiv.2203.03584

ARCE-GARCÍA, S., & MENÉNDEZ-MENÉNDEZ, M.-I. (2023): "Inflaming public debate: A methodology to determine origin and characteristics of hate speech about sexual and gender diversity on Twitter". Profesional de la Informacion, 32(1). https://doi.org/10.3145/epi.2023.ene.06.

ARCILA, C., AMORES, J., SÁNCHEZ-HOLGADO, P. & BLANCO-HERRERO, D. (2021): "Using shallow and deep learning to automatically detect hate motivated by gender reasons and sexual orientation on Twitter in Spanish". Multimodal Technologies and Interactions, 5(10), 63. https://doi.org/10.3390/mti5100063.

BABAKOV, N., LOGACHEVA, V., & PANCHENKO, A. (2022): Beyond plain toxic: detection of inappropriate statements on flammable topics for the Russian language. arXiv:2203.02392. https://doi.org/10.48550/arXiv.2203.02392

BADJATIYA, P., GUPTA, M., & VARMA, V. (2019): Stereotypical Bias Removal for Hate Speech Detection Task using Knowledge-based Generalizations. In Proceedings of the World Wide Web Conference (WWW '19) (pp. 49-59). https://doi.org/10.1145/3308558.3313504

BALAJI, M. J., & CHINMAYA, H. S. (2022): A Study on Sentimental Analysis, Homophobia-Transphobia Detection for Dravidian Languages. In CEUR Workshop Proceedings. https://ceur-ws.org/Vol-3395/T2-7.pdf

BANERJEE, S., & NGUYEN, H. (2023): "Dataset for identification of queerphobia". Journal of Student Research, 12(1). https://doi.org/10.47611/jsrhs.v12i1.4405

BEL-ENGUIX, G., GÓMEZ-ADORNO, H., SIERRA, G., VÁSQUEZ, J., ANDERSEN, S. T., & OJEDA-TRUEBA, S. (2023): "Overview of HOMO-MEX at Iberlef 2023: Hate speech detection in online messages directed towards the MEXican spanish speaking LGTBQ+ population". Procesamiento del lenguaje natural, 71, 361-370. http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6566

BROWNE, K. & NASH, C. (2014): "Resisting LGBT rights where We Have Won: Canada and Great Britain". Journal of Human Rights, 13(3), 322-336. https://doi.org/10.1080/14754835.2014.923754

CARVALHO, P., CALED, D., SILVA, C., BATISTA, F., & RIBEIRO, R. (2023): "The expression of hate speech against Afro-descendant, Roma, and LGTBQ+ communities in YouTube comments". Journal of Language Aggression and Conflict. https://doi.org/10.1075/jlac.00085.car

CASTILLO, V.L.G. (2020): "El control europeo del ciberespacio ante el discurso de odio: análisis de las medidas de lucha y prevención". Araucaria. Revista Iberoamericana de Filosofía, Política, Humanidades y Relaciones Internacionales, 22(45), 291–310. https://doi.org/10.12795/araucaria.2020.i45.12

CHAKRAVARTHI, B. R. (2020): HopeEDI: A multilingual hope speech detection dataset for equality, diversity, and inclusion. In Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media (pp. 41-53). Barcelona, Spain (Online). Association for Computational Linguistics. https://aclanthology.org/2020.peoples-1.5

CHAKRAVARTHI, B. R. (2022a): "Hope speech detection in YouTube comments". Social Network Analysis and Mining, 12(1), 75. https://doi.org/10.1007/s13278-022-00901-z

CHAKRAVARTHI, B. R. (2022b): "Multilingual hope speech detection in English and Dravidian languages". International Journal of Data Science and Analytics, 14, 389–406. https://doi.org/10.1007/s41060-022-00341-0

CHAKRAVARTHI, B. R. (2023): "Detection of homophobia and transphobia in YouTube comments". International Journal of Data Science and Analytics. Advance online publication. https://doi.org/10.1007/s41060-023-00400-0

CHAKRAVARTHI, B. R., HANDE, A., PONNUSAMY, R., KUMARESAN, P. K., & PRIYADHARSHINI, R. (2022): "How can we detect Homophobia and Transphobia? Experiments in a multilingual code-mixed setting for social media governance". International Journal of Information Management Data Insights, 2(2), 100119. https://doi.org/10.1016/j.jjimei.2022.100119

CHAKRAVARTHI, B. R., PONNUSAMY, R., SUBRAMANIAN, M., BUITELAAR, P., GARCÍA-CUMBRERAS, M. A., JIMÉNEZ-ZAFRA, S. M., GARCÍA-DÍAZ, J. A., VALENCIA-GARCÍA, R., & JINDAL, N. (2023): Overview of Second Shared Task on Homophobia and Transphobia Detection in Social Media Comments. In LT-EDI 2023. Third Workshop on Language Technology for Equality, Diversity and Inclusion (pp. 38–46). Varna, Bulgaria. https://doi.org/10.26615/978-954-452-084-7_006

CHAKRAVARTHI, B. R., PRIYADHARSHINI, R., PONNUSAMY, R., KUMARESAN, P. K., SAMPATH, K., THENMOZHI, D., ... & McCRAE, J. P. (2021): Dataset for identification of homophobia and transophobia in multilingual YouTube comments. arXiv:2109.00227. https://doi.org/10.48550/arXiv.2109.00227

CHIRIL, P., PAMUNGKAS, E. W., BENAMARA, F., MORICEAU, V., & PATTI, V. (2022): "Emotionally Informed Hate Speech Detection: A Multi-target Perspective". Cognitive Computation, 14, 322–352. https://doi.org/10.1007/s12559-021-09862-5

COLUSSI, J., GARCÍA-ESTÉVEZ, N., & BALLESTEROS-AGUAYO, L. (2024): "Polarización y odio contra la Ley Trans de España en TikTok". Revista ICONO 14. Revista científica de Comunicación y Tecnologías emergentes, 22(1), https://doi.org/10.7195/ri14.v22i1.2088

ÇETINKAYA, A., DONDURUCU, Z.B. & YETKIN CILIZOĞLU, G., (2021): Instances of Hate Speech Directed toward LGBTIQ+ People on Social Media Platforms in Turkey and Measures Taken against Discrimination. In N. Akıncılar KöseoĞu & D. Apak (Eds.) Challenging Discrimination in Different Areas: Turkey (pp. 45-80), Berlin, Bern, Bruxelles, New York, Oxford, Warszawa & Wien: Peter Lang.

DACON, J., SHOMER, H., CRUM-DACON, S., & TANG, J. (2022): Detecting Harmful Online Conversational Content towards LGBTQIA+ Individuals. arXiv:2207.10032. https://doi.org/10.48550/arXiv.2207.10032

DOĞANÇ, M., & MARKOV, I. (2023): From Generic to Personalized: Investigating Strategies for Generating Targeted Counter Narratives against Hate Speech. In Proceedings of the 1st Workshop on CounterSpeech for Online Abuse (CS4OA) (pp. 1-12).

DONZELLI, S. (2021): "Countering Harmful Speech Online. (In)effective Strategies and the Duty to Counterspeak". Phenomenology and Mind, 20, 76-87. https://doi.org/10.17454/pam-2007

ELSHERIEF, M., ZIEMS, C., MUCHLINSKI, D., ANUPINDI, V., SEYBOLT, J., DE CHOUDHURY, M., & YANG, D. (2021): Latent hatred: A benchmark for understanding implicit hate speech. arXiv:2109.05322. Latent hatred: A benchmark for understanding implicit hate speech. arXiv:2109.05322

EUROPEAN COMMISSION (2019): Special Eurobarometer 493. Discrimination in the European Union: The Social Acceptance of LGBTI People in the EU. Brussels. http://data.europa.eu/euodp/en/data/dataset/S2251_91_4_493_ENG

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:2107.08720. https://doi.org/10.48550/arXiv.2107.08720

FORTUNA, P., & NUNES, S. (2019): "A survey on automatic detection of hate speech in text". ACM Computing Surveys, 51(4), 1–30. https://doi.org/10.1145/3232676

FRANZA, J., EVKOSKI, B., FIŠER, D. (2022): "Emotion analysis in socially unacceptable discourse". SlovenšĞina 2.0, 10(1), 1–22. https://doi.org/10.4312/slo2.0.2021.2.41-70

FRANZA, J., & FIŠER, D. (2019): "The lexical inventory of Slovene socially unacceptable discourse on Facebook". In Proceedings of the 7th Conference on CMC and Social Media Corpora for the Humanities (CMC-Corpora2019), Cergy-Pontoise University, France. pp.43-47. https://hal.archives-ouvertes.fr/hal-02292616/document

GARCÍA-BAENA, D., GARCÍA-CUMBRERAS, M. Á., & JIMÉNEZ-ZAFRA, S. M. (2023): "Hope speech detection in Spanish: The LGTB case". Language Resources and Evaluation, 57(6), 1487–1514. https://doi.org/10.1007/s10579-023-09638-3

GEVERS, I., MARKOV, I., & DAELEMANS, W. (2022). Linguistic analysis of toxic language on social media. Computational Linguistics in the Netherlands Journal, 12, 33-48.

GÓMEZ BELLVIS, A.B. & CASTRO TOLEDO, F.J. (2022): "Los delitos de expresión política en redes sociales desde los efectos de la sanción penal: ¿Efecto disuasorio o efecto desafío?". Revista chilena de Derecho y tecnología, pp 323- https://doi.org/10.5354/0719-2584.2022.66547

GONÇALVES, J., WEBER, I., MASULLO, G. M., TORRES DA SILVA, M., & HOFHUIS, J. (2021): "Common sense or censorship: How algorithmic moderators and message type influence perceptions of online content deletion". New Media & Society, 25(10), 2595-2617. https://doi.org/10.1177/14614448211032310

GUILLEN-NIETO, V. (2023): La lingüística del discurso de odio; perspectivas lingüísticas. DeGruyter. https://doi.org/10.1515/9783110672619-202

GUPTA, R., DESAI, S., GOEL, M., BANDHAKAVI, A., CHAKRABORTY, T., & AKHTAR, M. S. (2023): Counterspeeches up my sleeve! intent distribution learning and persistent fusion for intent-conditioned counterspeech generation. arXiv:2305.13776. https://doi.org/10.48550/arXiv.2305.13776

ILGA WORLD: MENDOS, L.R, BOTHA, K., CARRANO LELIS, R., LÓPEZ DE LA PEÑA, E., SAVELEV, I., & TAN, D. (2020): State-Sponsored Homophobia 2020: Global Legislation Overview Update (Geneva: ILGA, December 2020). https://ilga.org/es/informe-homofobia-estado/

ISHITA, S. (2019): "Contextualizing Hate Speech: A Study of India and Malaysia" Journal of International Studies, 15, 133-144. https://doi.org/10.32890/jis2019.15.9

KHAN, A.N., & RAFIQ, R.I. (2023): A Preliminary Analysis of Twitter's LGBTQ+ Discussions. In: Lossio-Ventura, J.A., Valverde-Rebaza, J., Díaz, E., Alatrista-Salas, H. (eds) Information Management and Big Data. SIMBig 2022. Communications in Computer and Information Science, vol 1837. Springer, Cham. https://doi.org/10.1007/978-3-031-35445-8_1

KUMAR, D., KELLEY, P. G., CONSOLVO, S., MASON, J., BURSZTEIN, E., DURUMERIC, Z., ... & BAILEY, M. (2021): Designing toxic content classification for a diversity of perspectives. In Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021) (pp. 299-318). https://dl.acm.org/doi/abs/10.5555/3563572.3563588

KUMARESAN, P. K., PONNUSAMY, R., PRIYADHARSHINI, R., BUITELAAR, P., & CHAKRAVARTHI, B. R. (2023): "Homophobia and transphobia detection for low-resourced languages in social media comments". Natural Language Processing Journal, 5, 100041. https://doi.org/10.1016/j.nlp.2023.100041

KUNST, M., PORTEN-CHEÉ, P., EMMER, M., & EILDERS, C. (2021): "Do "Good Citizens" fight hate speech online? Effects of solidarity citizenship norms on user responses to hate comments". Journal of Information Technology & Politics, 18(3), 258-273. https://doi.org/10.1080/19331681.2020.1871149

LEMMENS, J., MARKOV, I., & DAELEMANS, W. (2021): Improving hate speech type and target detection with hateful metaphor features. In Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda (pp. 7-16).

LJUBEŠIĆ, N., FIŠER, D., & ERJAVEC, T. (2019): The FRENK datasets of Socially Un-acceptable Discourse in Slovene and English. In International Conference on Text, Speech, and Dialogue. Springer, Cham. https://doi.org/10.1007/978-3-030-27947-9_9

LOCATELLI, D., DAMO, G., & NOZZA, D. (2023): A cross-lingual study of homotransphobia on Twitter. In Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP) (pp. 16-24).

MARTÍNEZ VALERIO, L., & MAYAGOITÍA SORIA, A.M. (2021): "Influencers y mensajes de odio: Jóvenes y consumo de contenidos autocensurados nuevas formas de comunicación". Revista prisma social, 34, 4-39. https://revistaprismasocial.es/article/view/4343/5007

MIRÓ LINARES, F., & GOMEZ BELLVIS, A.B. (2021): "Freedom of expression in social media and criminalization of hate speech in Spain: evolution, impact and empirical analysis of normative compliance and self-censorship". Spanish Journal of Legislative Studies, 1. https://doi.org/10.21134/sjls.v0i1.1837

MOLINA-VILLEGAS, A., CATTIN, T., GAZCA-HERNANDEZ, K., & ALDANA-BOBADILLA, E. (2023): "High-Quality Data from Crowdsourcing towards the Creation of a Mexican Anti-Immigrant Speech Corpus". Applied Sciences, 13(14), 8417. https://doi.org/10.3390/app13148417

MOZAFARI, M., FARAHBAKHSH, R., & CRESPI, N. (2020): Hate speech detection and racial bias mitigation in social media based on BERT model. PLoS ONE, 15(8), e0237861. https://doi.org/10.1371/journal.pone.0237861

MÜLLER, A., & LOPEZ-SANCHEZ, M. (2021): Countering Negative Effects of Hate Speech in a Multi-Agent Society. En M. Villaret, T. Alsinet, C. Fernández, & A. Valls (Eds.), Frontiers in Artificial Intelligence and Applications (Vol. 339, pp. 103-112). IOS Press Ebooks. ISBN 978-1-64368-210-5 (print) | 978-1-64368-211-2 (online). DOI: 10.3233/FAIA210122

NASCIMENTO, F.R.S., & CAVALCANTI, G.D.C., & DA COSTA-ABREU, M. (2023): "Exploring Automatic Hate Speech Detection on Social Media: A Focus on Content-Based Analysis". Sage Open, 13(2). https://doi.org/10.1177/21582440231181311

OLIVA, T. D., ANTONIALLI, D. M., & GOMES, A. (2021): "Fighting hate speech, silencing drag queens? Artificial intelligence in content moderation and risks to LGBTQ voices online". Sexuality & Culture, 25, 700–732. https://doi.org/10.1007/s12119-020-09790-w

PAPCUNOVÁ, J., MARTONČIK, M., FEDÁKOVÁ, D., KENTOŠ, M., BOZOGÁŇOVÁ, M., SRBA, I., MORO, R., PIKULIAK, M., ŠIMKO, M., & ADAMKOVIČ, M. (2021): "Hate speech operationalization: A preliminary examination of hate speech indicators and their structure". Complex & Intelligent Systems, 9, 2827-2842. https://doi.org/10.1007/s40747-021-00561-0

PAZ, M. A., MONTERO-DÍAZ, J., & MORENO-DELGADO, A. (2020): "Hate speech: A systematized review". Sage Open, 10(4), 1-12. https://doi.org/10.1177/2158244020973022

PELICON, A., SHEKHAR, R., ŠKRLJ, B., PURVER, M., & POLLAK, S. (2021): "Investigating cross-lingual training for offensive language detection". PeerJ Computer Science, 7, e559. https://doi.org/10.7717/peerj-cs.559

SAHA, K., CHANDRASEKHARAN, E., & DE CHOUDHURY, M. (2019): Prevalence and Psychological Effects of Hateful Speech in Online College Communities. In Proceedings of the 10th ACM Conference on Web Science (WebSci '19), 255-264. https://doi.org/10.1145/3292522.3326032

SÁNCHEZ-HOLGADO, P., ARCILA-CALDERÓN, C., & GOMES-BARBOSA, M. (2023): Hate Speech and Polarization Around the" Trans Law" in Spain. Politics and Governance, 11(2), 187-197. https://doi.org/10.17645/pag.v11i2.6374

SÁNCHEZ-SÁNCHEZ, A. M., RUIZ-MUÑOZ, D., & SÁNCHEZ-SÁNCHEZ, F. J. (2024): Mapping Homophobia and Transphobia on Social Media. Sexuality Research and Social Policy, 21, 210–226. https://doi.org/10.1007/s13178-023-00879-z

SHARMA, D., GUPTA, V., & SINGH, V. K. (2023): Detection of Homophobia & Transphobia in Malayalam and Tamil: Exploring Deep Learning Methods. In 2nd International Conference on Advanced Network Technologies and Intelligent Computing, ANTIC 2022 (Vol. 1798, pp. 217-226). Communications in Computer and Information Science. https://doi.org/10.1007/978-3-031-28183-9_15

SILVA, M. P. DA, & SILVA, L. S. da. (2021): "Hate speech dissemination in news comments: analysis of news about LGTB universe on Facebook cybermedia from Mato Grosso do Sul". Revista Brasileira de Ciências da Comunicação, 44(2), May-Aug, https://doi.org/10.1590/1809-5844202127

SORAL, W., BILENWICZ, M., & WINIEWSKI, M. (2018): "Exposure to hate speech increases prejudice through desensitization". Aggressive Behavior, 44(2), 136–146. https://doi.org/10.1002/ab.21737

SPONHOLZ, L., & CHRISTOFOLETTI, R. (2019): "From preachers to comedians: Ideal types of hate speakers in Brazil". Global Media and Communication, 15(1), 67-84. https://doi.org/10.1177/1742766518818870

ŞTEFĂNIŢĂ, O., & BUF, D. (2021): "Hate Speech in Social Media and Its Effects on the LGBT Community: A review of the current research". Romanian Journal Of Communication and Public Relations, 23(1), 47-55. https://doi.org/10.21018/rjcpr.2021.1.322

TEIJÓN ALCALÁ, M. (2022): "La radicalización y el (auto)adoctrinamiento como valores causalmente relevantes de acciones terroristas. Un estudio empírico en el marco de la teoría subcultural de la violencia". Revista Electrónica de Estudios Penales y de la Seguridad, 11. https://www.ejc-reeps.com/La%20radicalizacion%20teijon.pdf

TEKIROGLU, S. S., BONALDI, H., FANTON, M., & GUERINI, M. (2022): Using pretrained language models for producing counter narratives against hate speech: a comparative study. arXiv:2204.01440. https://doi.org/10.48550/arXiv.2204.01440

THOMAS, D.R., & WAHEDI, L.A. (2023): Disrupting hate: The effect of deplatforming hate organizations on their online audience. Proceedings of the National Academy of Sciences of the USA, 120(24):e2214080120. https://doi.org/10.1073/pnas.2214080120.

VALERIO, L. M. (2022): "Hate messages towards the LGTBQ+ community: Instagram profiles of the Spanish press analysis during Pride Week". Revista Latina de Comunicacion Social, 80, 363-388. https://doi.org/10.4185/RLCS-2022-1749.

WACHS, S., COSTELLO, M., WRIGHT, M. F., FLORA, K., DASKALOU, V., MAZIRIDOU, E., KWON, Y., NA, E.-Y., SITTICHAI, R., BISWAL, R., SINGH, R., ALMENDROS, C., GÁMEZ-GUADIX, M., GÖRZIG, A., & HONG, J. S. (2021): "DNT LET 'EM H8 U!: Applying the routine activity framework to understand cyberhate victimization among adolescents across eight countries". Computers & Education, 160, 104026. https://doi.org/10.1016/j.compedu.2020.104026

WACHS, S., WRIGHT, M. F., SITTICHAI, R., SINGH, R., BISWAL, R., KIM, E.-M., YANG, S., GÁMEZ-GUADIX, M., ALMENDROS, C., FLORA, K., DASKALOU, V., & MAZIRIDOU, E. (2019): "Associations between Witnessing and Perpetrating Online Hate in Eight Countries: The Buffering Effects of Problem-Focused Coping". International Journal of Environmental Research and Public Health, 16(20), 3992. https://doi.org/10.3390/ijerph16203992

WINDISCH, S., WIEDLITZKA, S., OLAGHERE, A., & JENAWAY, E. (2022): "Online interventions for reducing hate speech and cyberhate: A systematic review". Campbell Systematic Reviews, 18(2), e1243. https://doi.org/10.1002/cl2.1243

ZHANG, Y., & TRIFIRO, B. (2022): "Who Portrayed It as the Chinese Virus? An Analysis of the Multiplatform Partisan Framing in U.S. News Coverage About China in the COVID-19 Pandemic". International Journal of Communication, 16, 1027-1050. https://ijoc.org/index.php/ijoc/article/view/17916