Discurso de odio en Twitter: Análisis de la LGTBIQ-fobia antes y después de Elon Musk
Hate Speech on Twitter: Analysis of LGBTIQ-phobia Before and After Elon Musk

Trabajo de Fin de Grado
Curso 2024–2025

Autor
María del Mar Ramiro Ortega

Director
Samer Hassan Collado

Calificación: *10*

Doble Grado en Ingeniería Informática y Matemáticas
Facultad de Informática
Universidad Complutense de Madrid

Discurso de odio en Twitter: Análisis de la LGTBIQ-fobia antes y después de Elon Musk
Hate Speech on Twitter: Analysis of LGBTIQ-phobia Before and After Elon Musk

Trabajo de Fin de Grado en Ingeniería Informática

Autor
María del Mar Ramiro Ortega

Director
Samer Hassan Collado

Convocatoria: *Junio 2025*

Calificación: ***10***

Doble Grado en Ingeniería Informática y Matemáticas
Facultad de Informática
Universidad Complutense de Madrid

26 de mayo de 2025

Agradecimientos

A la educación pública, para que siga siendo de calidad. A mis amigos de clase, por su ayuda y apoyo durante estos cinco años. A los amigos y familiares que han revisado "voluntariamente" este trabajo. A los profesionales que me han ayudado a realizar este proyecto. Muchas gracias a todas y todos por vuestro tiempo.

Y en especial a Samer, que me ha permitido orientar el trabajo a mis intereses, dándome la libertad de decidir, siempre desde el acompañamiento. Ha sido un placer y una suerte tenerte como tutor, gracias por enseñarme tu vocación.

Resumen

Discurso de odio en Twitter: Análisis de la LGTBIQ-fobia antes y después de Elon Musk

Las redes sociales han supuesto tanto beneficios como perjuicios. Han facilitado el acceso a la información, así como generado reflexión y debate, pero también han propiciado desinformación y discurso de odio online. Tras la compra de Twitter por Elon Musk, se eliminaron y modificaron las políticas para la detección y moderación del contenido de odio. Esto ha generado que muchos investigadores y usuarios tengan la plataforma en el punto de mira. Denuncian un aumento de contenido ofensivo, como el racismo o la LGTBIQ-fobia. Este cambio ha desencadenado la migración de usuarios a otras plataformas.

Este proyecto es el primer estudio en español que analiza la situación de la LGTBIQ-fobia en Twitter tras la llegada de Musk. El marco temporal abarca desde 2015 hasta 2024, centrando su estudio en los tuits en español del 28 de junio de cada año (Día del Orgullo). Para ello se ha recolectado un dataset de 653.000 tuits sobre el colectivo. Como punto de referencia para hacer comparaciones se ha recolectado un dataset de 395.000 sobre contenido aleatorio.

Los resultados muestran las consecuencias de las distintas modificaciones en las políticas de moderación del contenido de odio. En particular, tras la llegada de Elon Musk, se observa un aumento del 46,81 % en el número de tuits tóxicos sobre la comunidad LGTBIQ+. De hecho, si solo se consideran tuits extremadamente tóxicos, el aumento es del 232 %. En términos de visibilidad e interacción, el contenido tóxico ha subido un 159,2 % en "likes" y un 81,59 % en "retuits".

Palabras clave

Análisis de datos, Comunidad LGTBIQ+, Conjuntos de datos, Discurso de odio, Investigación cuantitativa, LGTB, LGTBIQ-fobia, LGTBQIA, Twitter, X.com.

Abstract

Hate Speech on Twitter: Analysis of LGBTIQ-phobia Before and After Elon Musk

Social media has brought both benefits and harm. It has facilitated access to information and generated reflection and debate, but it has also led to misinformation and online hate speech. Following Elon Musk's purchase of Twitter, some policies for detecting and moderating hateful content were removed and others modified. This made many researchers and users to focus their attention on the platform. They report an increase in offensive content, such as racism and LGBTIQ-phobia. As a result many users have migrated to other platforms.

This is the first study in Spanish to analyze the situation of LGBTIQ-phobia on Twitter after Musk's arrival. It examines Spanish-language tweets from June 28 (Pride Day), covering 2015 to 2024. To do so, a dataset of 650,000 tweets about the LGBTIQ+ community was collected. As a benchmark for comparison, a dataset of 390,100 random content tweets was collected.

The results highlight the consequences of changes in hate speech moderation policies over time. Specifically, the results show a 46.81% increase in the number of toxic community-related tweets, a 159.2% in "likes" and a 81.59% in "retweets" following Elon Musk's acquisition of Twitter.

Keywords

Data Analysis, Datasets, Hate Speech, LGBT, LGBTIQ+ community, LGBTIQ-phobia, LGBTQIA, Quantitative Research, Twitter, X.com.

Índice

Índice de figuras

Índice de tablas

Glosario de Términos

API	Application Programming Interface
BERT	Bidirectional Encoder Representations from Transformers
LGTBIQ-fobia	rechazo, miedo, prejuicio o discriminación hacia personas LGTBIQ+ o aquellas asociadas a ellas
LGTBIQ+	Lesbianas, Gays, Bisexuales, Transgéneros, Intersexuales, Queer y otras identidades de género y orientaciones sexuales

Capítulo 1

Introducción

"La libertad de uno acaba donde empieza la del otro"
— Rousseau

1.1. Motivación

En las últimas décadas hemos sido testigos de un gran avance por la inclusión y la mejora de los derechos de las personas del colectivo LGTBIQ+. La llegada de Internet y, con ella, la globalización de la información, ha facilitado el reconocimiento de diversas minorías, entre ellas este colectivo. Muchas aplicaciones como Twitter han sido nido de crítica social y han favorecido el estudio y avance de la inclusión de estas categorías sociales. Esta plataforma ha permitido el diálogo entre personas de todo tipo de pensamientos, generando reflexión y debate. Además, a lo largo de los años se han implementado políticas que buscan reducir el discurso de odio y la desinformación, contribuyendo a la creación de un ambiente más inclusivo.

Este ambiente de avance sufrió un giro abrupto tras los cambios en las políticas de moderación de contenido de odio realizados por Elon Musk, quien compró la plataforma en octubre de 2022. Estas modificaciones fueron motivadas por lo que el empresario define como "libertad de expresión", afirmando que, pese a los cambios, la exposición a contenido de odio y desinformación había disminuido [50, 111]. Sin embargo, estas declaraciones carecen de transparencia, pues X podría tener una definición muy diferente al resto de la sociedad de lo que son las expresiones "discurso de odio" o "libertad de expresión". Un claro ejemplo de ello es la reciente prohibición de la palabra "cisgender", traducido al español, cisgénero (persona que se siente conforme con su identidad de género asumida al nacer [70]), tras ser considerada un insulto. Sin embargo, esta es una palabra utilizada comúnmente por médicos [101] y personas queer [82].

Los cambios de las políticas unidos a la creciente polarización de nuestra sociedad han convertido Twitter en un lugar hostil para ciertas minorías, incluida la

comunidad LGTBIQ+. Esta situación ha motivado la migración de muchos usuarios a otras redes sociales como son Bluesky, Threads o Mastodon [115, 54], además de un auge del interés por parte de los investigadores, quienes buscan alertar de la situación actual y los posibles impactos que sucederían tanto a corto como a largo plazo si dicha situación se alargara en el tiempo.

En este proyecto se estudia el aumento de la LGTBIQ-fobia en Twitter tras la llegada de Elon Musk, además de una discusión de las posibles causas y consecuencias. Esto invita a la presión social para la implementación de medidas que moderen contenido claramente ofensivo para esta minoría, no sólo en Twitter, sino también en otras redes sociales. Se busca tanto la introducción de nuevas políticas de moderación del contenido de odio como el mantenimiento de aquellas que ya se encuentran vigentes.

Este proyecto se ha realizado sobre tuits en español y estudia las consecuencias de los cambios en las políticas a lo largo de los últimos diez años. Además, se contempla el discurso de odio como un espectro, no como un valor único. De esta manera, se trata de cubrir las limitaciones metodológicas de trabajos similares causadas por los sesgos lingüísticos, temporales y de concepto (como se explicará en el estado del arte, sección 2.2.5).

1.2. Objetivos

El objetivo principal de este proyecto es el estudio de las repercusiones de la llegada de Elon Musk a Twitter sobre el colectivo LGTBIQ+, observando la tendencia del contenido y exponiendo la situación actual. Para ello, se han concretado los siguientes tres objetivos:

- Análisis cuantitativo del aumento del discurso de odio, en general, y hacia el colectivo LGTBIQ+, entre 2015 y 2024, con atención a distintos periodos significativos.
- Análisis cuantitativo del aumento de la interacción y visibilidad del contenido de odio, en general, y hacia el colectivo, en el mismo periodo.
- Estudiar cómo la compra de Elon Musk ha repercutido en el discurso de odio en general y en particular hacia el colectivo, analizando la significación estadística de las diferencias encontradas.

Para abordar el primer objetivo, se ha analizado la evolución del contenido de odio a lo largo del tiempo. Para tratar el segundo objetivo, se ha examinado la interacción con dicho contenido, esto es, los "likes" y "retuits" recibidos. Por último, se ha dividido el marco temporal en diferentes períodos: pre-pandemia, pandemia, post-pandemia y post-Musk. Los resultados hacen especial énfasis en los cambios tras la adquisición de Twitter por parte de Elon Musk, lo que permite tratar el tercer objetivo.

1.3. Plan de trabajo

La primera fase, tal y como se hace en cada proyecto de investigación, consistió en la comprensión del problema. Para ello, se realizaron reuniones con distintos investigadores de ciencias sociales, que ayudaron a la comprensión y concreción del tema de estudio. Además, se leyeron trabajos parecidos, lo que daría algunas claves, inspiración y recursos que se utilizarían después.

En segundo lugar, se afrontó la que sería la mayor problemática del proyecto: la recolección de los tuits. Se realizó una primera recolección de tuits y un primer análisis básico. Dicho estudio confirmó nuestras hipótesis, lo que dio comienzo al grueso del proyecto.

En este contexto, se recolectaron todos los tuits que contenían ciertas palabras clave (se explicará en 3.2.2) del 28 de junio de los últimos diez años. Para llevar a cabo los estudios, fue necesario recolectar un conjunto de datos de tuits aleatorios como referencia, con el fin de compararlos con el corpus del colectivo. Para ello, se seleccionaron tres palabras comunes en español ("ser", "algo" y "todos"), y se recolectaron y procesaron los tuits.

El siguiente paso importante fue por consiguiente el análisis de los datos definitivos, para estudiar la subida de la LGTBIQ-fobia por medio de gráficos y porcentajes de aumento para facilitar las comparaciones, para lo cual usamos Jupyter Notebook. Por último, se comprobó que estos datos fueran estadísticamente significativos.

El último mes estuvo dedicado a la redacción y revisión de la memoria.

1.4. Estructura de la memoria

La memoria se divide en cinco capítulos:

- Introducción: se presenta el trabajo, así como la importancia del mismo. Además, se exponen los motivos que hacen que este sea un tema relevante y los objetivos principales que trata de cubrir.
- Fundamentos teóricos: busca dar al lector/a las herramientas necesarias para entender el proyecto, describiendo los conceptos clave y repasando los trabajos realizados anteriormente en esta área.
- Tecnologías y metodología: presentan la tecnología usada y la metodología seguida para la obtención de los resultados.
- Resultados y su análisis: se expone el análisis estadístico de los datos estudiados. Así como los resultados de los estudios mencionados en la sección (sección 5.3): cantidad de tuits tóxicos y el impacto de los mismos a lo largo del tiempo.

- Conclusiones y trabajo a futuro: se hace una discusión razonada de la conclusión, exponiendo las limitaciones y posibles trabajos a futuro.

Capítulo 2

Fundamentos teóricos y estado del arte

Esta sección tiene como propósito la definición de conceptos clave para el entendimiento del proyecto. Además, se tratará de poner en contexto al lector con los trabajos previos realizados en el área.

2.1. Fundamentos teóricos

2.1.1. Colectivo LGTBIQ+ y LGTBIQ-fobia

Este trabajo se centra en explorar la LGTBIQ-fobia hacia el colectivo LGTBIQ+, por lo que su definición será de especial relevancia. Para no repetir demasiado, se usará en ocasiones "el colectivo" para referirse a la comunidad LGTBIQ+.

Como se define en [64], las siglas LGTBIQ+ representan a la diversidad de identidades de género y de orientaciones sexuales. La primera letra, "L", se refiere a las lesbianas, mujeres que sienten atracción afectiva y/o sexual hacia otras mujeres. La letra "G" se refiere a los gays, hombres que sienten atracción afectiva y/o sexual hacia otros hombres. La letra "T" incluye a las personas transgénero que es un término que engloba a aquellas personas cuya identidad y/o expresión de género es diferente de las expectativas culturales basadas en el sexo que se les asignó al nacer. Incluye personas transexuales, transgéneros, personas trans no binarias, con expresión de género fluido y otras variaciones de género [45]. La letra "B" hace referencia a las personas bisexuales, es decir, que sienten atracción afectiva y/o sexual por más de un género. La "I" representa a las personas intersexuales, aquellas que nacen con características biológicas que no se ajustan a las típicas categorías de masculino o femenino. La "Q" representa la palabra 'queer' y hace referencia a aquellas personas que se identifican más allá de las categorías tradicionales del sistema binario varón/mujer, heterosexualidad/homosexualidad.

Siguiendo con la definición de [64], el símbolo "+" busca añadir a los colectivos que no están representados en las siglas anteriores. De esta forma, están incluidas identidades como las personas no binarias (personas que no encajan dentro del modo binario de entender el género), así como otras orientaciones sexuales como la pansexualidad (sentir atracción sexual y/o romántica por las personas sin importar su sexo o género), la asexualidad (personas que no experimentan atracción sexual y/o no desean contacto sexual) o la demisexualidad (personas que solo sienten atracción sexual hacia alguien después de haber formado un fuerte lazo emocional). Al utilizar este símbolo, se busca crear un espacio inclusivo y acogedor para todas las personas [64].

La LGTBIQ-fobia hace referencia al rechazo, miedo, prejuicio o discriminación hacia personas LGTBIQ+ o aquellas asociadas a ellas [32].

2.1.2. Discurso de odio

El término de discurso del odio es un concepto discutido, y cuya definición está sujeta a debate político, jurídico y académico a nivel internacional [36]. En este proyecto vamos a tomar como definición la que da la ONU: "cualquier tipo de comunicación ya sea oral o escrita, —o también comportamiento— , que ataca o utiliza un lenguaje peyorativo o discriminatorio en referencia a una persona o grupo en función de lo que son, en otras palabras, basándose en su religión, etnia, nacionalidad, raza, color, ascendencia, género u otras formas de identidad" [109].

De hecho, el Código Penal tipifica, en su artículo 510.1 el conocido como "delito de odio". Concretamente, dicho artículo impone una pena de prisión de uno a cuatro años a quienes de manera pública impulsen, promuevan o inciten —-ya sea de forma directa o indirecta—- el odio, la hostilidad, la discriminación o la violencia hacia un grupo, parte de este, o hacia una persona por el hecho de pertenecer a dicho grupo, motivados por, entre otras, razones de orientación sexual, identidad sexual [3].

2.1.3. Análisis estadístico

Para sacar conclusiones de un conjunto de datos es de especial importancia realizar previamente un análisis estadístico. En el estudio realizado en la sección 4.1, se utilizan una serie de conceptos estadísticos. Dichos conceptos se definirán en esta sección.

2.1.3.1. P-valor

Cuando se realiza un análisis de datos, una práctica común y necesaria es el estudio de las distribuciones de las muestras a comparar. Este estudio nos permitirá ver de manera fiable las diferencias y similitudes entre los distintos conjuntos de

datos. Es decir, nos ayudará a concluir si existen evidencias de que una misma variable se distribuya de manera distinta entre dos grupos [91].

Cuando se comparan estadísticos entre dos muestras (media, cuartiles, mediana, etc.) una de las principales preguntas que surgen es si esas diferencias existen en realidad o se deben únicamente a variaciones aleatorias de los datos utilizados. Parece importante, pues, el uso de tests analíticos que calculen la probabilidad exacta de observar diferencias de esa magnitud o mayor si en realidad no existiese diferencia, esta probabilidad se conoce como p-valor [91]. La definición formal es $p - valor = Probabilidad$(resultado tan extremo o más |hipótesis nula). Que se lee como la probabilidad de que el resultado sea tan extremo o más supuesta la hipótesis nula. Este valor de probabilidad estará entre 0 y 1. Así, si el valor es pequeño, indicaría que, siendo cierta la hipótesis nula, era muy difícil que se produjera el valor que se ha observado. Ello obliga a poner muy en duda, y por tanto a rechazar, la hipótesis nula [37]. Cada test estadístico define una hipótesis nula distinta, dependiendo del tipo de diferencia que se quiera medir.

Por convenio, un p-valor $< 0{,}05$ o $0{,}01$ es suficiente para afirmar que los resultados son reales. Ya que esto representará que la hipótesis nula se puede rechazar con un 95 % o 99 % de confianza. Es decir, hay un 95 % o 99 % de confianza de que el resultado observado no es producto del azar. No obstante, un valor $> 0{,}05$ o $0{,}01$ no nos da la suficiente confianza como para poder negar los resultados.

2.1.3.2. Percentiles, cuantiles, mediana, media y moda

El percentil es un valor que indica el porcentaje de datos que son menores o iguales a ese valor en un conjunto de datos ordenados. Los cuartiles son percentiles especiales que dividen los datos en cuatro partes iguales. El primer cuartil es el percentil 25. El segundo cuartil es el percentil 50, también conocido como mediana. El tercer cuartil es el percentil 75.

Por otro lado, la media es el promedio de todos los valores en un conjunto de datos. Se obtiene sumando todos los valores y dividiendo el resultado entre la cantidad de datos. Por último, la moda es el valor que más se repite en un conjunto de datos.

2.1.3.3. Prueba de Kolmogorov-Smirnov para dos muestras

El estadístico Kolmogorov-Smirnov, conocido como KS, se define como la distancia vertical máxima entre las funciones de distribución acumulada empírica de dos muestras. La distribución acumulada empírica es una función escalonada que indica, en cada valor de la variable, el porcentaje de observaciones que son menores o iguales a ese valor. A medida que aumentas el tamaño de la muestra, esta función se ajusta más a la distribución real que genera los datos [91].

La hipótesis nula en este caso es que ambas muestras provienen de la misma población [5]. Si el p-valor es suficientemente pequeño podremos concluir que no, que su distribución es distinta. Esta prueba es relevante en nuestro proyecto, pues nos permite observar la diferencia en distribución con respecto al discurso de odio. Es decir, si queremos comparar la toxicidad del discurso en dos períodos distintos, no nos basta solo con observar la media y los percentiles (definidos en la sección 2.1.3.2). En este caso debemos comprobar adicionalmente que los datos se distribuyen de manera distinta y que las diferencias obtenidas no son producto del azar.

Un p-valor alto no quitará valor al estudio, sino que nos dará otra información distinta: los datos se distribuyen de la misma manera. Por lo que las comparaciones solo podrán realizarse en el tamaño de la muestra, y no entre los percentiles. Esto es, si los "likes" de dos datasets se distribuyen de la misma manera, no tiene sentido comparar sus medianas, pues deberían ser muy parecidas. Sin embargo, sí parece un estudio interesante calcular la suma total de los "likes", pues nos dirá qué conjunto de datos tiene más visibilidad. Es decir, que los "likes" se distribuyan de la misma manera no nos dice nada sobre el tamaño y relevancia de los datos; eso se debe estudiar por separado.

2.2. Estado del arte

Este proyecto se va a centrar en el estudio de la LGTBIQ-fobia en Twitter, por medio de un estudio temporal (2015 a 2024). Para ello se han estado investigando trabajos con propósitos parecidos. En esta sección se hará un recorrido de aquellos artículos y resultados relevantes para esta investigación.

2.2.1. Discurso de odio en redes sociales

La llegada de las redes sociales ha añadido una capa adicional de complejidad a la detección y denuncia del contenido de odio [112]. La inmensa cantidad de datos, la permanencia de los mismos, el uso de pseudónimos y el anonimato en ellas han potenciado aún más los desafíos de este problema social [52]. Es por ello que observamos un aumento del interés por parte de los investigadores [90, 15], convirtiendo la proliferación y promoción del odio en uno de los mayores desafíos de esta época [85, 100, 30]. Dichos investigadores tratan de denunciar que la moderación del discurso de odio en redes sociales es insuficiente [102]. Los diversos estudios se centran en las aplicaciones más usadas, como puede ser Reddit [17], Youtube [31], Facebook [40] o Twitter [48, 22, 62]. Además, muchos de estos estudios se han centrado en el colectivo LGTBIQ+ [79, 46, 62, 22], aunque abundan también de minorías étnicas o religiosas, migrantes, refugiados, menores de edad, etc. [76].

Las diferentes plataformas suelen moderar el contenido que puede ser altamente ofensivo para otros usuarios, como puede ser el discurso de odio [43]. No obstante,

el reciente aumento del discurso de odio en las mayores plataformas como Twitter [83, 103, 29, 104], es preocupante por muchos motivos:

- La existencia de delitos de odio online está directamente relacionada con los delitos de odio fuera de la red, es decir, en el mundo físico [69, 22].
- Las víctimas de discurso de odio suelen denunciar un claro descenso de su bienestar psicológico [96, 84, 24, 22].
- La exposición a ideologías de odio puede incrementar el prejuicio y la insensibilización [99], así como disminuir la empatía hacía las minorías [86].
- Aquellos usuarios que consumen contenido de odio son más propensos a participar en otras formas de discriminación y utilizar lenguaje despectivo [27].

Algunos artículos denuncian la existencia de contenido de odio hacia la comunidad LGTBIQ+ [102]; otros tratan las consecuencias del discurso de odio online a jóvenes que pertenecen al colectivo [11, 53, 57]; y otros han creado modelos para la detección y denuncia del contenido que atenta contra la identidad sexual o de género de los usuarios [31, 16, 23].

No obstante, estos esfuerzos se ven difuminados por la falta de consenso en la definición de discurso de odio. Pues los datasets anotados no son totalmente fiables, ya que cada uno se ciñe a una definición distinta y esta puede variar mucho [92]. Por otro lado, hay modelos de aprendizaje profundo que tienen buenos resultados en la detección de delitos de odio online. Sin embargo, la falta de explicabilidad hace que sean difíciles de comprender [116, 25, 117, 67]. Por último, cabe destacar la problemática del idioma, ya que, pese a existir muchos estudios, artículos y modelos que funcionan para el inglés, aún queda mucho trabajo en otras lenguas, como puede ser el español [95, 81].

2.2.2. Contexto de Twitter

Twitter, ahora renombrado a "X", es una de las redes sociales más usadas desde su creación en 2006 [98].

Twitter es una red social de micro-blogging donde los usuarios comparten mensajes llamados tuits. Estos tuits tienen un límite de 280 caracteres y pueden contener texto, imágenes, enlaces o videos. Los elementos principales que se pueden encontrar en un tuit son los siguientes:

- Mensaje: el texto del tuit, donde el usuario expresa una idea o comparte algo.
- "Hashtags": palabras o frases precedidas por el símbolo #, utilizadas para etiquetar temas y hacer que el tuit sea fácilmente localizado por otros usuarios interesados en esos mismos temas.

- "Likes" (Me gusta): Indican que a otros usuarios les ha gustado un tuit. Es una forma de mostrar apoyo o acuerdo con el mensaje.
- "Retuits": Permite compartir el tuit de otra persona en tu propia cuenta, mostrando que estás de acuerdo con él o que quieres difundirlo.
- Comentarios: Los usuarios pueden responder a un tuit, creando una conversación sobre el tema del mensaje.

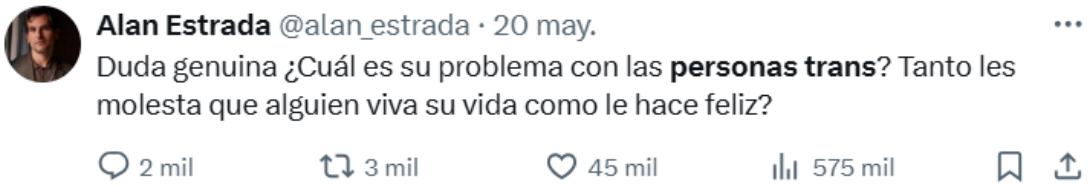

Figura 2.1: Ejemplo de tuit

Por ejemplo, la imagen 2.1 muestra un tuit con su respectivo mensaje. Justo debajo del texto, podemos observar que este tuit tiene 2 mil comentarios, 3 mil "retuits", 45 mil "likes" y 575 mil visualizaciones.

Su amplio uso hace que haya sido base para el estudio del discurso de odio online. De hecho, siempre ha estado muy vinculada a la investigación gracias a su API (antes) gratuita que permitía acceso a todos los tuits públicos y que fue muy útil para estudiar comportamientos y tendencias sociales [28].

Tras la compra de Twitter por parte de Elon Musk en octubre de 2022 [9], por un valor de 44 mil millones de dólares, la plataforma dejó de ser una empresa cotizada en bolsa. Al retirar la empresa de la bolsa, Twitter pasó a ser una empresa privada. Esta decisión permitió a Musk implementar diversas modificaciones en la plataforma sin la supervisión y las obligaciones regulatorias que enfrenta una empresa pública [42]. Destacamos las siguientes:

- El Consejo Asesor de Confianza y Seguridad (Trust and Safety Council), que asesoraba sobre la moderación de contenidos fue disuelto [78].
- Destitución de los altos ejecutivos de la empresa [74].
- Cobrar una cuota mensual a las cuentas por mantener el estatus de "verificado" [63], eliminando la estructura base de las redes sociales, cuyas cuentas verificadas suelen ser identidad de veracidad (independientemente de si se paga o no), y creando el nuevo concepto de "X Premium" [93, 60]. Este nuevo concepto consiste en un pago mensual de 8$ a cambio de obtener el verificado azul y una mayor visibilidad, entre otras ventajas [8].
- Nuevas condiciones laborales de la compañía, que muchos empleados no aceptaron, por lo que prefirieron renunciar. Como respuesta se ordenó el cierre de las oficinas [38, 62].

- Reactivación de la cuenta de Donald Trump y de otras 10.000 cuentas que habían sido suspendidas por la violación de los términos y servicios de la app [28].
- Posibilidad de publicación de posts con contenido falso que previamente habían sido deshabilitadas [28, 94, 49].
- La API de Twitter, que históricamente había sido gratuita ahora sería de pago y con precios elevados [105].

Tras estos cambios, la demografía de X cambió rápidamente, observando un descenso de actividad de muchos activistas, periodistas, artistas, minorías, y de muchos otros usuarios, que migraron a otras aplicaciones como Mastodon, Threads y Bluesky [115, 77, 54]. Esta migración se ha acelerado más recientemente con el alineamiento ideológico y financiación de Musk a Donald Trump, con muchas comunidades de Reddit baneando links a Twitter directamente [10].

Este tema es altamente sensible, pues la división de la sociedad en las distintas apps y los algoritmos utilizados por estas plataformas para gestionar la visibilidad del contenido contribuyen a la fragmentación del discurso. Estos algoritmos, al priorizar ciertos tipos de contenido o comunidades, refuerzan las divisiones ideológicas y dificultan el entendimiento común [88, 72]. Como resultado, se reducen los lugares de encuentro y debate entre distintas posturas, lo que favorece el aislamiento y la polarización de las opiniones.

Un informe publicado por el Center for Countering Digital Hate (CCDH) reveló la falta de medidas para la eliminación de tuits con contenido de odio publicados por cuentas "Twitter Blue" (de pago). El artículo indica que el 99 % de tuits publicados por cuentas verificadas y que fueron reportados a la aplicación usando las propias herramientas de la app para señalar discurso de odio, no fueron eliminados, es decir, solo el 1 % fue eliminado [58]. Esta falta de consenso no se queda aquí. En 2023, Elon Musk y el CCDH estuvieron implicados en un proceso judicial, tras la denuncia del magnate a la organización de investigación sin fines de lucro. Esta denuncia alega que las previas acusaciones del CCDH sobre la no eliminación de contenido de odio afectan de manera directa a la reputación de la compañía X. No obstante, esta demanda fue desestimada en marzo de 2024, tras ser considerada como intimidatoria e infundada, pues las críticas que realizó la "nonprofit" estaban respaldadas por la libertad de expresión [59].

El excesivo coste de la API ($5000 mensuales) y los sucesivos cambios para impedir el "web scraping" en la aplicación suponen grandes complicaciones para acceder a estos datos, con miles de repositorios de "scraping" de hace tan solo unos meses obsoletos [68]. Esto hace que solo aquellos investigadores que estén familiarizados con Python o R y el "data scraping" tengan acceso a dichos datos, y provoca la rendición de muchos investigadores sociales [87].

2.2.3. Cambio de políticas en Twitter

Históricamente, en 2009, Twitter introdujo sus "Rules" ("Reglas") a respetar por sus usuarios. Aunque la compañía prohibía la publicación de "amenazas directas y específicas de violencia contra otros", no se hizo referencia explícita a las políticas de prohibición de contenido de odio hasta finales de 2017 [106].

A finales de 2018 las investigaciones realizadas en el discurso de odio exponían que había ciertos grupos de personas que eran más propensos a ser víctimas del discurso de odio online. Es por ello que Twitter añadió a su "política de conducta de odio" una mención explícita a esas minorías, denominándolos "grupos protegidos". Estos son: "mujeres, personas de color, lesbianas, gays, bisexuales, transgénero, queer, intersexuales, asexuales, comunidades marginadas e históricamente subrepresentadas" [107]. La especificación de grupos protegidos es común en las políticas sobre discurso de odio, como refleja el ordenamiento jurídico español, que protege a colectivos considerados tradicionalmente vulnerables.

Durante la pandemia se produjo un aumento de la actividad, y con este un aumento del contenido de odio, lo que mostró una ineficiencia de las medidas. En respuesta, entre finales de 2020 y principios de 2021 Twitter ejecutó las medidas de prevención del discurso de odio, mostrando un aumento del 77 % de las cuentas sancionadas por no cumplir las políticas de conducta de odio. Además, amplió el número de "grupos protegidos" [113].

Un estudio sugiere que las políticas de contenido de odio establecidas en 2018 que prometían eliminar los estereotipos contra los grupos establecidos no fueron aplicadas hasta 2020, dos años después [73]. Sin acceso a los datos de la empresa sobre el contenido procesado según cada disposición de la política de conducta de odio es imposible saber cómo se aplica la política, así como el impacto de los cambios en la práctica. Este ejemplo ilustra la importancia de que los investigadores tengan acceso a los datos de la plataforma, para poder evaluar y auditar exhaustivamente las afirmaciones de las plataformas sobre la aplicación de las políticas [73].

No fue hasta principios de 2023, una vez comprado Twitter por Musk, cuando se produjo otro fuerte cambio en las políticas. Ahora no se haría una mención explícita a las minorías señaladas anteriormente en su "política de conducta de odio" [114]. Aunque esto afecta a todo el colectivo LGTBIQ+, el más afectado sin duda fue el subgrupo trans, pues el recién rebautizado X, eliminó la política que evitaba el misgendering (referirse a una persona con pronombres con los que no se identifica) y el deadnaming (referirse a una persona con el nombre que tenía previamente a su transición de género) [55], además de la consideración de cisgénero como insulto, y por consiguiente, la prohibición de su uso [82].

2.2.4. Discurso de odio en twitter

En este contexto comenzamos examinando aquellos artículos que hacen mención del aumento del discurso de odio tras la llegada de Elon Musk [61, 26], observando un elevado contenido de odio en las semanas próximas a la adquisición de Twitter por Musk [62].

Es importante estudiar si este aumento se debe a un pequeño pico o si es una dinámica que se ha mantenido durante todo el periodo de Musk como CEO de la plataforma. Es por ello que en el estudio realizado por Hickey et al. [62] se amplía el periodo de estudio hasta el 9 de junio de 2023, abarcando todo el tiempo en el que Musk fue CEO de Twitter [71]. Los resultados muestran un 50 % del aumento del contenido de odio, en comparación con el 8 % de aumento en el resto de actividad. A esto se suma un aumento del 70 % de "likes" al contenido de odio, en contraste con el 4 % del aumento de "likes" al resto de actividad, lo que entra en contradicción con las afirmaciones de Elon Musk, en las que confirmaba un descenso en la interacción con los posts con contenido de odio, a pesar del aumento de los mismos (el famoso "Freedom of Speech, Not Reach", "libertad de expresión, no de alcance", traducido al español) [108].

Sumado a lo anterior, este mismo artículo estudió la actividad de cuentas falsas tras la llegada de Elon Musk, pues la eliminación de estas fue una de sus primeras promesas al comprar la app [75]. Los resultados muestran que no hay descenso en su número y actividad, sino un aumento. El incremento del número de "bots" es peligroso, pues a lo largo de los últimos años han sido usados para influenciar votos, manipular mercados y difundir información falsa [62].

2.2.5. Contribuciones de este trabajo

Los estudios mencionados en la sección anterior (sección 2.2) tienen ciertas limitaciones que se han tratado de cubrir en este proyecto:

- Sesgo lingüistico: se han realizado diversos estudios en inglés, no obstante, esta es la primera vez en la que se estudian las repercusiones de la llegada de Elon Musk en países y comunidades online hispanohablantes.
- Sesgo conceptual (no categorización del discurso de odio): en los artículos mencionados anteriormente el discurso de odio recibía una connotación muy general. Sin embargo, en este proyecto se contempla el discurso de odio como un espectro, dando diferenciando los distintos tipos de odio. Esto es, se han realizado análisis independientes para el contenido "tóxico", "extremadamente tóxico", "ataque a la identidad", "insulto", "amenaza" y "grosería".
- Limitaciones temporales: el trabajo [62] se centra en las semanas próximas a la compra de Twitter por Musk. Sin embargo, el proyecto presentado en este documento expone una visión más global del problema del discurso de odio,

estudiando las consecuencias de los cambios en las políticas a lo largo de los últimos diez años. Resaltando la eficacia de las mismas cuando se han llevado a cabo y denunciando las repercusiones de la eliminación de las mismas tras la llegada de Elon Musk.

- Falta de contexto: El aumento y descenso del contenido objeto de estudio puede deberse a un aumento del contenido en general. Es por ello que los estudios deben tener un punto de referencia con el que comparar. Por ello, se han seleccionado seis conjuntos de datos: tuits sobre el colectivo, tuits tóxicos sobre el colectivo, tuits no tóxicos sobre el colectivo, tuits aleatorios, tuits tóxicos aleatorio y tuits no tóxicos aleatorio. Esto le da una capa extra de calidad al trabajo realizado en [62] que solo compara contenido tóxico sobre el colectivo con el contenido aleatorio.
- Se ha realizado otro estudio específico sobre subcomunidad trans, pues ha sido la mayor perjudicada a los cambios realizados en 2023.

Capítulo 3

Tecnologías y metodología

3.1. Tecnologia

3.1.1. Web scraping

Podemos definir web scraping o "desguace web" como el proceso de extraer y combinar contenidos de internet de la Web de una manera automatizada. En ese proceso, un agente, también conocido como robot web, imita el comportamiento de un humano online. Paso a paso, el robot accede a todas las páginas Web de las que necesite recopilar información, selecciona los datos deseados, los descarga en formatos específicos y los estructura de la manera preestablecida [56].

Una aplicación común del web scraping es la descarga de tuits. En este caso, el agente automatizado imita la acción de un usuario en Twitter, navegando por las publicaciones de la red social. El robot "lee" los tuits que se ajustan a los criterios establecidos, extrayendo y organizando los datos relevantes como el texto del tuit, el número de "likes", los "retuits" y otras métricas asociadas, de manera automatizada.

En este proyecto, se realizaron múltiples intentos de probar distintos repositorios de "web scraping" para Twitter, como snscrape [66], tweepy [7] o twikit [35], lo cual requirió trabajo de prueba y ajuste. No obstante, las diversas modificaciones realizadas en Twitter para evitar este tipo de actividad, hacen que estos repositorios, aunque relativamente recientes, se volvieran obsoletos en un corto periodo de tiempo. Después de un largo proceso de búsqueda y prueba de diversas soluciones, se encontró TweetScraperR, que finalmente resultó ser funcional.

3.1.1.1. TweetScraperR

TweetsScraperR es un paquete de GitHub que "proporciona funciones para extraer datos de X/Twitter, incluidos tweets, usuarixs y metadatos asociados, permitiendo realizar la extracción y manejo de estos datos de manera conveniente en R" [80].

Este repositorio utiliza web scraping para la recolección de tuits, por lo que los datos obtenidos no son tan limpios ni tienen tantos atributos como los extraídos por la API. No obstante, esta es una alternativa flexible, gratuita y de código abierto, lo que la convierte en la mejor opción para la recolección de datos de manera gratuita actualmente.

También se ha usado un cuaderno llamado "TweetScraperR notebook" que facilita la descarga y visualización de los tuits obtenidos con las librerías de TweetScraperR [34].

Este "web scraper" se ha usado en este proyecto para una primera versión de los datasets, que fue muy útil para confirmar que las preguntas de investigación definidas en objetivos (secciones 5.3) parecían ir en la dirección esperada.

3.1.2. API

Una API (Application Programming Interface, en español Interfaz de Programación de Aplicaciones) es una interfaz que permite a dos aplicaciones software comunicarse entre sí para compartir información y funcionalidades [47].

Gracias a las APIs, los desarrolladores pueden interactuar con servicios y aplicaciones complejas sin tener que conocer los detalles internos de su funcionamiento, ya que la API actúa como un intermediario que facilita esta comunicación. De esta manera, las APIs permiten la integración de diversas tecnologías y servicios, haciendo más eficiente el desarrollo de nuevas aplicaciones y funcionalidades.

Puesto que en este trabajo hemos necesitado clasificar los tuits como contenido de odio, se probaron diversas APIs con dicho objetivo. En primer lugar se trató de usar los diversos modelos de BERT de hugging face [1, 2]. Sin embargo, la mayoría funcionan solo para el inglés y aquellos hechos en español carecen de pruebas y uso. En segundo lugar, se trató de usar SocialHaterBert [110], pero su dificultad de uso y la falta de recursos explicativos supusieron el descarte de su uso. Por último, se observó que diversos artículos de investigación [62, 22] usaban la Perspective API de Google y que esta ofrecía su uso para texto en español.

Otra dificultad fue que pese a que TweetScraperR fuera una solución para la descarga de datos, esta sufría alguna carencia. El "desguace web" es una práctica muy útil, pero está sujeta a diversas problemáticas, como paradas abruptas y falta de contenido. Por ello, se trató de buscar una solución alternativa. Lo primero fue

tratar de usar la API de Twitter, pero la versión gratuita es muy limitada y no permite acceder a tuits antiguos. La versión que permite acceder a dichos tuits es de $5000, lo que nos hizo descartar la alternativa. En este contexto, se buscaron datasets públicos en diversas plataformas, como Kaggle y HuggingFace. Ahí se anunciaba la API que finalmente se usó: Twitterapi.io.

3.1.2.1. Perspective API

"Perspective es una API que emplea modelos de aprendizaje automático para asignar una puntuación a cada comentario según la repercusión que podría tener en una conversación" [19]. Para ello, utiliza una serie de atributos que tendrán un valor del 0 al 1, siendo 0 el valor mínimo y 1 el valor máximo. Los atributos, definidos en [20], son los siguientes:

- TOXICITY (toxicidad): Comentarios groseros, irrespetuosos o inadmisibles que podrían incitar a los usuarios a abandonar una conversación.
- SEVERE TOXICITY (toxicidad severa): Comentarios llenos de odio, agresivos, irrespetuosos o que es muy probable que, de alguna manera, hagan que un usuario abandone una conversación o renuncie a expresar su punto de vista. Este atributo es mucho menos sensible a formas más suaves de toxicidad, como los comentarios que incluyen usos positivos de palabras malsonantes.
- IDENTITY ATTACK (ataque a la identidad): Comentarios negativos o llenos de odio dirigidos a una persona debido a su identidad.
- INSULT (insulto): Comentarios insultantes, incendiarios o negativos hacia una persona o un grupo.
- PROFANITY (grosería): Palabras malsonantes de cualquier tipo.
- THREAT (amenaza): Describe la intención de causar dolor, lesionar o actuar con violencia contra un individuo o un grupo.

La Perspective API ha sido de vital importancia en este proyecto, pues ha ayudado a detectar y clasificar el contenido de odio de los diversos tuits.

3.1.2.2. Twitterapi.io

Twitterapi.io es una API externa a Twitter diseñada para facilitar el acceso a datos de Twitter de manera eficiente y sencilla. Permite a los desarrolladores extraer información de tuits, como el número de "likes", "retuits", seguidores y más información relevante [21].

Esta API es de pago, no obstante, su precio es de $0,15 por cada 1.000 tuits, por lo que es una alternativa mucho más económica que la API oficial de Twitter.

En este trabajo se usó para recopilar los tuits utilizados para el estudio.

3.1.3. Jupyter Notebook

Como se define en [65], Jupyter Notebook es una aplicación de código abierto que permite crear y compartir documentos que contienen código, visualizaciones y texto en formato interactivo. Es ampliamente utilizado para análisis de datos, visualización, aprendizaje automático, modelado matemático, entre otras aplicaciones.

Permite ejecutar bloques de código en tiempo real, lo que facilita la experimentación y el análisis interactivo. Jupyter soporta varios lenguajes de programación, como Python, R y Julia, y ofrece una interfaz amigable para integrar código, texto, gráficos y resultados en un solo documento. En este proyecto, el lenguaje utilizado ha sido Python.

Esta tecnología ha sido la herramienta elegida para realizar el análisis estadístico y el análisis de los datos durante todo el proyecto. Ha facilitado el uso de librerías de Python así como la construcción y visualización de gráficas que ayudan a la comprensión del trabajo.

3.1.4. Librerías Python

Se han utilizado las siguientes librerías de Python:

- Pandas: "es una librería de Python especializada en el manejo y análisis de estructuras de datos" [14].
- Matplotlib: "es una librería de Python especializada en la creación de gráficos en dos dimensiones" [12].
- Numpy: "es una librería de Python especializada en el cálculo numérico y el análisis de datos, especialmente para un gran volumen de datos." [13].
- Scipy: "añade una capa a NumPy -sobre la que está construída- ofreciendo funciones científicas y estadísticas por encima de las funciones puramente matemáticas disponibles en ésta" [4].

3.2. Metodología

3.2.1. Modus operandi

Para poder llevar a cabo este proyecto, la alumna ha estado en contacto con el tutor en todo momento. Se han realizado emails periódicos cada semana y videolla-

madas cuando se ha necesitado, para resolver dudas y decidir los distintos caminos que tomaría el trabajo.

Unido a la importante labor realizada por el tutor, este puso en contacto a la alumna con diversos investigadores de ciencias sociales. Esto fue de vital importancia, pues se recibieron consejos que serían decisivos para el resultado final, como por ejemplo el marco temporal del estudio.

Esta base de ayuda fue clave para la elección de las APIs utilizadas y definidas en la sección 3.1. Pues se contrastaron diversas opiniones. Así, se consiguió obtener un corpus de calidad, formado por dos datasets:

- Dataset sobre la comunidad LGTBIQ+: constituido por 653,000 tuits en español, cada uno con 34 atributos. Con un tamaño de 256.7 MB. En el texto se usará para referirse a él "dataset colectivo" o "dataset LGTBIQ+".
- Dataset sobre contenido aleatorio: constituido por 395,000 tuits en español, cada uno con 34 atributos. Con un tamaño de 151.3 MB. En el texto se usará para referirse a él "dataset aleatorio" o "dataset generalista".

Estos datasets están alojados en un repositorio público de ciencia abierta (Zenodo), y con licencia libre Creative Commons Atribución 4.0 Internacional:

https://zenodo.org/records/15639492

Una vez obtenidos los datos, debían ser procesados por la Perspective API, para lo que se utilizó Python como herramienta. Una vez asignada una calificación de toxicidad, se realizó un análisis estadístico de los datos, así como distintas gráficas para su visualización. Para esto último, se usaron dos Jupyter Notebooks. Todo el código se puede encontrar en un repositorio de GitHub, con licencia de software libre MIT:

https://github.com/MarRamiro/Analisis-Twitter

3.2.2. Extracción de tuits

En primer lugar hemos hecho un glosario de palabras en español que suelen ir ligadas a contenido LGTBIQ-fóbico, que se pueden encontrar en la tabla A.1 del Apéndice A. Estas han sido extraídas del Glosario de Términos LGBT del Proyecto Rainbow, cofinanciado por el Programa de Derechos Fundamentales y Residencia de la Unión Europea [89]. Se han añadido también otros términos que no están definidos en el glosario pero que se consideraron necesarios, como por ejemplo "mariquita", "asexual" o "tortillera".

Para hacer una primera recolección de estos datos se usó TweetScraperR, lo que permitió hacer un pequeño análisis que mostró que las hipótesis iban en buena dirección. Una vez observada una clara tendencia de incremento del contenido objeto de estudio, se hizo una recolección de datos más extensa, asegurando así la obtención de todos los tuits en el período deseado. Por ello, en la recolección final se usó Twitterapi.io, ya que este ofrece más información de cada tuit, y es más veloz. Cada tuit tiene 34 atributos, que están especificados en el Apéndice C.

Puesto que se quiere que haya suficiente contenido sobre el colectivo, los datos son recolectados de todos los 28 de junio de cada año, día en el que se celebra el Día Internacional del Orgullo LGTBIQ+. Además, para que el período temporal sea suficientemente relevante, se han recolectado los datos desde 2015 hasta 2024.

El aumento de la LGTBIQ-fobia podría ser un resultado de un aumento del contenido en general. Para demostrar que este no es el caso, se ha recolectado otro dataset que pretende representar cómo ha cambiado la actividad de Twitter a lo largo del tiempo. Para ello se han recolectado los tuits que contienen tres de las palabras más comunes del español: "ser", "todos" y "algo" [18]. En este dataset, y para tener un corpus de suficiente tamaño, se han recolectado todos aquellos tuits que se hayan publicado durante los mismos seis minutos aleatorios para cada hora de cada 28 de junio de cada año.

La recolecta de tuits por palabras clave no es el único método de proceder. De hecho, este método tiene algunas limitaciones, como pérdida de tuits relevantes por no contener dichas palabras, así como la captura de otros irrelevantes para el proyecto (que sí contengan palabras clave). Hay otros métodos de proceder que también son válidos y cubren algunas de las limitaciones, como por ejemplo descargar por hashtags o descarga de los tuits de cuentas específicas.

3.2.3. Estudio de toxicidad

Una vez recolectados los datos, se procesaron con la Perspective API para asignar a cada tuit un valor de toxicidad, toxicidad severa, ataque a la identidad, insulto, grosería y amenaza. Cada valor es un número entre 0 y 1: si es más cercano a cero, significa que ese atributo no se da en el texto. Por el contrario, si el número es cercano a 1, significa que ese atributo está muy presente en el texto. Por ejemplo, para el atributo toxicidad, un valor cercano a cero indicará que el tuit es poco tóxico, y un valor cercano a 1 indicará que el tuit es muy tóxico.

Esta API ha sido validada y usada por distintas investigaciones, y establecen como umbral para considerar un tuit su atributo en 0,7, lo que coincide con lo establecido en la documentación de la API [51, 97, 62, 19]. Es decir, para el atributo toxicidad, se considerarán tóxicos todos aquellos tuits cuyo valor para el atributo "toxicidad" sea mayor a 0,7.

3.2.4. División en periodos

Aunque el objetivo principal es el estudio del aumento antes y después de la adquisición de Musk, no se pueden tomar todos los años antes de Elon Musk como un mismo período, ya que durante esos años ha habido también un significativo cambio de las medidas de moderación del lenguaje de la aplicación.

Tal y como se ha expuesto en el apartado 2.2.3 de este trabajo, Twitter hizo el primer comunicado contra el delito de odio y las políticas de actuación en 2017. No obstante, dichas políticas no fueron realmente implementadas hasta 2020. Es por ello que en el análisis de este proyecto, los datos se agrupan en los siguientes periodos:

- 2015 a 2019, representando el contenido previo al establecimiento de políticas explícitas sobre contenido de odio.
- 2020, como caso aislado, pues durante la pandemia la actividad en las redes sociales aumentó aproximadamente un 27 % [44]. De hecho, este aumento fue aún más exagerado durante el 28 de junio, pues el Día del Orgullo, que año a año une a miles de personas en las grandes ciudades, fue celebrado online [33].
- 2021-2022, tras la ejecución de medidas contra el discurso de odio que se establecieron en diciembre de 2020 [113].
- 2023-2024, que marca la nueva era de Twitter, la era Musk, y marca las consecuencias de la eliminación de ciertas medidas para la moderación del discurso de odio.

3.2.5. Análisis de datos

Una vez recolectados los datos, se usaron varios archivos Jupyter Notebook para realizar gráficas y estudios estadísticos. Para resolver este problema de manera más eficiente, se ha hecho uso complementario de la inteligencia artificial generativa ChatGPT como asistente, con el objetivo de aumentar la comprensión y verificación del trabajo desarrollado. Cuando se ha usado, ha sido poniendo énfasis en la posibilidad de errores y alucinaciones que pueda cometer, y en la generación de código, adaptando y entendiendo el código antes de usarlo.

En algún estudio se ha separado la comunidad trans del resto del colectivo, puesto que los recientes cambios de políticas en la moderación del contenido de odio en Twitter afectan en especial a este subgrupo del colectivo [39]. Para realizar esto se han considerado los tuits del dataset del colectivo que poseen las palabras de la tabla A.2 del Apéndice A.

Para estudiar el aumento del contenido de odio, se ha observado la cantidad absoluta de tuits tóxicos hacia el colectivo, así como su cantidad relativa al número total de tuits. Para el aumento de la visibilidad, se ha tomado como medida de

referencia la cantidad de "likes" y "retuits", pues el "número de visualizaciones" no existió hasta 2022. Además, esta correlación ha sido probada por otros estudios anteriormente [62].

Por último, se ha estudiado la significancia estadística de los resultados, usando la prueba de Kolmogorov-Smirnov, que nos dirá si dos distribuciones son independientes. Además, se han estudiado los percentiles, la media y la moda de los distintos datasets.

Los estudios concretos han sido los siguientes:

- Cantidad de tuits tóxicos a lo largo del tiempo: muestra la cantidad de tuits en contra del colectivo. Un aumento supone una mayor urgencia en la toma de medidas para la moderación del discurso.
- Variación relativa de odio al colectivo con respecto al total: manifiesta la visión general de los usuarios de Twitter acerca del colectivo. Un aumento supone que cada vez la visión de los usuarios sobre el colectivo es más negativa.
- Media del valor de toxicidad a lo largo del tiempo: exhibe el tono usado para tratar temas relacionados con la comunidad LGTBIQ+. Un aumento supone una mayor tensión, es decir, indica que el tema se ha vuelto más polémico.
- Media de "likes" y "retuits" en tuits tóxicos a lo largo del tiempo: Esto puede deberse a distintos motivos, como una menor diversidad de pensamientos o un aumento de la visión negativa hacia el colectivo. Un aumento del valor de estos atributos expone un incremento de la interacción de este contenido, y por tanto niega que tenga menos visibilidad.
- Agregación de "likes" y "retuits" de tuits tóxicos a lo largo del tiempo: revela la visibilidad del contenido de odio hacia el colectivo como un conjunto. La suma mostrará el alcance global que ha tenido el tuit, es decir, la interacción total que ha habido con tuits de odio hacia el colectivo.
- Número de tuits tóxicos de usuarios "Verificados Azules": advierte el impacto que tiene la nueva manera de obtener el "verificado azul", es decir, presenta las consecuencias de hacer que la visibilidad y el "distintivo de veracidad" sean de pago.

Todo esto ha sido estudiado en primer lugar de una manera más general a lo largo de los años, para tener así una visión global de la situación. A continuación, se ha realizado el mismo estudio pero para los distintos periodos separados, indicados en el apartado 3.2.4.

Estos estudios se han realizado para los siguientes conjuntos de datos:

- Dataset de contenido sobre el colectivo LGTBIQ+ (653.000 tuits). Para referirse a este se usarán los terminos: colectivo o LGTBIQ+.

- Datasets de contenido de odio sobre el colectivo LGTBIQ+ (aquellos tuits del dataset del colectivo cuyo atributo Toxicidad $> 0{,}7$): 49.000 tuits tóxicos, 1.600 muy tóxicos, 3.800 de ataque a la identidad, 45.000 de insultos, 600 de amenazas y 44.000 de grosería).
- Datasets de contenido neutro sobre el colectivo LGTBIQ+ (aquellos tuits del dataset del colectivo cuyo atributo Toxicidad $< 0{,}05$): 145.000 tuits no tóxicos, 483.000 no muy tóxicos, 193.000 sin ataque a la identidad, 218.000 sin insultos, 617.000 sin amenazas y 236.000 sin grosería.
- Dataset de contenido aleatorio en español: 395.000 tuits. Para referirse a este se usarán los términos: aleatorio, generalista.
- Datasets de contenido tóxico aleatorio en español (aquellos tuits del dataset del aleatorio cuyo atributo Toxicidad $> 0{,}7$): 9.000 tuits tóxicos, 191 muy tóxicos, 135 de ataque a la identidad, 10.000 de insultos, 381 de amenazas y 13.000 de grosería).
- Datasets de contenido no tóxico aleatorio en español (aquellos tuits del dataset del aleatorio cuyo atributo Toxicidad $< 0{,}05$): 247.000 tuits no tóxicos, 370.000 no muy tóxicos, 361.000 sin ataque a la identidad, 284.000 sin insultos, 383.000 sin amenazas y 300.000 sin grosería.

Para el análisis estadístico se han observado las distribuciones de los distintos datasets descritos con respecto a ciertos atributos, como la toxicidad o los "likes". Además, se ha tratado la independencia de los mismos por medio del test de Kolmogorov-Smirnov, definido en la sección 2.1.3.3. En este estudio rechazaremos la hipótesis nula si el p-valor es menor a 0.01, siguiendo lo realizado por artículos similares [62].

Capítulo 4

Resultados y su análisis

4.1. Análisis estadístico

Para poder sacar conclusiones de los datos, será necesario explorar la distribución de los mismos con respecto a los distintos atributos objeto de estudio. Es por ello que, en esta sección se van a usar diversos tests estadísticos sobre los atributos "likes", "retuits" y "toxicidad". Si se quisiera demostrar lo mismo con respecto al resto de atributos de medición de discurso de odio, como pueden ser el átaque a la identidadó la 'toxicidad extrema', bastaría con repetir las pruebas realizadas.

4.1.1. Distribuciones de toxicidad

4.1.1.1. Dataset del colectivo

En primer lugar, observamos la distribución del atributo toxicidad a lo largo de los años en la figura 4.1. Cada gráfico muestra cómo se distribuyen los datos respecto al atributo toxicidad, es decir, muestra la densidad a lo largo del eje X. La altura de la curva en el eje Y indica cuán comunes son esos valores. Es decir, si la curva sube mucho en un área, significa que esos valores son más frecuentes en los datos. Además, el área total debajo de la curva es igual a 1, lo que representa toda la distribución de los datos. Es por ello que lo relevante de este gráfico no son los valores del eje Y, sino el área bajo la curva.

En este contexto, el descenso del área en los valores bajos de toxicidad a lo largo de los años muestra una menor frecuencia de tuits no tóxicos. Esa pérdida de área debe verse reflejada en un aumento de la misma en otra parte del gráfico (recordemos que el área total es uno siempre). Este es el motivo del engruesamiento del área en valores próximos a toxicidades extremas en los últimos años.

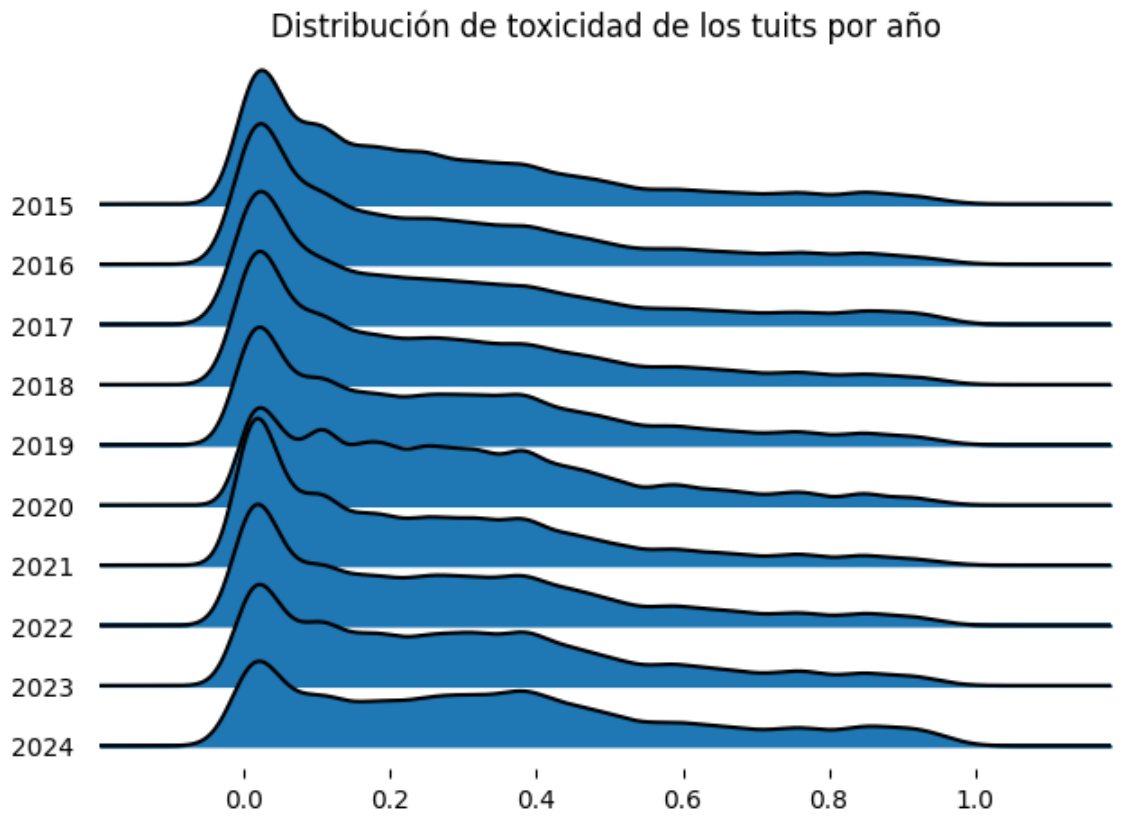

Figura 4.1: Distribuciones por años - dataset colectivo tóxico

Puesto que la figura 4.1 trata de mostrar la densidad de valores, es difícil observar las diferencias en valores extremos. No obstante, nótese que el área en cualquier intervalo del eje X nos dice qué porcentaje del total de los datos está en ese intervalo. Es decir, el área azul representada entre los valores 0.7 y 1 del eje X, equivale aproximadamente a la cantidad de tuits que poseen esos valores para el atributo toxicidad, dividido entre el total de tuits (para cada año). Es por ello que en la sección 4.2.2, concretamente en la figura 4.12 se observará mejor ese aumento de densidad o porcentaje para los tuits tóxicos (toxicidad $> 0{,}7$).

Si volvemos a examinar la figura 4.1 notamos que 2020 es un año atípico, así como la similitud entre ciertas distribuciones. Las gráficas de 2015 a 2019 muestran una tendencia semejante. Por otro lado, las densidades de 2021 y 2022 se asemejan mutuamente en estructura. Por último, las distribuciones de 2023 y 2024 anuncian una bajada en el número de tuits no tóxicos y una subida de los tóxicos de la misma manera. Esta observación es de especial relevancia, pues no sólo nos aporta información de las distintas tendencias del contenido a lo largo de los años, sino que nos reafirma en la división por periodos establecida en el apartado 3.2.4.

Tiene sentido pues, observar los cambios en las distribuciones por periodos. Es por ello que en las próximas gráficas de esta sección vamos a estudiar los datos por dichos intervalos de tiempo.

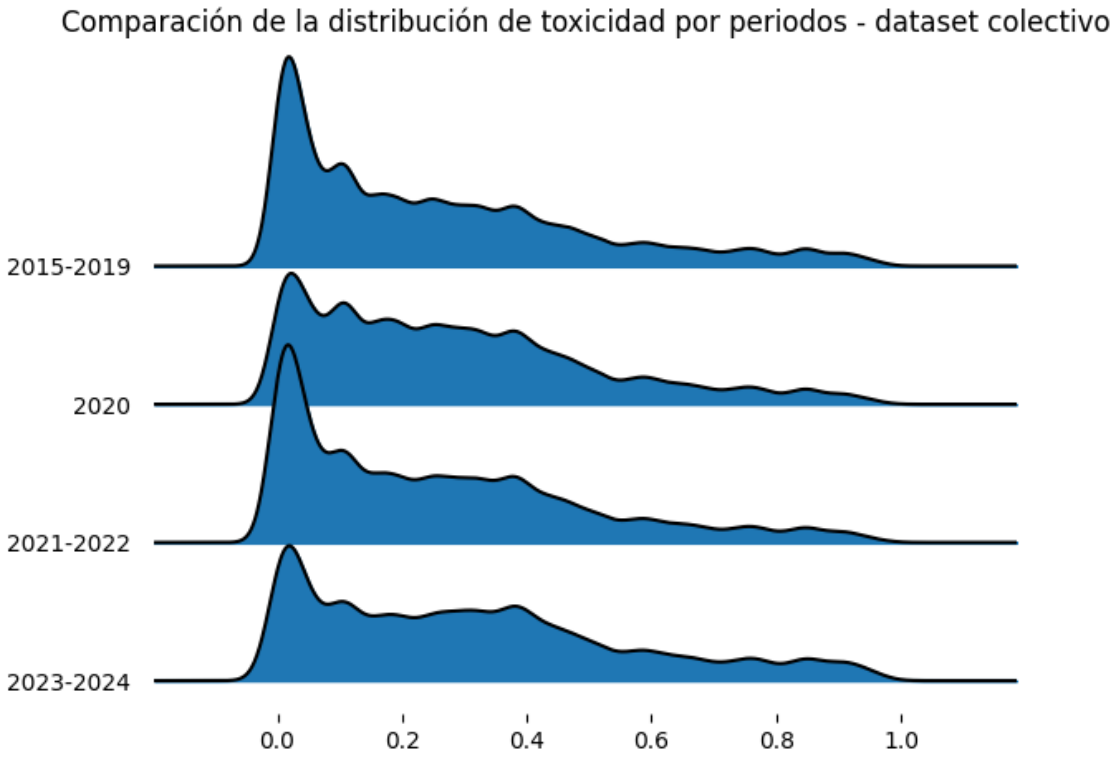

Figura 4.2: Distribuciones por periodos del dataset del colectivo tóxico

En la figura 4.2 se aprecian de manera más clara las diferencias para las distintas etapas: pre-pandemia, pandemia, post-pandemia y post-Musk.

Si nos fijamos en la cola de la distribución, la fase post-Musk es claramente más gruesa que el resto. No obstante, de igual manera que para las distribuciones por año realizadas en la figura 4.1 y que se han mencionado anteriormente, lo importante de estas gráficas es que el área total es uno. Por ello, un descenso del área en valores bajos de toxicidad viene dado por un aumento de la frecuencia o porcentaje en el resto de valores del eje X.

Atributo	2015-2019	2020	2021-2022	2023-2024
Cuartil 25 %	0.05	0.10	0.05	0.10
Mediana (50 %)	0.20	0.25	0.20	0.28
Cuartil 75 %	0.40	0.41	0.40	0.46
Percentil 90 %	0.63	0.61	0.60	0.71
Percentil 95 %	0.77	0.75	0.75	0.84
Percentil 99 %	0.92	0.90	0.90	0.93
Moda	0.25	0.25	0.25	0.38
Media	0.26	0.28	0.26	0.31

Tabla 4.1: Porcentaje de aumento en número de tuits del dataset colectivo entre los diferentes periodos

Las conclusiones que se derivan de las distribuciones, cuyos atributos están defi-

nidos en el apartado 2.1.3.2, son más claras si observamos los percentiles de la tabla 4.1. Un ejemplo de uso, para quien no esté familiarizado/a con este concepto: para 2020 el percentil 90 es 0.61, esto quiere decir que el 90 % de los tuits tienen un valor menor a 0.61 para el atributo toxicidad.

Los valores del cuartil 25 son 0.05 para los periodos pre-pandemia y post-pandemia y 0.10 para los períodos de pandemia post-Musk. Esto nos indica un menor número de tuits no tóxicos para los datasets de 2020 y 2023-2024. Otro ejercicio interesante consiste en observar la diferencia que hay entre el periodo post-Musk en comparación con el resto de periodos para los percentiles 90, 95 y 99, lo que muestra una mayor cantidad de valores extremos. Para el percentil 90 el dataset diferenciado muestra que el valor ya está por encima de 0.71, es decir, al menos el 10 % (los datos no incluidos en el percentil 90 representan el 10 % restante) son considerados tóxicos. Sin embargo, el resto de percentiles 90 rondan entre 0.60 y 0.63, lo que muestra un claro aumento del contenido tóxico.

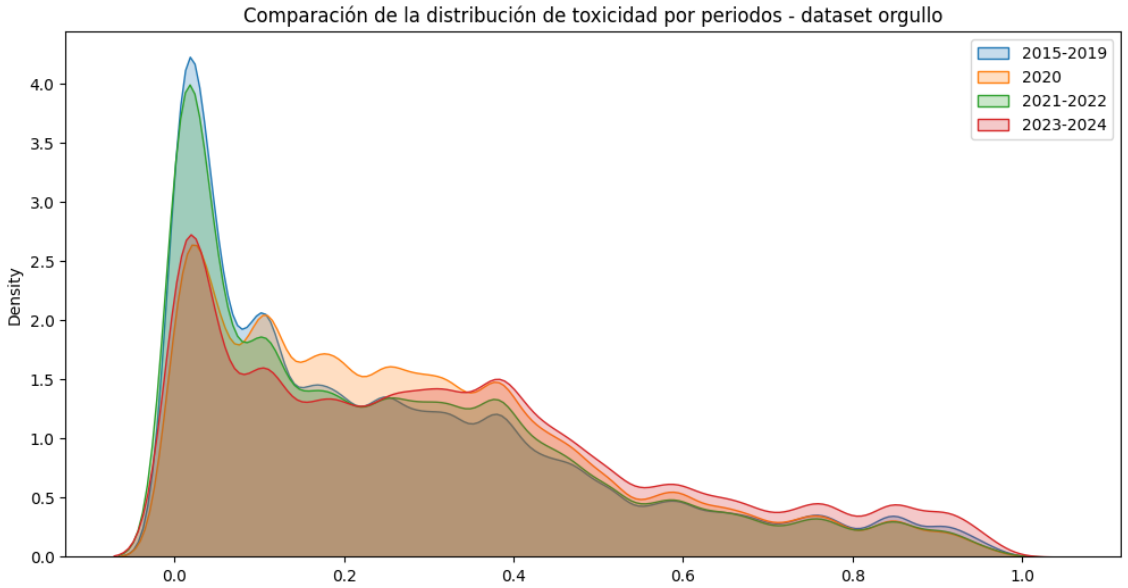

Figura 4.3: Distribuciones por periodos del dataset del colectivo tóxico superpuestas

La simulitud en el cuartil 25 (0.1 en ambos) que se produce entre las épocas de pandemia y post-Musk sugieren que el número de tuits no tóxicos ha disminuido en ambos, lo que está representado con la disminución del área del gráfico 4.2 en la parte de pico. Sin embargo, obsérvese la diferencia en los percentiles 90 (0.61 para 2020 y 0.71 para 2023-2024) y el 95 (0.77 vs 0.84). Esto evidencia que el aumento del área se refleja en los valores de toxicidad medios en el caso del 2020 y el aumento de los valores extremos en el caso de 2023-2024. De hecho, podemos verificar esto de manera visual superponiendo las gráficas de la figura 4.2 en una misma gráfica, veáse en la figura 4.3. El eje Y en el gráfico representa la densidad de los datos, lo que indica la frecuencia relativa o la probabilidad de encontrar valores dentro de un intervalo específico en el eje X.

Para estudiar la posibilidad de que las diferencias observadas se hayan dado por

Comparación	p-valor
2015-2019 vs 2020	<0.01
2015-2019 vs 2021-2022	<0.01
2015-2019 vs 2023-2024	<0.01
2020 vs 2021-2022	<0.01
2020 vs 2023-2024	<0.01
2021-2022 vs 2023-2024	<0.01

Tabla 4.2: Resultados de las comparaciones entre diferentes períodos con sus p-valores para el dataset colectivo

azar, necesitamos estudiar el p-valor, que tendrá un valor entre 0 y 1. Si el valor es < 0.01 podremos afirmar con suficiente confianza que los datos comparados son independientes. Para ello vamos a usar la prueba de Kolmorov-Smirnov, explicada en el apartado 2.1.3.3, que se centra en ver si las distribuciones de dos muestras son estadísticamente diferentes. Con este objetivo obtenemos la tabla 4.2.

Como todos los valores son menores a 0.01, podemos concluir que los diferencias en las distribuciones obtenidas en esta sección están estadísticamente probadas. Dicho de otra manera, la probabilidad de que dichas diferencias sean una casualidad es lo suficientemente baja.

4.1.1.2. Dataset aleatorio

Se va a realizar el mismo estudio para el dataset aleatorio, aunque de manera menos extensa, pues los diversos conceptos importantes de las gráficas y tablas ya han sido explicados en el apartado anterior (apartado 4.1).

Se comienza observando sus gráficas de densidad en la figura 4.4.

Es inmediata la observación de que el período menos tóxico es el de 2015-2019, pues el área acumulada es mayor para los valores bajos del eje X. Además, al igual que para el dataset del colectivo, este estudio muestra un descenso del contenido no tóxico y un aumento del mismo durante la era post-Musk. Esto se debe a que el descenso del área en la parte de contenido menos tóxico supone un aumento del área para valores más altos del atributo toxicidad. De hecho, podemos superponer las gráficas, obteniendo la figura 4.5. El eje Y en el gráfico representa la densidad de los datos, lo que indica la frecuencia relativa o la probabilidad de encontrar valores dentro de un intervalo específico en el eje X.

Para ver los cambios en la distribución de manera más clara, se muestra la tabla 4.3. Si nos fijamos en los percentiles 90, 95 y 99, estos son menores en general para los períodos esperados, pre-pandemia y post-pandemia, y de igual manera la era post-Musk es más tóxica. No obstante, este dataset muestra valores menos alarmantes que el del colectivo, estudiados en la sección 4.1 en todos los intervalos de tiempo, incluido el post-Musk. De hecho, si comparamos los percentiles 90 y 95, la tabla 4.3

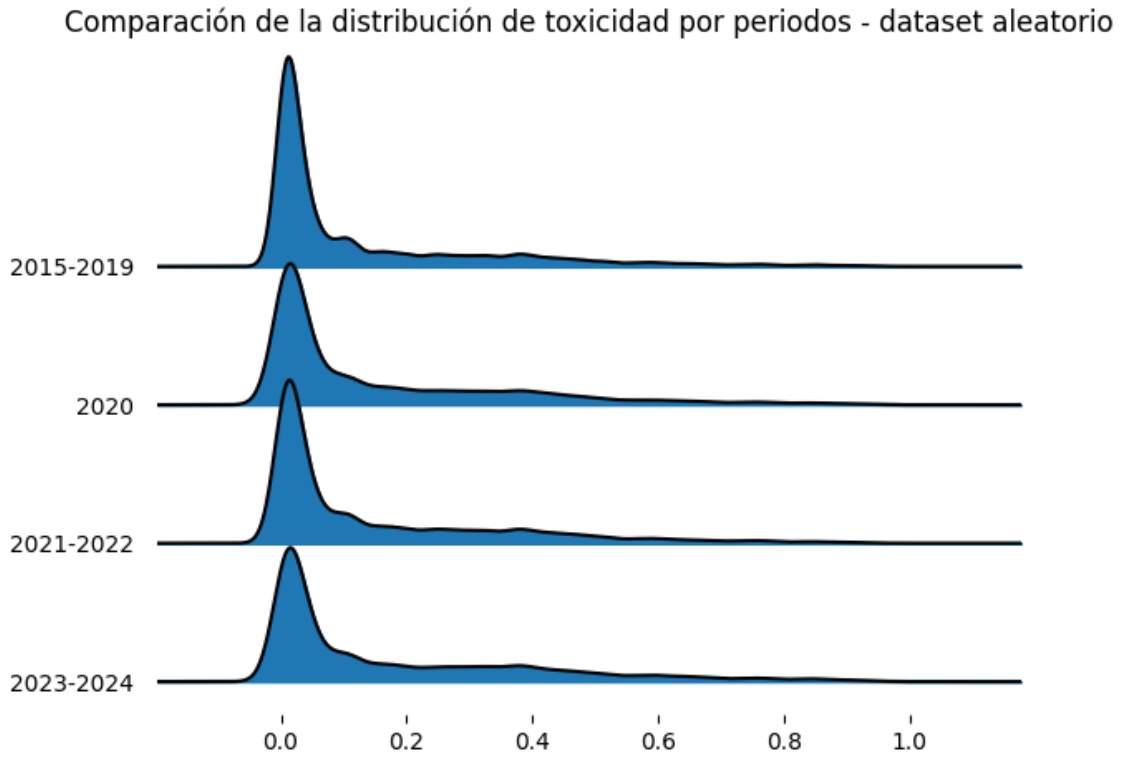

Figura 4.4: Distribuciones por períodos del dataset aleatorio tóxico

(dataset aleatorio) muestra valores 0.5 y 0.64 respectivamente, mientras que la tabla 4.1 contiene valores de 0.71 y 0.84 respectivamente.

Para estudiar la independencia de los datos sacamos el p-valor, tal y como se ha indicado en el apartado anterior, un p-valor menor a 0.01 supone que la probabilidad de que los resultados se deban al azar sea menor al 1 %.

Tal y como se muestra en la tabla 4.4 todos los p-valores son menores que 0.01, luego podemos afirmar que las diferencias observadas entre los datasets son reales con una probabilidad del 99 %.

Por último, se ha comprobado que ambos datasets sean independientes entre sí, también por medio del test de Kolmogorov-Smirnov obteniendo un p-valor de 0. Luego las comparaciones que hagamos entre ellos tendrán solo un 1 % de probabilidad de ser casualidad.

4.1.2. Distribuciones de "likes" y "retuits"

En la sección 4.1.1 se ha visto en profundidad el análisis estadístico de los datos para el atributo toxicidad. Puesto que el modo de proceder es el mismo, el estudio estadístico de los "likes" y "retuits" se va a ver de manera menos detallada. No obstante, si el lector/a quisiera ver los datos estadísticos concretos, estos se encuentran en el Apéndice B.

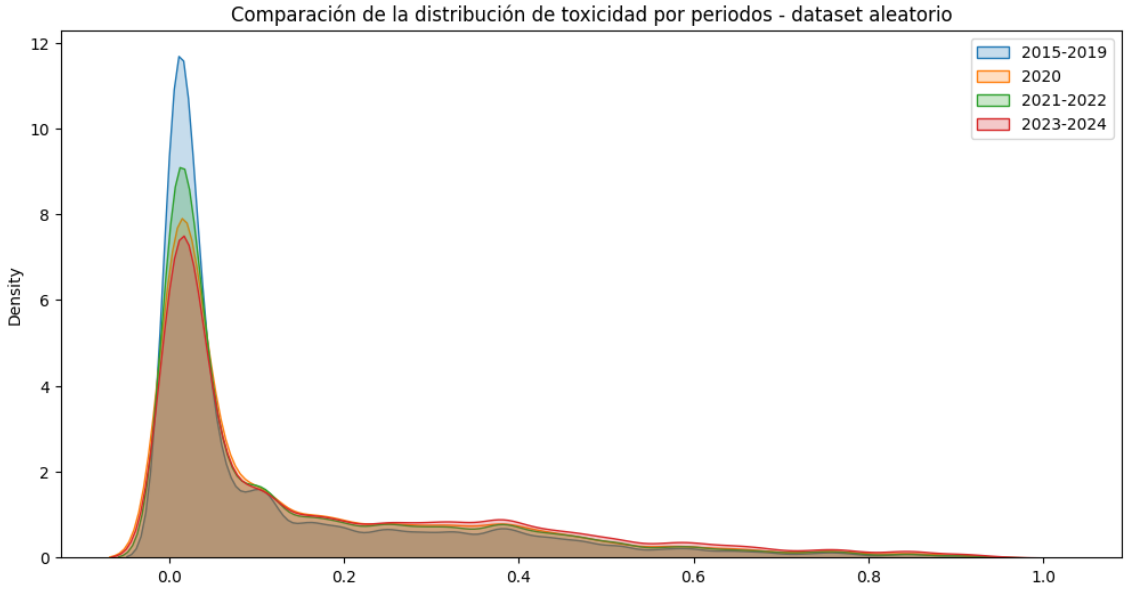

Figura 4.5: Distribuciones por períodos del dataset aleatorio tóxico superpuestas

Atributo	2015-2019	2020	2021-2022	2023-2024
Cuartil 25 %	0.01	0.01	0.01	0.01
Mediana (50 %)	0.03	0.05	0.04	0.06
Cuartil 75 %	0.17	0.25	0.23	0.30
Percentil 90 %	0.40	0.44	0.44	0.50
Percentil 95 %	0.52	0.59	0.57	0.64
Percentil 99 %	0.79	0.82	0.79	0.85
Moda	0.25	0.25	0.25	0.38
Media	0.12	0.15	0.14	0.17

Tabla 4.3: Resultados de las estadísticas descriptivas para cada periodo - dataset aleatorio

El objetivo de este estudio es ver si las distribuciones de los distintos datasets son parecidas o no. Como vemos, la mayoría de los valores son menores a 0.01; luego podremos concluir que las distribuciones son independientes con probabilidad del 99 %. Tal y como se ha explicado en el apartado 2.1.3.3, el hecho de que el p-valor sea mayor que 0.01 para alguna comparación no quiere decir que los estudios que realicemos posteriormente en 4.2.5 como la suma de "likes" no sean válidos. Lo que nos dice es que hay más probabilidad de que las distribuciones se parezcan.

Un dato curioso, que podemos observar en la tabla B.13, es que al comparar la distribuciones con respecto a los "likes" del dataset tóxico aleatorio con el dataset tóxico del colectivo es de 0.19, y de 0.99 para los "retuits". Esto nos sugiere que pese a haber mayor porcentaje de tuits tóxicos hacia el colectivo (se verá en la sección 4.2.2), el impacto de cada tuit es parecido al de los tuits tóxicos hacia otras minorías.

Las distribuciones de los "likes" y "retuits" para los dos datasets antes de filtrar por toxicidad, sí son distintas a las distribuciones de los datos tóxicos. Es decir, los

Comparación	p-valor
2015-2019 vs 2020	<0.01
2015-2019 vs 2021-2022	<0.01
2015-2019 vs 2023-2024	<0.01
2020 vs 2021-2022	<0.01
2020 vs 2023-2024	<0.01
2021-2022 vs 2023-2024	<0.01

Tabla 4.4: Resultados de las comparaciones entre diferentes períodos con sus p-valores - dataset aleatorio

datasets del colectivo y aleatorio se distribuyen con respecto a los "likes" y "retuits" de manera distinta a los tóxicos. Véase la tabla B.13.

4.2. Resultados

4.2.1. Cantidad de tuits tóxicos a lo largo del tiempo

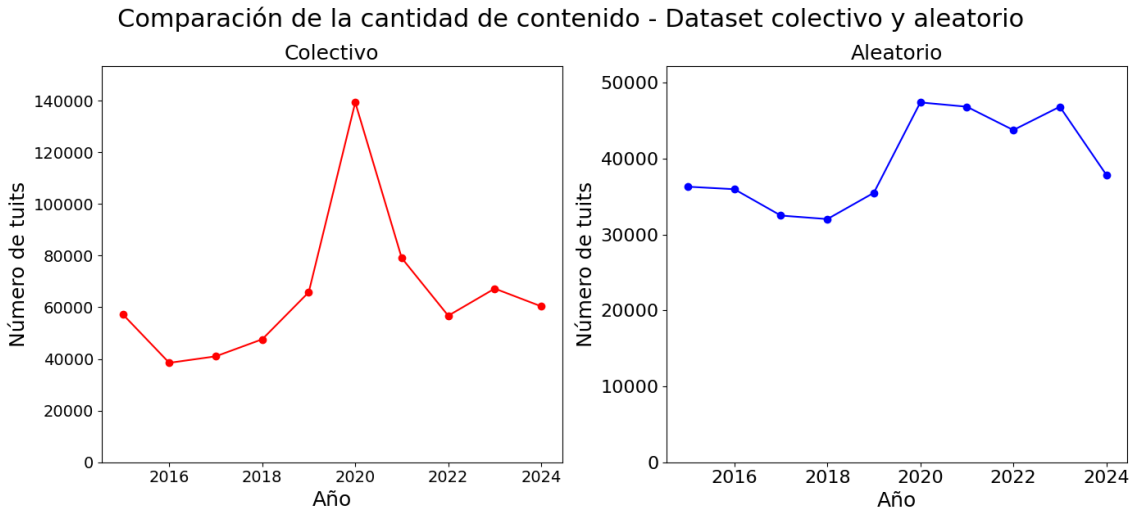

Figura 4.6: Comparación de la cantidad de contenido - dataset colectivo (gráfica de la izquierda) y aleatorio (gráfica de la derecha)

Comenzamos estudiando la cantidad de tuits que hay en los datasets del orgullo y de la comunidad LGTBIQ+ en la figura 4.6, llegando a las siguientes observaciones:

- 2020 es un año atípico en el contenido del colectivo, lo cual tiene sentido, pues el Orgullo se celebró online. Este pico se va a ver reflejado en otros estudios sobre este dataset, como la cantidad de discurso de odio, que muestra valores muy altos para 2020 en la figura 4.7. No obstante, esta cresta se evitará en los apartados que estudian porcentajes y medias relativas.

- Los primeros años se publicaba menos contenido en general (gráfica de la derecha) pero también mucho menos sobre el colectivo (gráfica de la izquierda). La pandemia aumentó el tiempo de uso del móvil y de las redes sociales, por lo que hay más actividad en ambos datasets. Pasada la pandemia, hay menos contenido sobre el colectivo pero el contenido general se mantiene alto. Por útimo, 2024 muestra una bajada en el uso de Twitter para ambos datasets.

Esta bajada entra en contraste con la varianza del contenido de discurso de odio sobre el colectivo LGTBIQ+, pues en la figura 4.7 observamos que la tendencia del contenido tóxico en los últimos años es ascendente. Sobre todo para los atributos "muy tóxico" y "ataque a la identidad", lo que es altamente preocupante, pues indica un descenso de la moderación del contenido lleno de odio y que ataca directamente a la identidad de las personas.

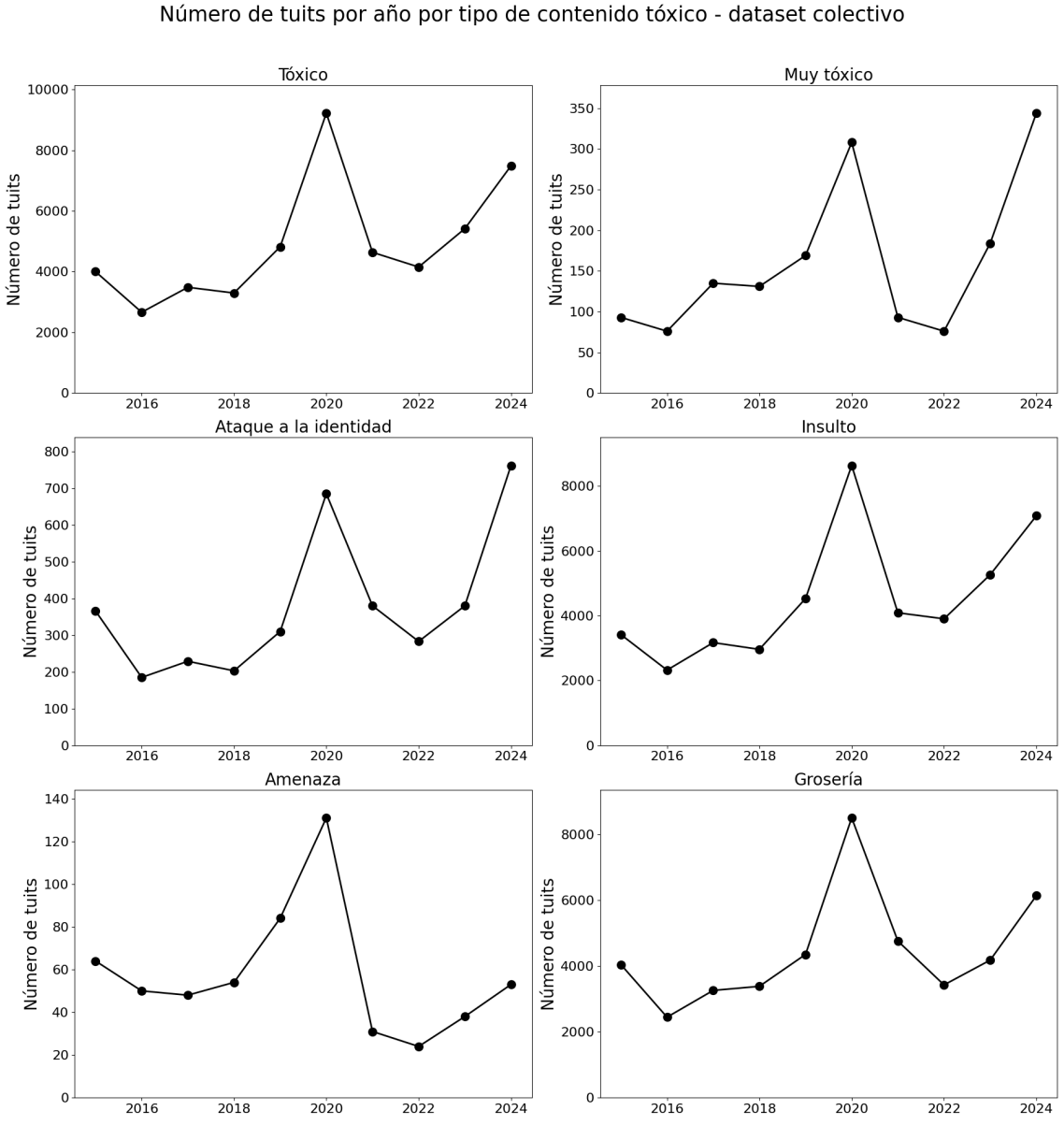

Figura 4.7: Comparación de la cantidad de contenido de odio - dataset colectivo

Unido a este aumento, en la figura 4.8 observamos un descenso de los tuits neutros sobre el colectivo. Esto es una muestra de las consecuencias de una exposición prolongada a contenido de odio, una desensibilización del tema, reduciendo la neutralidad de las opiniones y favoreciendo un ambiente tóxico. Tiene sentido que en este contexto aquellos usuarios identificados con la comunidad LGTBIQ+ migraran a otras aplicaciones en las que se sintieran más seguras. Esto podría provocar un futuro descenso aún más abrupto de usuarios en los años próximos.

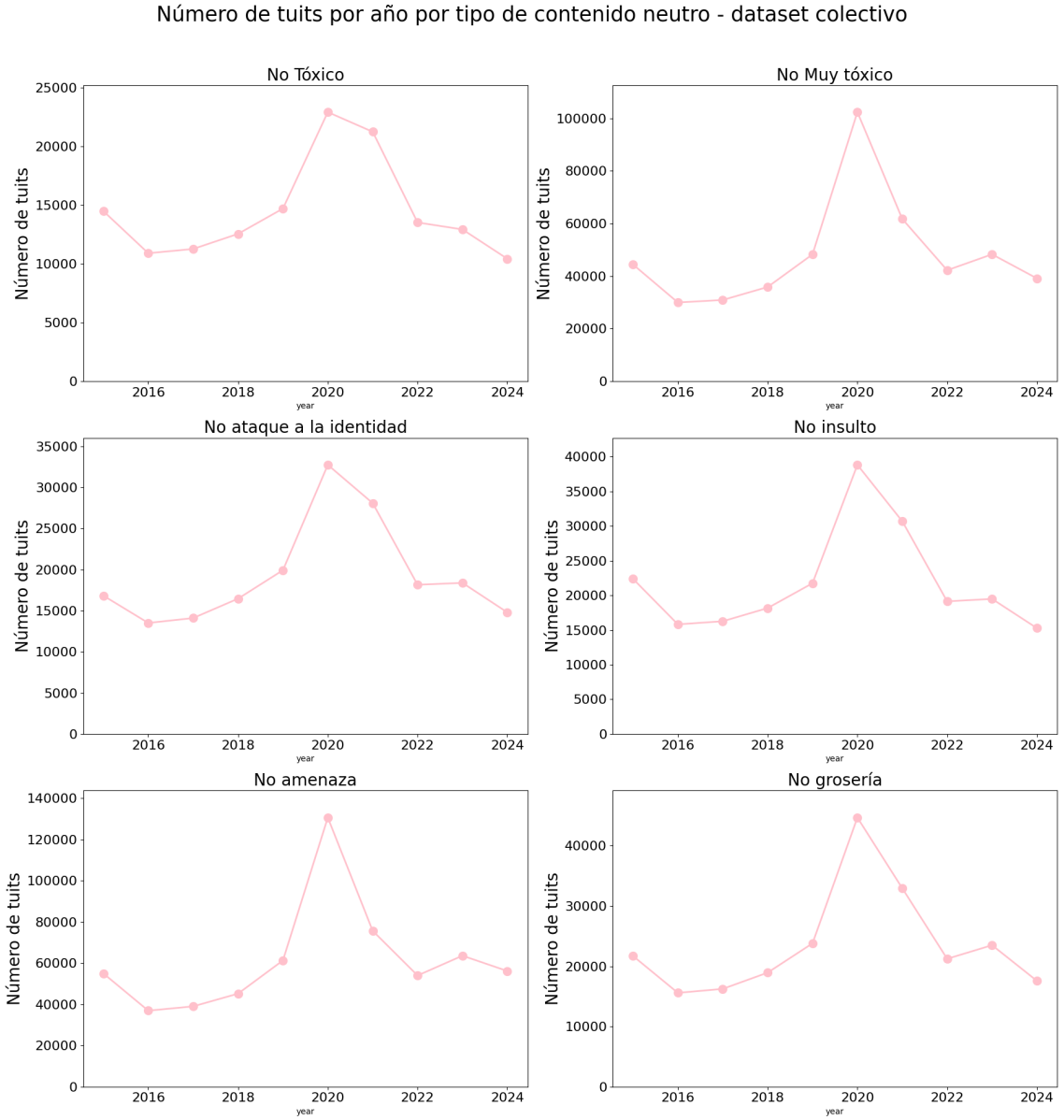

Figura 4.8: Comparación de la cantidad de tuits neutros - dataset colectivo

De hecho, el aumento de toxicidad no se produce solo sobre la minoría objeto de estudio en este trabajo; en la figura 4.9 observamos, aunque en menor medida, un aumento del discurso de odio en el contenido aleatorio en general, además de un descenso del contenido neutro en la figura 4.10, convirtiendo Twitter no solo en

un lugar hostil para usuarios LGTBIQ+, sino también para el resto de usuarios. De hecho, las gráficas con mayor pendiente en los últimos años vuelven a ser las de contenido "muy tóxico" y "ataque a la identidad".

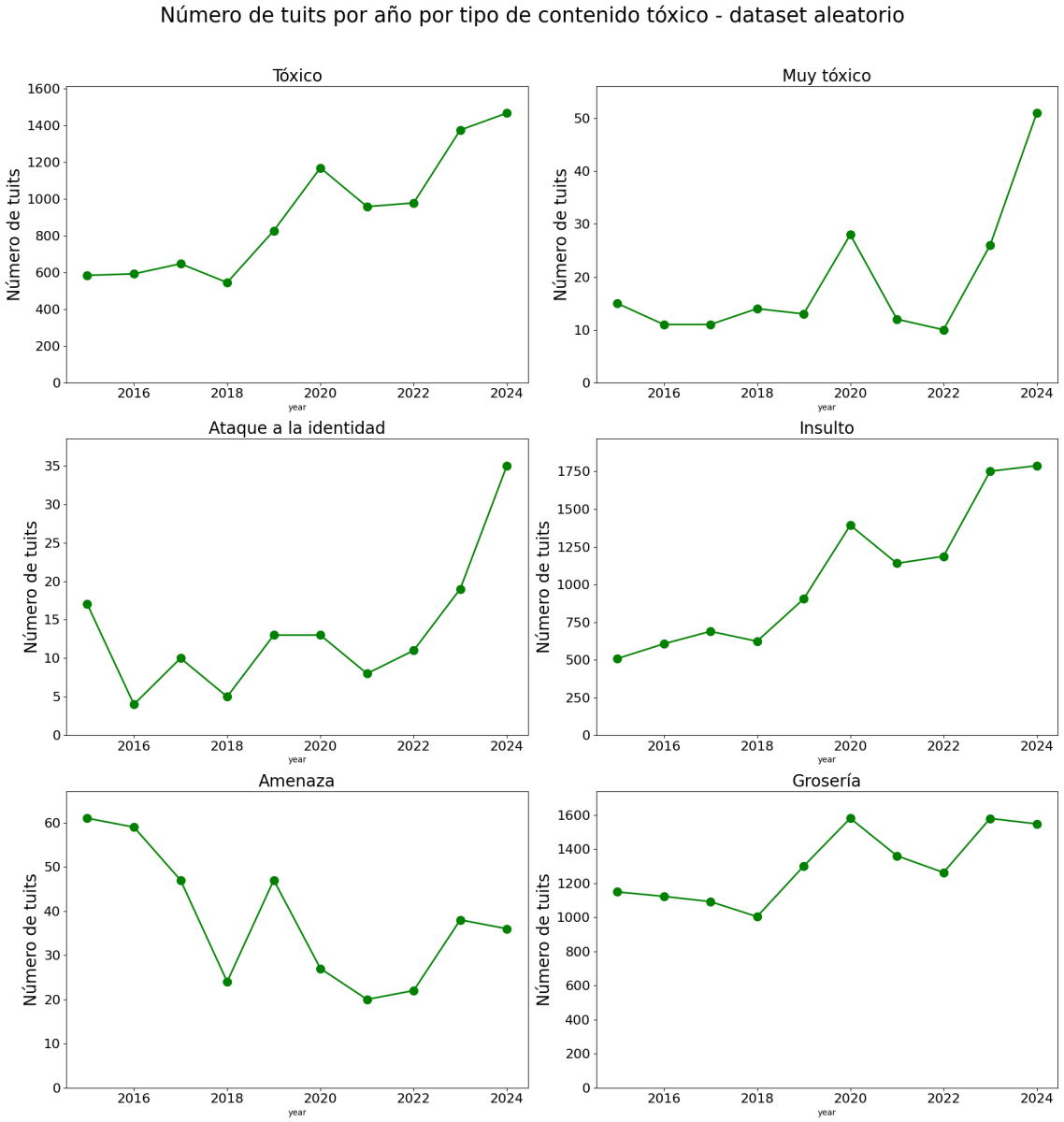

Figura 4.9: Comparación de la cantidad de contenido de odio - dataset aleatorio

Observamos que las gráficas de contenido tóxico hacia el colectivo y las del contenido tóxico en general crecen de manera parecida en 2023 para los atributos "tóxico", "insulto", "amenaza" y "grosería". Sin embargo, la toxicidad en contenido aleatorio (figura 4.9) se mantiene estable de 2023 a 2024 mientras que la toxicidad hacia el colectivo (figura 4.7) sigue creciendo en este último año. Este aumento de toxicidad hacia el colectivo plantea la posibilidad de que de 2023 a 2024 se hayan eliminado algunas políticas de detección y eliminación de contenido de odio hacia el colectivo, o que el clima sea más hostil frente a éste debido a otros factores (e.g. desensibilización,

cambio de población de la plataforma).

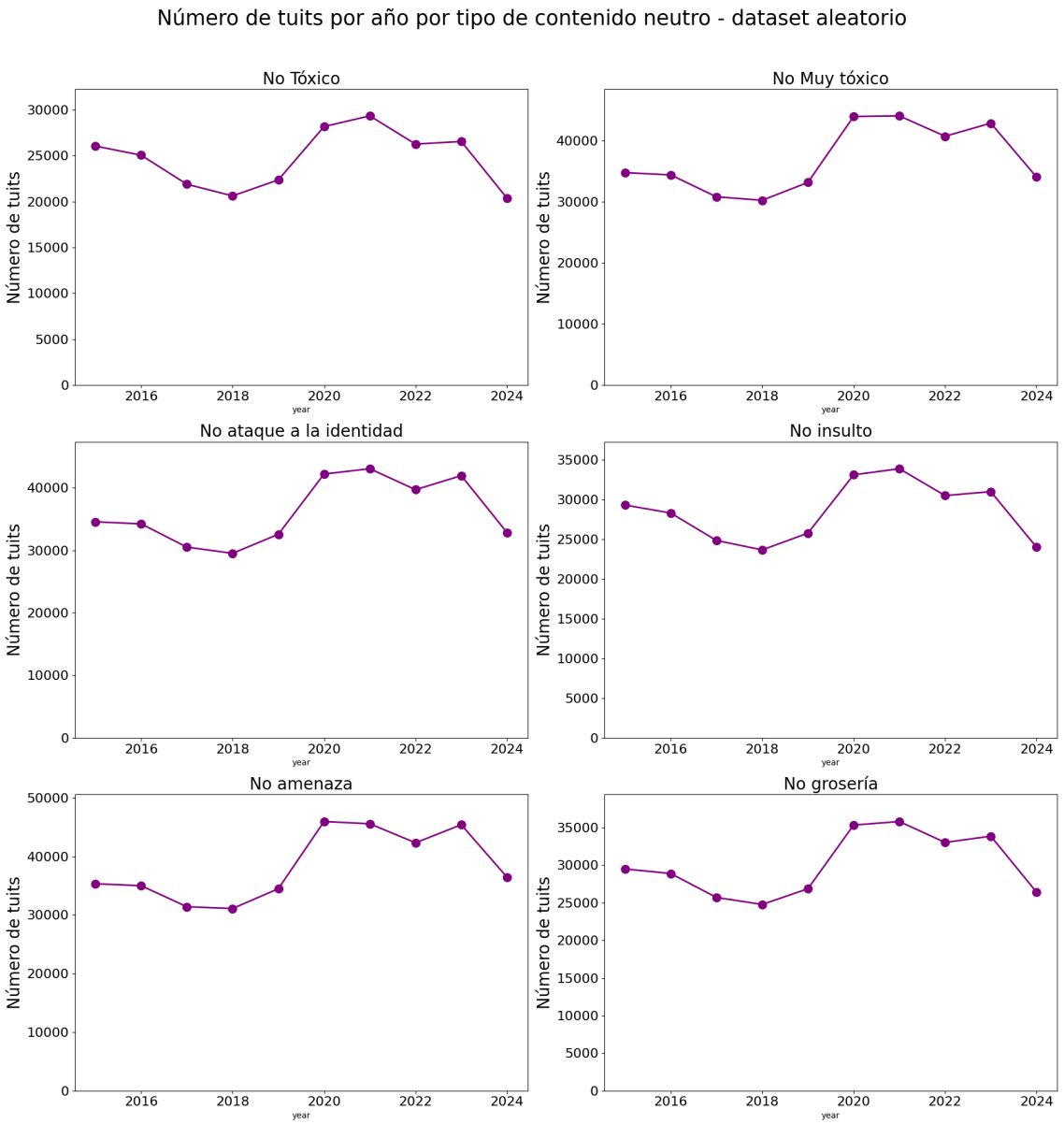

Figura 4.10: Comparación de la cantidad de contenido neutro - dataset aleatorio

Concluimos este apartado observando el porcentaje de aumento entre los distintos períodos descritos en el apartado metodología, esto es, la sección 3.2.

La tabla 4.5 nos muestra la variación de la cantidad de contenido entre los distintos períodos. Observamos que en 2020, como se mencionó anteriormente, se produce un pico en el contenido LGTBIQ+ (+178,3 %), y por tanto del contenido LGTBIQ-fóbico (+152,56 %). Por otro lado, la toxicidad del dataset aleatorio sube un 84,94 %, mientras que la subida del contenido aleatorio sube sólo un 37,56 %.

Tras la imposición de medidas en 2021 para impedir el contenido de odio observamos que los números descienden. De la misma manera, la eliminación de medidas

en 2023 supone un aumento del 46,8 % en el contenido de odio en ambos datasets.

Periodo	**Dataset LGTBIQ+**	**Dataset aleatorio**	**Dataset toxicidad LGTBIQ+**	**Dataset toxicidad aleatorio**
2015-2019 vs 2020	+178.3 %	+37.56 %	+152.56 %	+82.94 %
2020 vs 2021-2022	-51.26 %	-4.44 %	-52.35 %	-17.19 %
2021-2022 vs 2023-2024	-6.03 %	-6.51 %	+46.81 %	+46.80 %

Tabla 4.5: Porcentaje de aumento en número de tuits entre diferentes periodos de los distintos datasets

En lo que hace referencia a la figura 4.11, si estudiamos los porcentajes de aumento de todos los atributos, no solo el de toxicidad, para el dataset del colectivo LGTBIQ+, se observa que todas las gráficas siguen una misma tendencia durante esos periodos. Si algo llama la atención es el exagerado aumento de los tuits muy tóxicos contra el colectivo LGTBIQ+ tras la llegada de Elon Musk: 212,43 %.

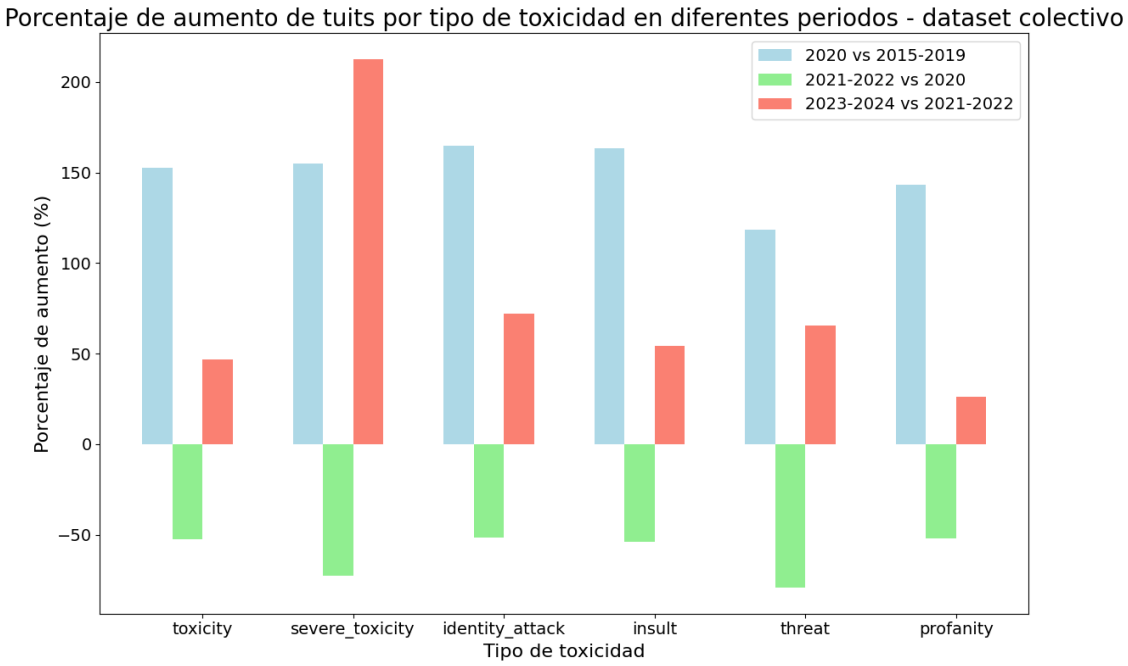

Figura 4.11: Comparación de la cantidad de contenido de odio - dataset colectivo

4.2.2. Variación relativa de tuits de odio al colectivo respecto al total

El apartado anterior ya muestra un claro aumento del discurso de odio en el último período; no obstante, las cantidades relativas al total de tuits, es decir, los porcentajes de tuits tóxicos respecto al total de tuits, reflejan resultados aún más alarmantes.

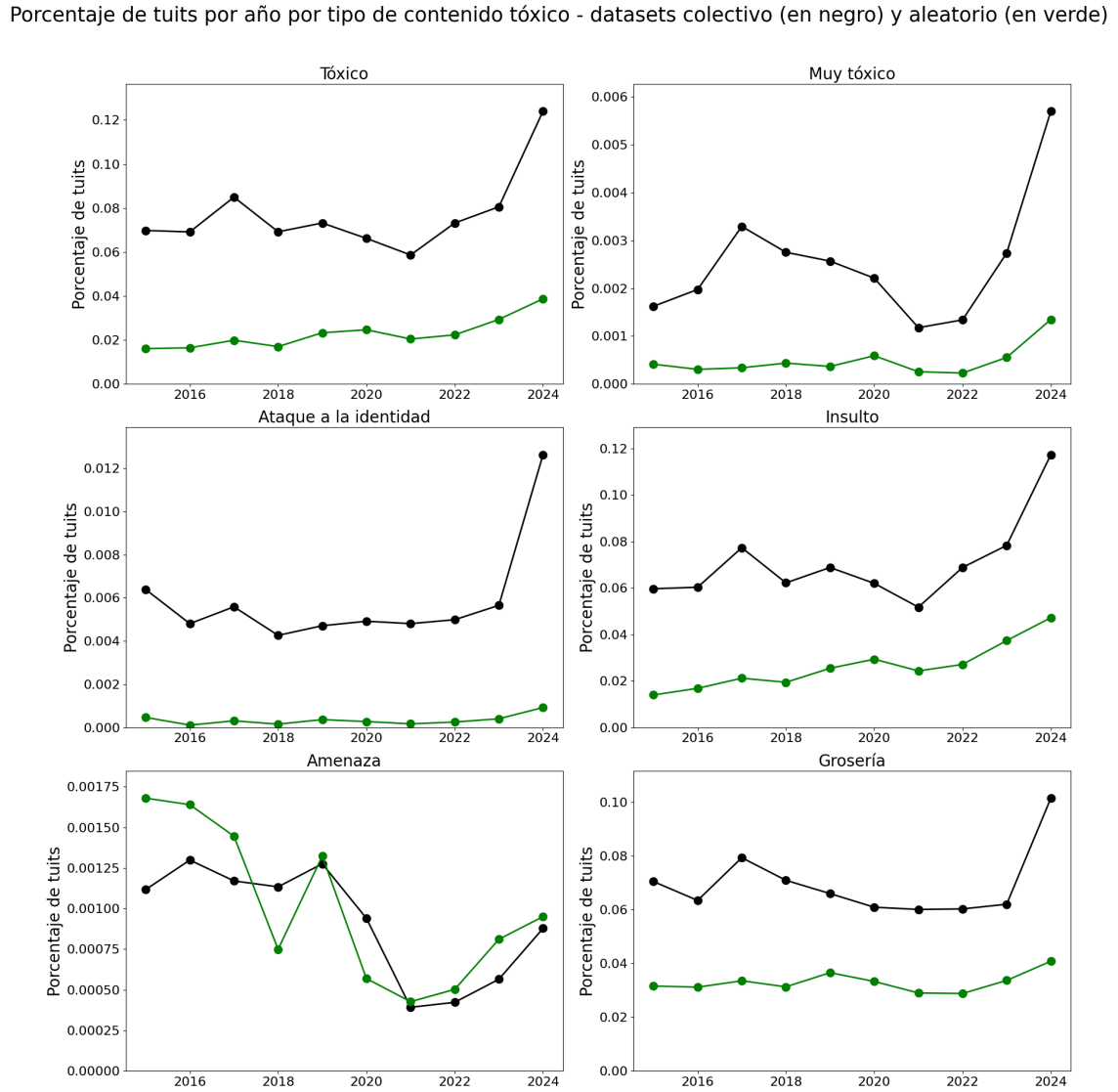

Figura 4.12: Porcentajes de tuits de contenido de odio - datasets colectivo y aleatorio

La figura 4.12 compara dos porcentajes:

- Porcentaje de tuits que cumplen algún atributo de toxicidad respecto del total de tuits del dataset de datos aleatorios (en verde).
- Porcentaje de tuits que cumplen algún atributo de toxicidad respecto del total

de tuits del dataset de datos del colectivo (en negro).

En primer lugar, observamos que durante todos los años el porcentaje de tuits que son tóxicos respecto del total es claramente mayor para aquellos tuits relacionados con el colectivo que para el de contenido aleatorio, salvo la excepción del atributo "amenaza". De hecho, esa diferencia de porcentajes se vuelve hasta cinco veces mayor en 2024 para los atributos que más nos conciernen: "tóxico", "muy tóxico" y "ataque a la identidad". También cabe resaltar el crecimiento desproporcionado de 2023 a 2024, cuando en teoría no ha habido ningún cambio en las políticas.

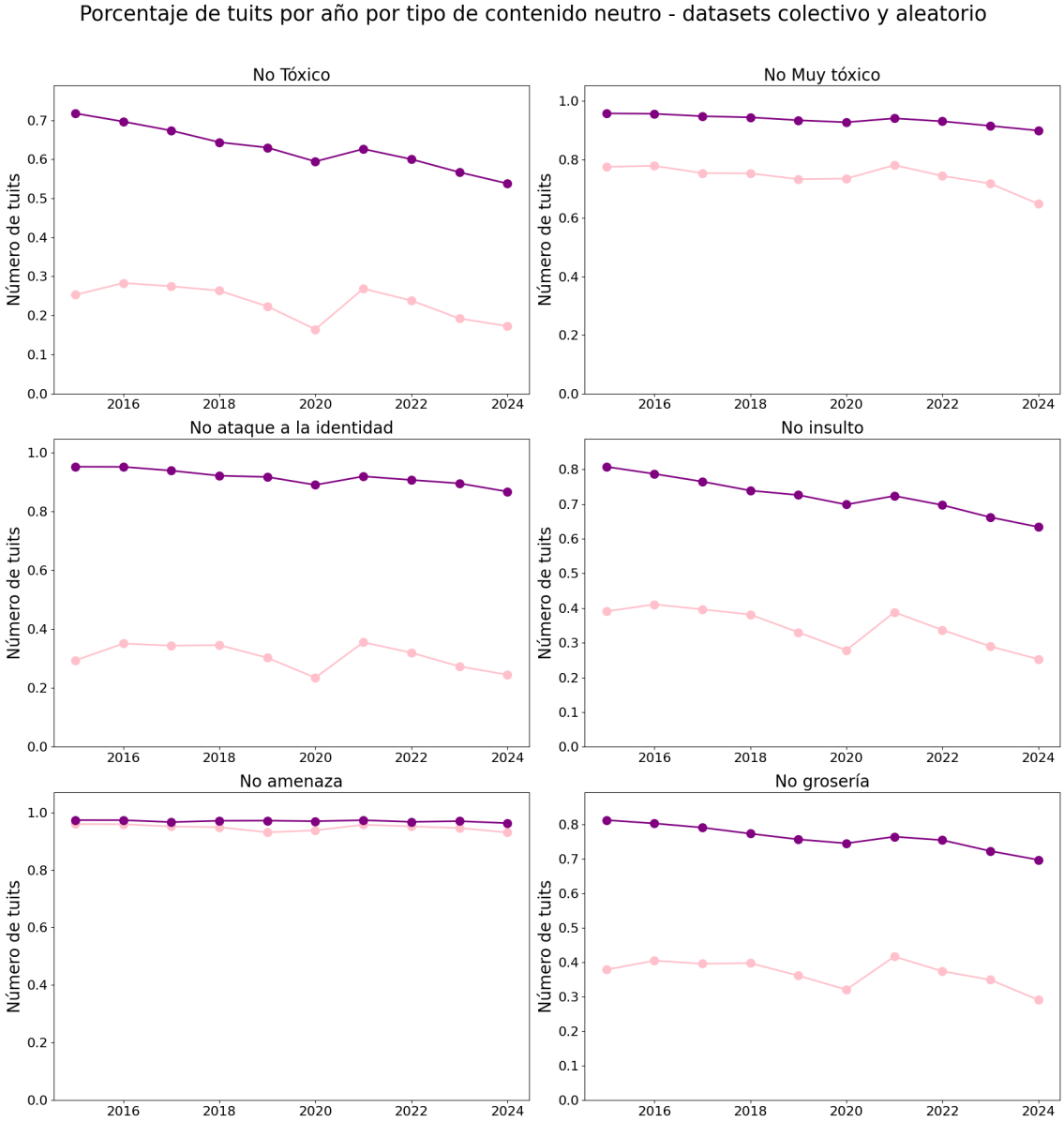

Figura 4.13: Porcentaje de tuits de contenido neutro - datasets colectivo (rosa) y aleatorio (morado)

En la figura 4.13 se comparan otros dos porcentajes:

- Porcentaje de tuits del dataset de tuits aleatorios que no cumplen atributos de toxicidad (en morado).
- Porcentaje de tuits del dataset de tuits del colectivo que no cumplen atributos de toxicidad (en rosa).

A la inversa de lo que observamos en la imagen anterior, es claro que el porcentaje de tuits neutros respecto del total de tuits es menor para el dataset del colectivo. Además, aunque observamos un descenso de la cantidad, es cierto que el descenso es muy parecido entre los dos datasets. Esto nos indica que Twitter está perdiendo neutralidad en sus publicaciones.

Tipo de toxicidad	2015-2019 vs 2020	2020 vs 2021-2022	2021-2022 vs 2023-2024
Tóxico	-9.26 %	-2.23 %	+56.23 %
Muy tóxico	-8.40 %	-43.71 %	+232.47 %
Ataque a la identidad	-4.83 %	-0.70 %	+83.30 %
Insulto	-5.43 %	-5.01 %	+64.35 %
Amenaza	-21.56 %	-56.93 %	+76.07 %
Grosería	-12.67 %	-1.22 %	+34.11 %

Tabla 4.6: Aumento del porcentaje de tuits de discurso de odio entre diferentes periodos - dataset del colectivo

Tipo de toxicidad	2015-2019 vs 2020	2020 vs 2021-2022	2021-2022 vs 2023-2024
Tóxico	+32.99 %	-13.35 %	+57.01 %
Muy tóxico	+59.02 %	-58.89 %	+274.36 %
Ataque a la identidad	-3.57 %	-23.53 %	+203.99 %
Insulto	+52.05 %	-12.59 %	+62.67 %
Amenaza	-58.77 %	-18.61 %	+88.45 %
Grosería	+1.39 %	-13.16 %	+27.50 %

Tabla 4.7: Aumento del porcentaje de tuits de discurso de odio entre diferentes periodos - dataset aleatorio

Así los porcentajes de aumento de tuits tóxicos respecto del total se reflejan en la tabla 4.6 para el dataset del colectivo y en la tabla 4.7 para el dataset aleatorio. En la primera tabla observamos que la moderación del contenido LGTBIQ-fóbico ha sido efectiva hasta la llegada de Elon Musk, momento en el que los porcentajes suben entre el 34 % (grosería) y el 232 % (muy tóxico). La segunda tabla muestra un aumento del discurso de odio en general durante la pandemia, así como el descenso tras la imposición de las políticas contra el delito de odio de 2021. La última columna, al igual que la tabla anterior, muestra las consecuencias de la no moderación del contenido de odio.

El aumento es aún más exagerado si miramos los datos del subgrupo trans del colectivo de la figura 4.14, cuyos porcentajes se reflejan en la tabla 4.8. Destacamos también el descenso producido en 2021, lo que indica que las medidas ejecutadas en ese año incluían la protección del subgrupo trans.

En la figura 4.14 y la tabla 4.8, queda patente que, tras la llegada de Musk, se ha producido un aumento de los tuits que cumplían algún atributo de toxicidad respecto

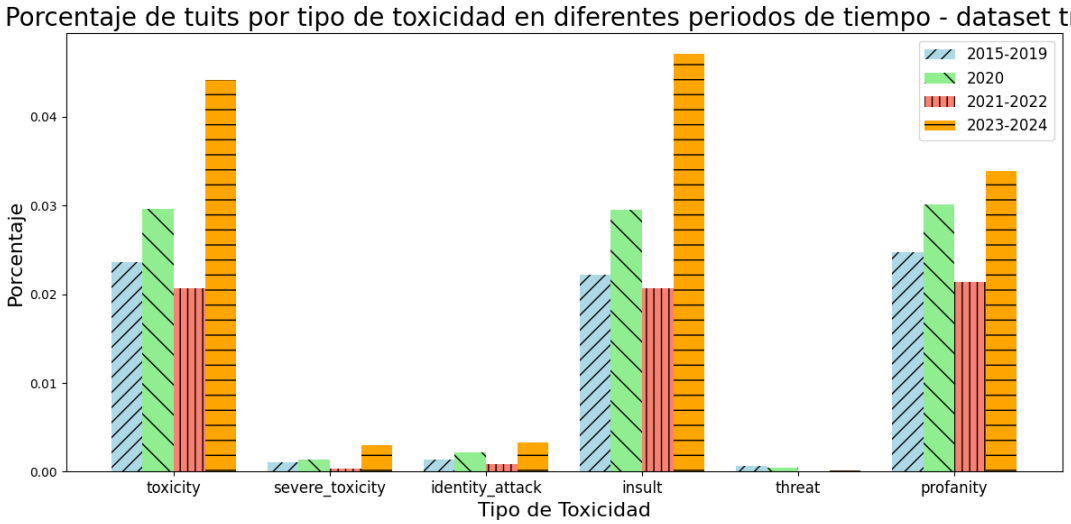

Figura 4.14: Comparación del porcentaje del contenido de odio hacia personas trans

Tipo de toxicidad	2015-2019 vs 2020	2020 vs 2021-2022	2021-2022 vs 2023-2024
Tóxico	+25.27 %	-30.25 %	+113.50 %
Muy tóxico	+32.25 %	-73.24 %	+690.10 %
Ataque a la identidad	+67.86 %	-60.27 %	+276.85 %
Insulto	+32.94 %	-29.75 %	+127.05 %
Grosería	+21.44 %	-28.93 %	+58.27 %

Tabla 4.8: Aumento del porcentaje de tuits de discurso de odio entre diferentes periodos para el dataset trans

de la totalidad de tuits del dataset trans, con porcentajes como 690.1 % para los tuits "muy tóxicos" o 276.85 % para los "ataques a la identidad" son una evidencia más que clara de la exclusión de este subgrupo como categoría protegida. No se tiene en cuenta el atributo "amenaza" porque el conjunto es demasiado pequeño.

4.2.3. Media del valor de toxicidad a lo largo del tiempo

En los apartados anteriores se han estudiado los atributos tóxicos para los valores $< 0{,}05$ (tuits neutros) y para $> 0{,}7$ (tuits tóxicos) ya que no se podía introducir en ninguna categoría al resto de los valores intermedios. Es por ello que parece interesante observar cuál es la tendencia de los valores de los atributos.

En este contexto, la figura 4.15 nos muestra que el valor de los atributos asciende a lo largo de los años tanto para los tuits sobre el colectivo como los aleatorios. Además, también observamos que el tono usado para referirse al colectivo ha sido históricamente más negativo que el contenido generalista.

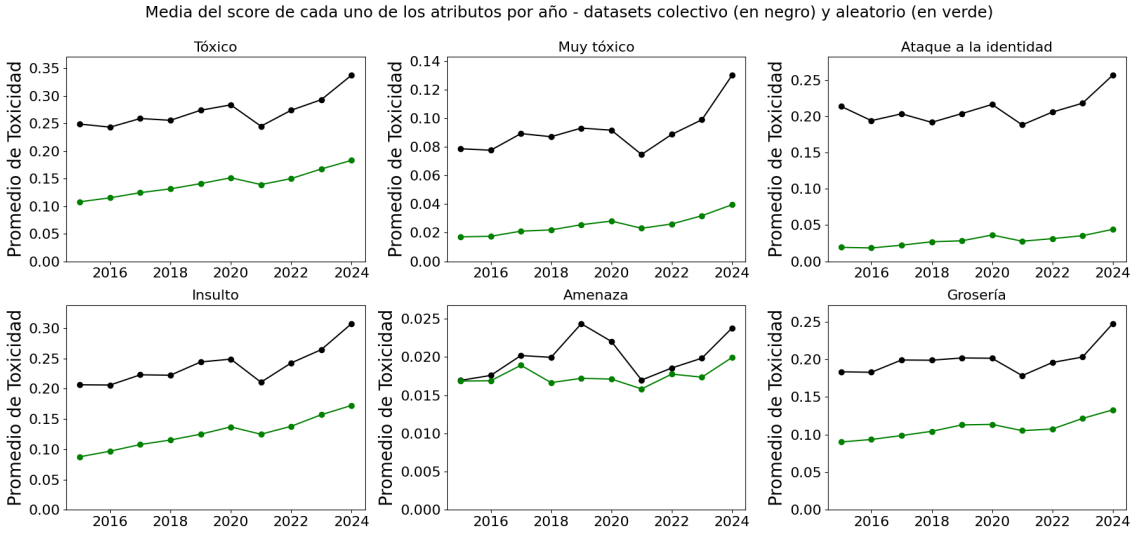

Figura 4.15: Comparación del valor del atributo - datasets colectivo vs aleatorio

4.2.4. Media de "likes" y "retuits" en tuits tóxicos a lo largo del tiempo

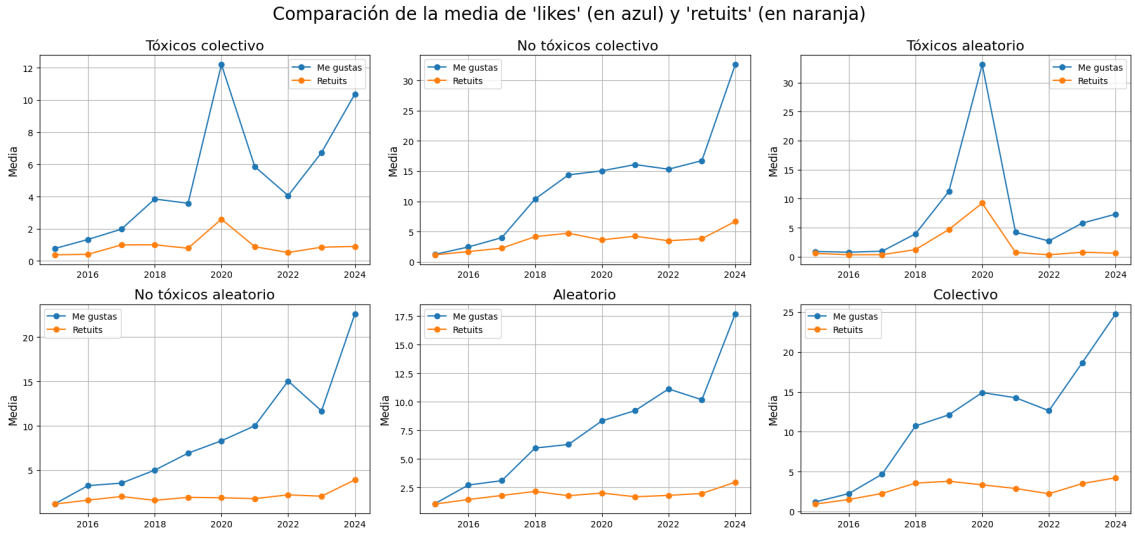

Figura 4.16: Comparación de la cantidad de "likes" y "retuits"

Puesto que toxicidad es el atributo mejor desarrollado de la Perspective API, y además es el atributo que más tuits contiene (lo que impedirá que un tuit viral modifique en exceso los resultados), tal y como se puede ver en la figura 4.7, vamos a centrarnos en este atributo. Las conclusiones para el resto de atributos son muy parecidas, aunque si el lector/a estuviera interesado/a en ver los resultados, puede consultar el repositorio de este trabajo, que se encuentra en [6].

El primer gráfico 4.16 compara la interacción de los tuits tóxicos y no tóxicos tanto del dataset aleatorio como el del colectivo. Destacamos las siguientes conclusiones:

- 2020 muestra una gran interacción con el contenido tóxico, tanto para el colectivo como el generalista. Esto entra en contraste con la interacción del contenido no tóxico, que no muestra un aumento destacable en 2020.
- En 2021 se produce un claro descenso de la visibilidad del contenido tóxico, lo que resalta, además de un menor uso de la aplicación que en 2020, el buen funcionamiento en la ejecución de las medidas que se llevaron a cabo en este año.
- Durante 2023 y 2024 se produce un aumento claro en la cantidad de "likes" y "retuits". Destacan por su subida los tres datasets con contenido sobre el colectivo. De hecho, observamos un mayor aumento de la interacción con el contenido tóxico hacia el colectivo que del tóxico en general.

Dataset	2015-2019 vs 2020	2020 vs 2021-2022	2021-2022 vs 2023-2024
LGTBIQ+	+125.49 %	-8.88 %	+58.68 %
Aleatorio	+120.42 %	+21.82 %	+33.44 %
Tóxico LGTBIQ+	+412.29 %	-58.92 %	+76.55 %
Tóxico aleatorio	+712.33 %	-89.6 %	+91.17 %

Tabla 4.9: Porcentaje de aumento de la media de "likes" a tuits del colectivo y aleatorios en diferentes períodos

Dataset	2015-2019 vs 2020	2020 vs 2021-2022	2021-2022 vs 2023-2024
LGTBIQ+	+34.25 %	-21.94 %	+47.56 %
Aleatorio	+23.56 %	-13.36 %	+38.47 %
Tóxico LGTBIQ+	+259.2 %	-72.5	+23.69 %
Tóxico aleatorio	+458.75 %	-94.37	+29.46 %

Tabla 4.10: Porcentaje de aumento de la media de "retuits" a tuits del colectivo y aleatorios en diferentes períodos

En las tablas 4.9 y 4.10 podemos observar los porcentajes de aumento y descenso de los "likes" y los "retuits" durante los diferentes periodos. Lo primero que llama la atención es el gran crecimiento que se produce en 2020. Esto puede ser por el aumento del uso de la actividad de Twitter que se produjo en 2020, aumentando la densidad y por tanto las interacciones con el contenido. Sin embargo, la diferencia entre el aumento del contenido tóxico con respecto al no tóxico puede indicar que se produjo un cambio del algoritmo y que este buscaba mostrar contenido más polémico, que suele implicar más interacciones y por tanto tiempo en la plataforma. Todo esto son suposiciones. No obstante, algo que se ve claro es cómo se reducen las interacciones con el contenido de odio tras la ejecución de políticas en 2021 y cómo vuelve a ampliar tras la relajación de las mismas con la llegada de Elon Musk.

Durante la última época el incremento de "retuits" es mayor para los tuits no tóxicos, sin embargo, el aumento de "likes" es mayor en los tuits tóxicos, por lo que la media de visibilidad se mantiene estable. Es decir, la visibilidad de contenido de odio crece de la misma manera en la que aumenta la actividad en general, lo que muestra que no se hacen esfuerzos por reducir la visibilidad del contenido tóxico, que es lo que prometió Musk con su "Freedom of Speech, Not Reach" (explicado en subsección 2.2.4).

4.2.5. Agregación de "likes" y "retuits" de tuits tóxicos a lo largo del tiempo

Para el apartado anterior, 100 tuits que han recibido un solo "likes", tienen una media mucho menor que un tuit de diez "likes". Parece lógico pues calcular también el aumento de la visibilidad del contenido de odio como la suma de todas las interacciones con tuits tóxicos. Es decir, en vez de tomar la media de "retuits" por tuit, se toma la suma de todos los "retuits" recibidos por todos los tuits del dataset, e igual con los "likes".

Puesto que los datasets son de distinto tamaño, no nos vamos a centrar en los valores de la suma, sino en el porcentaje de aumento o descenso de esa suma en los distintos períodos.

Dataset	2015-2019 vs 2020	2020 vs 2021-2022	2021-2022 vs 2023-2024
LGTBIQ+	+527,64 %	+55.59 %	+49.12 %
Aleatorio	+203.22 %	+16.42 %	+24.76 %
Tóxico LGTBIQ+	+1193.86 %	-80.43 %	+159.2 %
Tóxico aleatorio	+1386.1 %	-91.39 %	+180.65 %

Tabla 4.11: Porcentaje de aumento de la suma de "likes" a tuits del colectivo y aleatorios en diferentes períodos

Dataset	2015-2019 vs 2020	2020 vs 2021-2022	2021-2022 vs 2023-2024
LGTBIQ+	+273.69 %	-61.96 %	+38.66 %
Aleatorio	+69.97 %	-17.21 %	+29.47 %
Tóxico LGTBIQ+	+807.21 %	-86.89 %	+81.59 %
Tóxico aleatorio	+922.19 %	-95.34 %	+90.04 %

Tabla 4.12: Porcentaje de aumento de la suma de "retuits" a tuits del colectivo y aleatorios en diferentes períodos

Observamos en las tablas 4.11 y 4.12 que en este estudio el aumento del alcance de contenido de odio tras la llegada de Musk (columna 2021-2022 vs 2023-2024) es más de tres veces mayor para los "likes" y más de dos veces mayor para los "retuits" que el aumento de la visibilidad del contenido en general para los datasets del colectivo.

Cabe destacar el aumento del contenido tóxico para el dataset aleatorio, que es siete veces mayor en "likes" y tres veces mayor en "retuits" que el contenido en general.

4.2.6. Número de tuits tóxicos de usuarios "Verificados Azules"

La gráfica 4.17 muestra un aumento del 104 % del contenido tóxico hacia el colectivo por parte de las cuentas "Twitter Blue". Esto se debe al cambio en la manera de conseguir el verificado azul (antes como signo de veracidad y ahora por

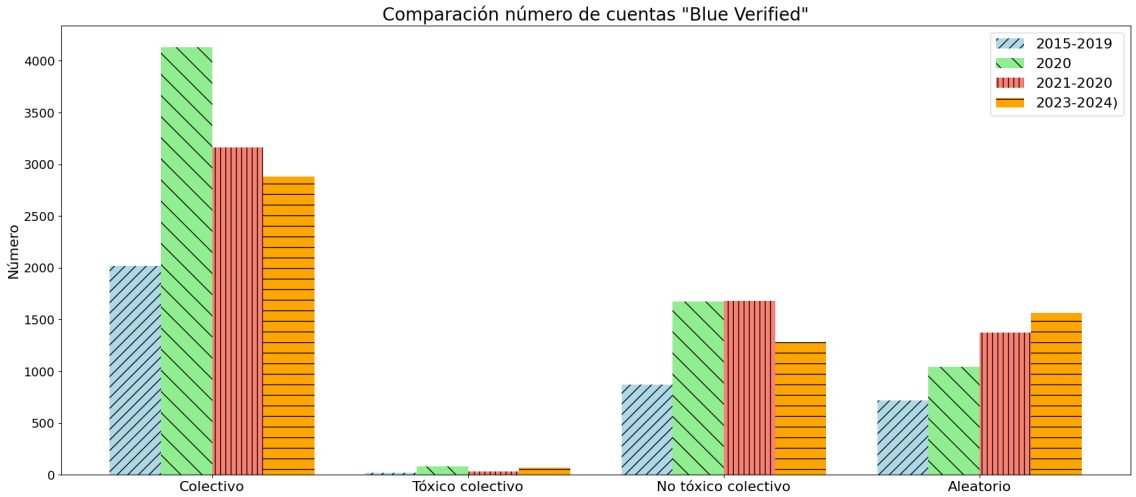

Figura 4.17: Comparación del número de cuentas "Blue Verified"

pago) y la menor moderación del contenido publicado por las cuentas verificadas. Este aumento se produjo a pesar del descenso de contenido sobre el colectivo de usuarios verificados. También es menor el número de publicaciones no tóxicas sobre el colectivo.

Capítulo 5

Conclusiones y trabajo futuro

5.1. Conclusiones y discusión razonada

Este proyecto consiste en el estudio de la LGTBIQ-fobia en Twitter antes y después de la llegada de Elon Musk. Para ello se han descargado los tuits en español del 28 de junio de cada año (Día del Orgullo) desde 2015 a 2024, obteniendo dos datasets. Uno con tuits relacionados con la comunidad LGTBIQ+ y otro de tuits aleatorios en español que se ha usado como punto de referencia, para comparar el contenido sobre el colectivo con el resto de actividad.

Este trabajo se ha centrado en seis estudios a lo largo del tiempo:

1. Cantidad de tuits tóxicos.
2. Variación relativa de odio al colectivo con respecto al total.
3. Media del valor de toxicidad a lo largo del tiempo.
4. Media de "likes" y "retuits" en tuits tóxicos a lo largo del tiempo.
5. Agregación de "likes" y "retuits" de tuits tóxicos a lo largo del tiempo.
6. Número de tuits tóxicos de usuarios "Verificados Azules".

El primer estudio, que observa la cantidad de tuits, nos muestra un aumento del número de tuits sobre el colectivo en 2020, así como un descenso en el siguiente periodo. Estos cambios provocan de manera proporcional un aumento y descenso natural en el contenido de odio. Sin embargo, en el periodo que marca la llegada de Elon Musk, notamos un aumento del 46,81 % del contenido de odio hacia la comunidad LGTBIQ+, que contrasta con un descenso del 6 % de los tuits sobre el colectivo. Este aumento llega incluso al 212,43 % si solo se consideran los tuits extremadamente ofensivos.

Esto es similar a los resultados obtenidos en [62], donde se detecta un aumento del 50 % del contenido tóxico tras la compra de Musk, que se contrasta con un aumento del 8 % de la actividad general.

Lo primero que llama la atención del segundo estudio, que trata de la variación relativa del contenido de odio, es que históricamente el porcentaje de tuits tóxicos hacia el colectivo es alrededor de cuatro veces mayor que el porcentaje de tuits tóxicos en general. Es decir, si un tuit es sobre la comunidad LGTBIQ+, entonces la probabilidad de que este sea contenido de odio es cuatro veces mayor. Otro detalle que es importante resaltar es que el porcentaje de tuits tóxicos ha ido disminuyendo en los distintos cambios de período, y no es hasta la compra de Elon Musk cuando sube de manera desproporcionada. Este aumento es del 56,23 % para tuits tóxicos, pero llega hasta 76,07 % si tomamos solo amenazas, hasta el 83,30 % si consideramos los ataques a la identidad y hasta el 232,47 % si consideramos contenido altamente tóxico. Por último, es importante mencionar que estos porcentajes son aún mayores si tomamos tuits sobre el subgrupo trans, llegando a un 203,99 % de aumento para los ataques a la identidad y a un 274,36 % en el contenido fuertemente tóxico.

El siguiente estudio, que estudia la media de toxicidad, nos muestra el aumento continuo (excepto en 2021, donde se produce un descenso) de la media de los valores de los atributos tóxicos para los tuits del colectivo de cada año. Esto representa las consecuencias de la exposición prolongada al discurso de odio online: incremento del prejuicio y la insensibilización, pérdida de empatía y neutralidad, y aumento de la toxicidad sobre el tema tratado.

Los dos estudios siguientes, que observan la media y agregación de "likes" y "retuits", desmienten el "Freedom of Speech, Not Reach" de Elon Musk. Traducido al español, significa "libertad de expresión, no alcance" y trata de justificar la eliminación de la moderación de contenido de odio con un supuesto descenso de su visibilidad.

En primer lugar, la media de "likes" y "retuits" a contenido de odio hacia el colectivo aumenta de manera proporcional al incremento de la interacción con el contenido en general. Por otro lado, observamos la suma de las interacciones por año como un valor único, pues este mide el alcance general del contenido de odio. Observamos que el crecimiento de la interacción con contenido de odio tras la llegada de Musk es del 159,2 % para "likes" y de 81,59 % para los "retuits", mientras que el aumento para la interacción con el contenido sobre el colectivo es del 49 % y del 38,66 % para los atributos mencionados anteriormente en ese orden. Además, cabe resaltar que para el contenido en general el aumento es del 24,79 % y 29,4 %, lo que nos dice que el interés por los tuits relacionados con el colectivo ha aumentado. Por último, observamos también un gran aumento del contenido tóxico en general, 180,65 % y 90,04 %, lo que nos indica que Twitter actualmente no es solo un lugar hostil para la comunidad LGTBIQ+ sino para todos los usuarios.

El último estudio muestra las consecuencias del sistema de obtención del verificado por medio de pago. Pese al descenso de las cuentas verificadas que publican contenido sobre el colectivo, se observa un aumento del 104 % de los tuits con con-

tenido de odio hacia la comunidad LGTBIQ+ que han sido publicados por cuentas verificadas en el periodo post-Musk. Esto reafirma la denuncia realizada por el CCDH, en la que exponían que el 99 % de los tuits con contenido de odio de cuentas "Twitter Blue" que fueron reportados no fueron eliminados. Es decir, exponían que los usuarios verificados eran "inmunes" a las consecuencias de publicación de contenido de odio.

Estos resultados muestran las consecuencias de la eliminación de las políticas que previenen contenido de odio: Twitter es hoy en día un territorio hostil para usuarios del colectivo LGTBIQ+. Además, estos resultados advierten de la tendencia que tomarían los distintos estudios realizados si no se instauran medidas contra contenido tóxico. Los análisis de este tipo pueden servir como advertencia de las consecuencias si otras redes sociales siguen el camino de Twitter, como por ejemplo Meta ha comenzado a hacer con Facebook e Instagram [41].

5.2. Limitaciones

En este trabajo se han descargado los tuits de cada 28 de junio de cada año. Debido a que Musk realizó la compra de Twitter en octubre de 2022 y aún no ha sucedido el 28 de junio de este año 2025, solo hay dos años que representen la era post-Musk, por lo que algunos estudios, como son la cantidad de tuits, solo poseen dos valores para el último período. No obstante, este trabajo es ilustrador de la tendencia que seguirían los estudios si no se toman medidas para la eliminación y reducción del contenido de odio.

El uso de la Perspective API hace que dependamos de su análisis de tuits tóxicos así como del conjunto de datos que hayan usado para entrenar su modelo. Puesto que, dada la cantidad de tuits que hay en cada dataset, no ha podido verificarse más que parcialmente.

La extracción de datos se basó en un conjunto de palabras clave definido a partir de un glosario externo que posteriormente se amplió. Este enfoque presenta la limitación de no capturar vocablos LGTBIQ-fóbicos de reciente aparición en Twitter. Para mitigar este sesgo, podrían aplicarse técnicas de *topic modelling* (modelado de temas) que permitan detectar de forma automática términos emergentes.

5.3. Trabajo futuro

Sería interesante tomar otros marcos temporales. Es decir, en vez de tomar días representativos, se podrían tomar días totalmente aleatorios, o toda la semana del orgullo, etc. Una forma fácil de ampliar este proyecto sería añadir un número de días por año, para así tener más variedad de información para cada uno de los años.

Tal y como se ha mencionado en 3.2.2, recolectar datos por palabras clave puede suponer perder información relevante (si esta no contiene ninguna de las palabras clave), así como obtener alguna irrelevante. Por ello, se podría utilizar un método distinto para la obtención de datos, como seguir a influencers o la búsqueda de tuits por hashtags.

Otra opción para ampliar el trabajo sería tomar los datos de más días que también sean representativos para el Orgullo (e.g. toda la semana), para así tener más variedad en la muestra para el estudio del número de tuits.

Introduction

"One's freedom ends where another's begins"
— Rousseau

Motivation

In recent decades, we have witnessed great progress in the inclusion and improvement of the rights of the LGBTIQ+ community. The arrival of the Internet and, with it, the globalization of information, has facilitated the recognition of various minorities, including this community. Many applications such as Twitter have been a place for social criticism and have promoted the study and advancement of the inclusion of these social categories. This platform has enabled dialogue between people with diverse ideologies, generating reflection and debate. In addition, over the years, policies have been implemented to reduce hate speech and misinformation, contributing to creating a more inclusive environment.

This environment of progress suffered a drastic setback following changes to hate speech moderation policies made by Elon Musk, who purchased the platform in October 2022. These changes were motivated by what the entrepreneur defines as "Freedom of Speech, Not Reach", claiming that, despite the changes, exposure to hateful content and misinformation had decreased [50, 111]. However, these statements lack transparency, as X may have a very different definition from the rest of society of what constitutes "hate speech" or "freedom of expression". A clear example of this is the recent ban on the word "cisgender" (a person who feels comfortable with their gender identity as assigned at birth [70]), after it was considered an insult. However, this is a word commonly used by doctors [101] and queer people [82].

Policy changes and the growing polarization of our society have made Twitter a hostile place for certain minorities, including the LGBTIQ+ community. This situation has led many users to migrate to other social networks such as Bluesky, Threads, and Mastodon [115, 54], as well as a spike in interest from researchers in trying to raise awareness of the current situation and the study of the possible short- and long-term impacts if this situation remains the same.

This project studies the increase in LGBTIQ-phobia on Twitter following the arrival of Elon Musk, as well as discussing the possible causes and consequences. This calls for social pressure to implement measures that moderate content that is clearly offensive to this minority, not only on Twitter, but also on other social networks. The aim is to introduce new policies for moderating hateful content and to maintain those already in place.

This project has been carried out on tweets in Spanish and studies the consequences of policy changes over the last ten years. Furthermore, hate speech is considered as a spectrum, not as a single value. In this way, we aim to address the methodological limitations of similar studies caused by linguistic, temporal, and conceptual biases (as will be explained in the state of the art, section 2.2.5).

Objectives

The main objective of this project is to study the impact of Elon Musk's arrival on Twitter on the LGBTIQ+ community, observing content trends and showing the current situation. To this end, the following two objectives have been set:

- Quantitative analysis of the increase in hate speech, in general, and towards the LGBTIQ+ community, between 2015 and 2024, with a focus on different significant periods.
- Quantitative analysis of the increase in interaction and visibility of hate content, in general, and towards the community, in the same period.
- To study how Elon Musk's purchase has affected hate speech in general and specifically towards the community, analyzing the statistical significance of the differences found.

To address the first objective, the evolution of hate content over time has been analyzed. To address the second objective, interaction with such content, i.e., the amount of "likes" and "retweets" received, was studied. Finally, the time frame was divided into different periods: pre-lockdown, lockdown, post-lockdown, and post-Musk. The results place special emphasis on the changes following Elon Musk's acquisition of Twitter, which allows us to address the third objective.

Work Plan

The first phase, as in any research project, consisted of understanding the problem. To do so, meetings were held with various social science researchers, who helped

to understand and define the topic of study. In addition, similar works were read, which provided some clues, inspiration, and resources that would be used later.

Secondly, we tackled what would be the biggest challenge of the project: collecting the tweets. An initial collection of tweets was carried out, followed by a basic analysis. This study confirmed our hypotheses, which led to the beginning of the bulk of the project.

In this context, all tweets containing certain LGBTIQ-related keywords (see 3.2.2 for a deeper explanation) from June 28 of the last ten years were collected. To do this, it was necessary to collect a dataset of random tuits as a benchmark to compare with the community dataset. To create the dataset, three common words in Spanish ("ser", "algo", and "todos", which means "to be", "thing" and "everybody") were selected as keywords, and the tweets were collected and processed.

The next step was therefore to analyze the data to study the rise of LGBTIQ-phobia. In order to facilitate comparisons, graphs and percentage increases were used, for which we used Jupyter Notebook. Finally, we checked that these data were statistically significant.

The last month was dedicated to writing and reviewing the report.

Structure of the Report

The report is divided into five chapters:

- Introduction: the work and its importance are presented. In addition, the reasons that make this a relevant topic and the main objectives are outlined.
- Theoretical foundations: try to provide the reader with the necessary tools to understand the project, describing the key concepts and reviewing previous work in this area.
- Technologies and methodology: presents the technology used and the methodology followed to obtain the results.
- Results and analysis: the statistical analysis of the data studied is presented, as well as the results of the objectives mentioned in section 5.3.
- Conclusions and future work: a reasoned discussion of the conclusion is provided, remarking the limitations and possible future work.

Conclusions and Future Work

Conclusions and Reasoned Discussion

This project consists of studying LGBTIQ-phobia on Twitter before and after the arrival of Elon Musk. To do so, Spanish tweets from June 28 of each year (Pride Day) from 2015 to 2024 were downloaded, obtaining two datasets. One with tweets related to the LGBTIQ+ community and another with random tweets in Spanish. The random one was used as a benchmark to compare the content about the community with the rest of the activity.

This work has focused on six studies over time:

1. Number of toxic tweets.
2. Relative variation in hate speech toward the community compared to the total.
3. Average toxicity value over time.
4. Average number of likes and retweets on toxic tweets over time.
5. Aggregation of "likes" and 'retweets' of toxic tweets over time.
6. Number of toxic tweets from "Verified Blue" users.

The first study, which focuses on the number of tweets, shows an increase in the number of tweets related to the LGBTIQ+ community in 2020, as well as a decrease in the following period. These changes cause a proportional increase and natural decrease in hateful content. However, in the post-Musk period, there is a 6% decrease in LGBTIQ-related tweets, which contrasts with a 46.81% increase when only offensive community-related tweets are being considered. This increase goes up to 212.43% if only extremely offensive tweets are considered.

This is similar to the results obtained in [62], where a 50% increase in toxic content was detected after Musk's purchase, contrasting with an 8% increase in overall activity.

The second study, which deals with the relative variation in hate content, shows that historically the percentage of toxic tweets towards the community is around four times higher than the percentage of toxic tweets in general. In other words, if a tweet is LGBTIQ-related, then the probability that it turns out to be hate speech is four times higher. Another detail to highlight is that the percentage of toxic tuits decreased over the different periods, and it is not until Elon Musk's purchase that it rises disproportionately. This increase is 56.23% for toxic tweets, but rises to 76.07% if we only consider threats, to 83.30% if we consider identity attacks, and to 232.47% if we consider severe toxicity content. Finally, it is important to mention that these percentages are even higher if we take tweets about the trans subgroup, reaching a 203.99% increase for identity attacks and 274.36% for severe toxicity content.

The following study, which looks at average score toxicity, shows a continuous increase during the years (except in 2021, when there was a decrease) in the average score of the toxic attributes for LGBTIQ-related tweets. This represents the consequences of prolonged exposure to online hate speech: increased prejudice and desensitization, loss of empathy and neutrality, and increased toxicity on the topic at hand.

The following two studies, which look at the average and aggregation of "likes" and "retweets", refute Elon Musk's "Freedom of Speech, Not Reach", which means that users are free to post anything, but hate speech tweets will have reduced visibility.

First, the average number of "likes" and "retweets" for hateful content toward the LGBTIQ+ community increases proportionally to the rise in interaction with the content in general. On the other hand, we look at the sum of interactions per year as a single value, as this measures the overall reach of hateful content. Results show that the growth in interaction with LGBTIQ-related content after Musk's arrival is of 49% for "likes" and of 38.66% for "retweets", which contrasts with the 159.2% and 81,59% if only toxic community-related tuits are considered. Furthermore, the random content increase in visibility is 24.79% ("likes") and 29.4% ("retuits"), which suggests that interest in LGBTIQ-related tweets has increased. Finally, a large increase in toxic random content has been observed, concretely 180.65% and 90.04%, which indicates that Twitter is currently a hostile place for all users, not only for the LGBTIQ+ community.

The latest study shows the consequences of the "pay to become Blue Verified system". Despite the decrease in LGBTIQ-related tweets posted by verified accounts, these accounts show a 104% increase in hateful community-related content during the post-Musk period. This reaffirms the complaint made by the CCDH, which stated that 99% of hateful tweets from "Twitter Blue" accounts that were reported were not removed. In other words, they exposed that verified users were "immune" to the consequences of posting hateful content.

These results show the consequences of removing policies that prevent hateful content: Twitter is now a hostile place for users of the LGBTIQ+ community. Furthermore, these results warn of the trend that the different studies done would

take if measures against toxic content are not executed. Analyses of this kind can be used as a warning of the consequences if other social networks follow Twitter's lead, as Meta has begun to do with Facebook and Instagram [41].

Limitations

In this study, tweets were downloaded on June 28 of each year. Since Musk purchased Twitter in October 2022, and 2025th pride day has not yet occurred, there are only two years representing the post-Musk era, this causes that some studies, such as the number of tweets, have only two values for the last period. However, this study helps to illustrate the trend that studies would follow if no measures were taken to remove and reduce hateful content.

The use of the Perspective API makes us dependent on its analysis of toxic tweets as well as the dataset they used to train their model. Given the number of tweets in each dataset, it has only been possible to verify it partially.

Data extraction was based on a set of keywords defined from an external glossary, which was expanded. This approach has the limitation of not capturing LGBTIQ-phobic words that have recently appeared on Twitter. To mitigate this bias, topic modeling techniques could be applied to automatically detect emerging terms.

Future Work

As mentioned in 3.2.2, collecting data by keywords can lead to a loss of relevant information (if it does not contain any of the keywords), as well as obtaining some irrelevant data. That is why another study using a different method for getting data could be done. Some example of methods would be studying tweets of some selected influencers or searching tweets by hashtags.

Another option to expand this work would be to take data from more days that are also representative of Pride (e.g., the whole week), in order to have a wider variety of data to study.

Bibliografía

[1] BERT. URL https://huggingface.co/docs/transformers/model_doc/bert.

[2] dccuchile/bert-base-spanish-wwm-cased · Hugging Face. URL https://huggingface.co/dccuchile/bert-base-spanish-wwm-cased.

[3] BOE-A-1995-25444 Ley Orgánica 10/1995, de 23 de noviembre, del Código Penal. URL https://www.boe.es/buscar/act.php?id=BOE-A-1995-25444.

[4] Presentación | Interactive Chaos. URL https://interactivechaos.com/es/manual/tutorial-de-scipy/presentacion.

[5] Two Sample Kolmogorov-Smirnov | Real Statistics Using Excel. URL https://real-statistics.com/non-parametric-tests/goodness-of-fit-tests/two-sample-kolmogorov-smirnov-test/.

[6] MarRamiro/Analysis-Twitter: This repository contains the code for analyzing LGTBIQ-phobia on Twitter, focusing on Spanish tweets from June 28th (Pride Day) between 2015 and 2024, with a special focus on the impact of Elon Musk's acquisition of the platform. It includes data collection, toxicity classification using the Perspective API, statistical analysis, and visualization. URL https://github.com/MarRamiro/Analysis-Twitter.

[7] tweepy/tweepy: Twitter for Python! URL https://github.com/tweepy/tweepy.

[8] Acerca de X Premium. URL https://help.x.com/es/using-x/x-premium.

[9] Elon Musk takes control of Twitter in $44bn deal. October 2022. URL https://www.bbc.com/news/technology-63402338.

[10] Over 100 Reddit groups ban X links in protest at Musk arm gesture, January 2025. URL https://www.bbc.com/news/articles/c77r1p887e5o.

[11] R. L. Abreu and M. C. Kenny. Cyberbullying and lgbtq youth: A systematic literature review and recommendations for prevention and intervention. *Journal of Child & Adolescent Trauma*, 11(1):81–97, 2017. doi: 10.1007/s40653-017-0175-7. URL https://doi.org/10.1007/s40653-017-0175-7. Published 2017 Jul 24.

[12] Alfredo Sánchez Alberca. La librería Matplotlib, . URL https://aprendeconalf.es/docencia/python/manual/matplotlib/.

[13] Alfredo Sánchez Alberca. La librería Numpy, . URL https://aprendeconalf.es/docencia/python/manual/numpy/.

[14] Alfredo Sánchez Alberca. La librería Pandas, . URL https://aprendeconalf.es/docencia/python/manual/pandas/.

[15] Malik Almaliki. Cyberhate dissemination: A systematic literature map. *IEEE Access*, 11:117385–117392, 2023. doi: 10.1109/ACCESS.2023.3326254.

[16] Hind Almerekhi, Supervised by Bernard J. Jansen, and co-supervised by Haewoon Kwak. Investigating toxicity across multiple reddit communities, users, and moderators. In *Companion Proceedings of the Web Conference 2020*, WWW '20, page 294–298, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450370240. doi: 10.1145/3366424.3382091. URL https://doi.org/10.1145/3366424.3382091.

[17] Hind Almerekhi, Haewoon Kwak, and Bernard Jansen. Statistical modeling of harassment against reddit moderators. In *Companion Proceedings of the Web Conference 2020*, WWW '20, page 122–123, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450370240. doi: 10.1145/3366424.3382729. URL https://doi.org/10.1145/3366424.3382729.

[18] Manuel Alvar Ezquerra. La frecuencia léxica y su utilidad en la enseñanza del español como lengua extranjera. In María Auxiliadora Castillo Carballo, Olga Cruz Moya, Juan Manuel García Platero, Juan Pablo Mora Gutiérrez, and Mª Regla Cordero Raffo, editors, *Las gramáticas y los diccionarios en la enseñanza del español como segunda lengua. Deseo y realidad: actas del XV Congreso Internacional de ASELE*, pages 19–39. ASELE, 2005. ISBN 84-472-0882-6.

[19] Perspective API. Perspective api documentation, . URL https://perspectiveapi.com/.

[20] Perspective API. Acerca de la api: Atributos y lenguajes, . URL https://developers.perspectiveapi.com/s/about-the-api-attributes-and-languages?language=es.

[21] Twitter API. Twitter api documentation, . URL https://docs.twitterapi.io/introduction.

[22] Carlos Arcila Calderón, Patricia Sánchez Holgado, Jesús Gómez, Marcos Barbosa, Haodong Qi, Alberto Matilla, Pilar Amado, Alejandro Guzmán, Daniel López-Matías, and Tomás Fernández-Villazala. From online hate speech to offline hate crime: the role of inflammatory language in forecasting violence against migrant and lgbt communities. *Humanities and Social Sciences Communications*, 11(1):1369, 2024. doi: 10.1057/s41599-024-03899-1. URL https://doi.org/10.1057/s41599-024-03899-1.

[23] Muhammad Arslan, Manuel Sandoval Madrigal, Mohammed Abuhamad, Deborah L. Hall, and Yasin N. Silva. Detecting lgbtq+ instances of cyberbullying, 2024. URL https://arxiv.org/abs/2409.12263.

[24] Imran Awan and Irene Zempi. The affinity between online and offline anti-muslim hate crime: Dynamics and impacts. *Aggression and Violent Behavior*, 27:1–8, 2016. ISSN 1359-1789. doi: https://doi.org/10.1016/j.avb.2016.02.001. URL https://www.sciencedirect.com/science/article/pii/S1359178916300015.

[25] Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion - WWW '17 Companion*, WWW '17 Companion, page 759–760. ACM Press, 2017. doi: 10.1145/3041021.3054223. URL http://dx.doi.org/10.1145/3041021.3054223.

[26] Bond Benton, Jin-A Choi, Yi Luo, and Keith Green. Hate speech spikes on twitter after elon musk acquires the platform, 2022. URL https://digitalcommons.montclair.edu/scom-facpubs/33.

[27] Michał Bilewicz and Wiktor Soral. Hate speech epidemic. the dynamic effects of derogatory language on intergroup relations and political radicalization. *Political Psychology*, 41(S1):3–33, 2020. doi: 10.1111/pops.12670. URL https://doi.org/10.1111/pops.12670. First published: 19 June 2020.

[28] James Bisbee and Kevin Munger. The vibes are off: Did elon musk push academics off twitter? *PS: Political Science & Politics*, 58(1):139–146, 2025. doi: 10.1017/S1049096524000416.

[29] Brookings. Why is elon musk's twitter takeover increasing hate speech?, November 2022. URL https://www.brookings.edu/articles/why-is-elon-musks-twitter-takeover-increasing-hate-speech/.

[30] Alexander Brown. What is hate speech? part 1: The myth of hate. *Law and Philosophy*, 36(4):419–468, 2017. doi: 10.1007/s10982-017-9297-1. URL https://doi.org/10.1007/s10982-017-9297-1.

[31] Bharathi Raja Chakravarthi, Ruba Priyadharshini, Rahul Ponnusamy, Prasanna Kumar Kumaresan, Kayalvizhi Sampath, Durairaj Thenmozhi, Sathiyaraj Thangasamy, Rajendran Nallathambi, and John Phillip McCrae. Dataset for identification of homophobia and transophobia in multilingual youtube comments, 2021. URL https://arxiv.org/abs/2109.00227.

[32] Observatorio LGTBI CLM. ¿qué es la lgtbifobia? URL https://observatoriolgtbiclm.com/que-es-la-lgtbifobia/.

[33] El Confidencial. Orgullo gay 2020: Madrid marcha ónline'(y en las calles) por los derechos lgtbi, 2020. URL https://www.elconfidencial.com/espana/2020-06-28/madrid-orgullo-gay-2020-manifestacion-lgtbi_2655428/.

[34] Congosto. Tweetscraperr notebook. URL https://github.com/congosto/TweetScraperR_notebook.

[35] d60. d60/twikit, May 2025. URL https://github.com/d60/twikit.

[36] Instituto de la Juventud (INJUVE). El discurso del odio en redes sociales, 2019. URL https://www.injuve.es/sites/default/files/2019/02/noticias/el_discurso_del_odio_en_rrss.pdf.

[37] Universidad de las Palmas de Gran Canaria. Prueba de Bondad de Ajuste de Kolmogorov-Smirnov (KS). URL https://www2.ulpgc.es/hege/almacen/download/5/5015/Complemento_3_Prueba_de_Bondad_de_Ajuste_de_Kolmogorov_Smirnov.pdf.

[38] El Heraldo de México. ¿será el fin de twitter? empleados renuncian masivamente tras ultimátum de elon musk, Noviembre 2022. URL https://heraldodemexico.com.mx/mundo/2022/11/17/sera-el-fin-de-twitter-empleados-renuncian-masivamente-tras-ultimatum-de-elon-musk-458710.html.

[39] El Diario. Twitter elimina una política clave sobre acoso hacia personas trans en la red social. URL https://www.eldiario.es/tecnologia/twitter-elimina-politica-clave-acoso-personas-trans-red-social_1_10134167.html.

[40] Elizabeth Dubois and Anna Reepschlager. How harassment and hate speech policies have changed over time: Comparing facebook, twitter and reddit (2005–2020). *Policy and Internet*, 16, 2024. doi: 10.1002/poi3.387. URL https://doi.org/10.1002/poi3.387. First published: 23 April 2024.

[41] Clare Duffy. Meta gets rid of fact checkers and says it will reduce 'censorship' | CNN Business, January 2025. URL https://www.cnn.com/2025/01/07/tech/meta-censorship-moderation.

[42] EFE. Twitter dejará de cotizar en bolsa el 8 de noviembre tras cerrar su venta a Elon Musk, October 2022. URL https://www.elperiodico.com/es/economia/20221028/twitter-cotizar-bolsa-elon-musk-77852234. Section: Economía.

[43] Sabine A. Einwiller and Sora Kim. How online content providers moderate user-generated content to prevent harmful online communication: An analysis of policies and their implementation. *Policy and Internet*, 12(2):184–206, 2020. doi: 10.1002/poi3.239. URL https://doi.org/10.1002/poi3.239. First published: 02 May 2020.

[44] Expansión. La pandemia dispara el uso de las redes sociales, un 27 URL https://www.expansion.com/economia-digital/innovacion/2021/02/10/6022c89de5fdea59448b459b.html.

[45] Federación Estatal de Lesbianas, Gais, Trans, Bisexuales, Intersexuales y más. Manual de instrucciones. URL https://felgtbi.org/trans/.

[46] FELGTBI. Discursos de odio en twitter, 2024. URL https://felgtbi.org/wp-content/uploads/2024/01/Discursos-de-odio-en-Twitter-FELGTBI-2.pdf.

[47] Pandora FMS. ¿para qué sirve una api?, 2019. URL https://web.archive.org/web/20190215165536/https://blog.pandorafms.org/es/para-que-sirve-una-api/.

[48] Fortune. People are fleeing elon musk's x for threads and bluesky. welcome to the era of social media fragmentation, November 2024. URL https://fortune.com/2024/11/14/x-elon-musk-leaving-election-trump-threads-bluesky-social-media-fragmentation/.

[49] Fortune. X's crowd-sourced 'community notes' fact checks fail to address flood of u.s. election misinformation, report says, October 2024. URL https://fortune.com/2024/10/31/x-community-notes-fact-checks-us-election-misinformation/.

[50] S. Frenkel and K. Conger. Hate speech's rise on twitter is unprecedented, researchers find, 2022. URL https://www.nytimes.com/2022/12/02/technology/twitter-hate-speech.html.

[51] Jeremy A Frimer, Harinder Aujla, Matthew Feinberg, Linda Skitka, karl Aquino, johannes C Eichstaedt, and Robb Willer. Incivility is rising among american politicians on twitter, Feb 2022. URL osf.io/preprints/psyarxiv/2hku3_v1.

[52] Iginio Gagliardone. Mapping and analysing hate speech online, 2014. URL https://ssrn.com/abstract=2601792. Available at SSRN: https://ssrn.com/abstract=2601792 or http://dx.doi.org/10.2139/ssrn.2601792.

[53] M. Garaigordobil and E. Larrain. Bullying and cyberbullying in lgbt adolescents: Prevalence and effects on mental health. *Comunicar*, 62:79–90, 2020. doi: 10.3916/C62-2020-07. URL https://doi.org/10.3916/C62-2020-07. Also available in Spanish: *Acoso y ciberacoso en adolescentes LGTB: Prevalencia y efectos en la salud mental.*

[54] D. Gewirtz. Bluesky vs. threads vs. mastodon: If you leave twitter, where will you go?, 2023. URL https://www.zdnet.com/article/bluesky-vs-threads-vs-mastodon-if-you-leave-twitter-where-will-you-go/.

[55] GLAAD. Glaad responds to twitter's rollback of long-standing lgbtq hate speech policy, 2023. URL https://glaad.org/releases/glaad-responds-twitters-roll-back-long-standing-lgbtq-hate-speech-policy#:%7E:text=transgender-,GLAAD%20RESPONDS%20TO%20TWITTER'S%20ROLL%2DBACK%20OF%20LONG,STANDING%20LGBTQ%20HATE%20SPEECH%20POLICY&text=GLAAD%3A%20%E2%80%9CTwitter's%20decision%20to%20covertly,for%20users%20and%20advertisers%20alike.%E2%80%9D.

[56] Daniel Glez-Peña, Anália Lourenço, Hugo López-Fernández, Miguel Reboiro-Jato, and Florentino Fdez-Riverola. Web scraping technologies in an api world. *Briefings in Bioinformatics*, 15(5):788–797, 04 2013. doi: 10.1093/bib/bbt026. URL https://doi.org/10.1093/bib/bbt026.

[57] Manuel Gámez-Guadix and Daniel Incera. Homophobia is online: Sexual victimization and risks on the internet and mental health among bisexual, homosexual, pansexual, asexual, and queer adolescents. *Computers in Human Behavior*, 119:106728, 2021. ISSN 0747-5632. doi: https://doi.org/10.1016/j.chb.2021.106728. URL https://www.sciencedirect.com/science/article/pii/S0747563221000509.

[58] Counter Hate. Twitter fails to act on twitter blue accounts tweeting hate, 2023. URL https://counterhate.com/research/twitter-fails-to-act-on-twitter-blue-accounts-tweeting-hate/.

[59] Counter Hate. Elon musk vs. center for countering digital hate: Nonprofit wins dismissal of 'baseless and intimidatory' lawsuit brought by world's richest man, marzo 2024. URL https://counterhate.com/blog/elon-musk-vs-ccdh-nonprofit-wins-dismissal-of-baseless-and-intimidatory-lawsuit/.

[60] X Help. X premium. URL https://help.x.com/es/using-x/x-premium.

[61] Daniel Hickey, Matheus Schmitz, Daniel Fessler, Paul E. Smaldino, Goran Muric, and Keith Burghardt. Auditing elon musk's impact on hate speech and bots. *Proceedings of the International AAAI Conference on Web and Social Media*, 17(1):1133–1137, Jun. 2023. doi: 10.1609/icwsm.v17i1.22222. URL https://ojs.aaai.org/index.php/ICWSM/article/view/22222.

[62] Daniel Hickey, Daniel M. T. Fessler, Kristina Lerman, and Keith Burghardt. X under musk's leadership: Substantial hate and no reduction in inauthentic activity. *PLOS ONE*, 20(2):1–24, 02 2025. doi: 10.1371/journal.pone.0313293. URL https://doi.org/10.1371/journal.pone.0313293.

[63] 24 Horas. Twitter cobrará 8 dólares por mes para certificar cuentas, 2022. URL https://24-horas.mx/negocios/twitter-cobrara-8-dolares-por-mes-para-certificar-cuentas/.

[64] Amnesty International. Lgbtiq: Significado de cada letra. URL https://www.es.amnesty.org/en-que-estamos/blog/historia/articulo/lgbtiq-significado-cada-letra/.

[65] Jupyter. Jupyter: Open-source software for interactive computing. URL https://jupyter.org/.

[66] JustAnotherArchivist. JustAnotherArchivist/snscrape, May 2025. URL https://github.com/JustAnotherArchivist/snscrape.

[67] Brendan Kennedy, Xisen Jin, Aida Mostafazadeh Davani, Morteza Dehghani, and Xiang Ren. Contextualizing hate speech classifiers with post-hoc explanation, 2020. URL https://arxiv.org/abs/2005.02439.

[68] LegalArmy. Nuevos límites en twitter para prevenir el data scraping, 2023. URL https://www.legalarmy.net/blog/nuevos-limites-en-twitter-para-prevenir-el-data-scraping.

[69] Ruth Lewis, Mike Rowe, and Clare Wiper. *Online/Offline Continuities: Exploring Misogyny and Hate in Online Abuse of Feminists*, pages 121–143. Springer International Publishing, Cham, 2019. ISBN 978-3-030-12633-9. doi: 10.1007/978-3-030-12633-9_5. URL https://doi.org/10.1007/978-3-030-12633-9_5.

[70] Andalucia Diversidad LGBT. Qué significa cisgénero, May 2022. URL https://andalucialgbt.com/que-significa-cisgenero/.

[71] R. Mac, T. Hsu, and B. Mullin. Twitter's new chief eases into the hot seat, 2023. URL https://www.nytimes.com/2023/.

[72] Edoardo Di Martino, Alessandro Galeazzi, Michele Starnini, Walter Quattrociocchi, and Matteo Cinelli. Characterizing the fragmentation of the social media ecosystem, 2024. URL https://arxiv.org/abs/2411.16826.

[73] Jacob Mchangama, Abby Fanlo, and Natalie Alkiviadou. Scope creep: An assessment of 8 social media platforms' hate speech policies, 2023. URL https://futurefreespeech.org/wp-content/uploads/2023/07/Twitter.pdf.

[74] BBC Mundo. Elon musk anuncia que twitter cobrará por verificación de cuentas tras afianzar su control de la compañía como director único, 2022. URL https://www.bbc.com/mundo/noticias-63463715.

[75] Elon Musk. Tweet by elon musk, 2022. URL https://x.com/elonmusk/status/1517215066550116354.

[76] United Nations. Targets of hate. URL https://www.un.org/en/hate-speech/impact-and-prevention/targets-of-hate. Publisher: United Nations.

[77] R. Nelson. Meta's threads crosses 150m downloads worldwide after seeing nearly 100m active users in first three days, 2023. URL https://www.dataai.com/blog/meta-threads-150m-downloads.

[78] AP News. Musk's twitter disbands its trust and safety advisory group, 2022. URL https://apnews.com/article/elon-musk-twitter-inc-technology-business-a9b795e8050de12319b82b5dd7118cd7.

[79] AP News. Twitter is the worst major social media platform when it comes to lgbtq+ safety, says glaad, 2023. URL https://apnews.com/article/elon-musk-lgbtq-social-media-twitter-glaad-aab0b650c858c177f34155ae3f5390c6.

[80] Agustín Nieto. Tweetscraperr. URL https://github.com/agusnieto77/TweetScraperR.

[81] Daniel Nkemelu, Harshil Shah, Michael Best, and Irfan Essa. Tackling hate speech in low-resource languages with context experts. In *Proceedings of the 2022 International Conference on Information and Communication Technologies and Development*, ICTD '22, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450397872. doi: 10.1145/3572334.3572372. URL https://doi.org/10.1145/3572334.3572372.

[82] M. Novak. Elon musk says 'cisgender' now considered a slur on twitter, 2023. URL https://www.forbes.com/sites/kimelsesser/2023/07/02/elon-musk-deems-cis-a-twitter-slurheres-why-its-is-so-polarizing/.

[83] NPR. Musk's twitter has dissolved its trust and safety council, December 2022. URL https://www.npr.org/2022/12/12/1142399312/twitter-trust-and-safety-council-elon-musk.

[84] M. Näsi, P. Räsänen, J. Hawdon, E. Holkeri, and A. Oksanen. Exposure to online hate material and social trust among finnish youth. *Information Technology & People*, 28(3):607–622, 2015. doi: 10.1108/ITP-09-2014-0198. URL https://doi.org/10.1108/ITP-09-2014-0198.

[85] Atte Oksanen, James Hawdon, Emma Holkeri, and Pekka Räsänen. *Exposure to Online Hate among Young Social Media Users*, volume 18, pages 253 – 273. 10 2014. ISBN 978-1-78441-060-5. doi: 10.1108/S1537-466120140000018021.

[86] A. Pluta, J. Mazurek, J. Wojciechowski, and et al. Exposure to hate speech deteriorates neurocognitive mechanisms of the ability to understand others' pain. *Scientific Reports*, 13:4127, 2023. doi: 10.1038/s41598-023-31146-1. URL https://doi.org/10.1038/s41598-023-31146-1.

[87] X Developer Portal. X developer portal - basic info. URL https://developer.x.com/en/portal/petition/essential/basic-info?plan=pro.

[88] Abhinav Pratap and Amit Pathak. From public square to echo chamber: The fragmentation of online discourse, 2025. URL https://arxiv.org/abs/2501.18441.

[89] Rainbow Project. Glosario de términos. URL http://www.rainbowproject.eu/material/es/glossary.htm.

[90] A. Ramírez-García, A. González-Molina, M. Gutiérrez-Arenas, and M. Moyano-Pacheco. Interdisciplinarity of scientific production on hate speech and social media: A bibliometric analysis. *Comunicar*, 72:129–140, 2022. doi: 10.3916/C72-2022-10. URL https://doi.org/10.3916/C72-2022-10. Also available in Spanish: *Interdisciplinariedad de la producción científica sobre el discurso del odio y las redes sociales: Un análisis bibliométrico*.

[91] Joaquín Amat Rodrigo. Comparación de distribuciones: test Kolmogorov–Smirnov. URL https://cienciadedatos.net/documentos/51_comparacion_distribuciones_kolmogorov%E2%80%93smirnov.

[92] Björn Roß, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. Measuring the reliability of hate speech annotations: The case of the european refugee crisis. 2016. doi: 10.17185/DUEPUBLICO/42132. URL https://duepublico2.uni-due.de/receive/duepublico_mods_00042132.

[93] RTVE. Mitos y realidades sobre las cuentas verificadas de twitter, November 2022. URL https://www.rtve.es/noticias/20221104/implicaciones-cuenta-twitter-verificada/2407980.shtml.

[94] Jim Rutenberg and Kate Conger. Elon musk is spreading election misinformation, but x's fact checkers are long gone. Gale Academic OneFile, January 2024. URL https://link.gale.com/apps/doc/A780599184/AONE?u=anon~b055f351&sid=googleScholar&xid=9e11fc08. Accessed 27 Apr. 2025.

[95] Kamal Safdar, Shibli Nisar, Waseem Iqbal, Awais Ahmad, and Yawar Abbas Bangash. Demographical based sentiment analysis for detection of hate speech tweets for low resource language. *ACM Transactions on Asian and Low-Resource Language Information Processing*, aug 2023. doi: 10.1145/3616867. URL https://doi.org/10.1145/3616867.

[96] Koustuv Saha, Eshwar Chandrasekharan, and Munmun De Choudhury. Prevalence and psychological effects of hateful speech in online college communities. In *WebSci '19: Proceedings of the 10th ACM Conference on Web Science*, pages 255–264, 2019. doi: 10.1145/3292522.3326032. URL https://doi.org/10.1145/3292522.3326032.

[97] Martin Saveski, Brandon Roy, and Deb Roy. The structure of toxic conversations on twitter. In *Proceedings of the Web Conference 2021*, WWW '21, page 1086–1097, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383127. doi: 10.1145/3442381.3449861. URL https://doi.org/10.1145/3442381.3449861.

[98] SimilarWeb. Top websites. URL https://www.similarweb.com/top-websites/.

[99] Wiktor Soral, Michał Bilewicz, and Mikołaj Winiewski. Exposure to hate speech increases prejudice through desensitization. *Aggressive Behavior*, 44 (2):136–146, 2018. doi: 10.1002/ab.21737. URL https://doi.org/10.1002/ab.21737.

[100] Wiktor Soral, James Liu, and Michał Bilewicz. Media of contempt: Social media consumption predicts normative acceptance of anti-muslim hate speech and islamoprejudice. *International Journal of Conflict and Violence (IJCV)*, 14:1–13, Mar. 2021. doi: 10.4119/ijcv-3774. URL https://www.ijcv.org/index.php/ijcv/article/view/3774.

[101] C. C. Tate, J. N. Bettergarcia, and L. M. Brent. Re-assessing the role of gender-related cognitions for self-esteem: The importance of gender typicality for cisgender adults. *Sex Roles*, 72(5):221–236, 2015. doi: 10.1007/s11199-015-0458-0. URL https://doi.org/10.1007/s11199-015-0458-0.

[102] Kurt Thomas, Devdatta Akhawe, Michael Bailey, Dan Boneh, Elie Bursztein, Sunny Consolvo, Nicola Dell, Zakir Durumeric, Patrick Gage Kelley, Deepak Kumar, Damon McCoy, Sarah Meiklejohn, Thomas Ristenpart, and Gianluca Stringhini. Sok: Hate, harassment, and the changing landscape of online abuse. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 247–267, 2021. doi: 10.1109/SP40001.2021.00028.

[103] Los Angeles Times. Elon musk's x faces scrutiny over harmful content amid transparency report, September 2024. URL https://www.latimes.com/business/story/2024-09-25/elon-musks-x-harmful-content-amid-transparcency-report-scrutiny-twitter.

[104] Social Media Today. New report suggests hate speech rising on twitter, November 2023. URL https://www.socialmediatoday.com/news/New-Report-Suggests-Hate-Speech-Rising-on-Twitter/645482/.

[105] Social Media Today. X increases its api access fees, October 2024. URL https://www.socialmediatoday.com/news/x-formerly-twitter-increases-api-access-fees/731151/.

[106] Twitter. Twitter rules, 2017. URL https://web.archive.org/web/20171218210508/https://help.twitter.com/en/rules-and-policies/twitter-rules.

[107] Twitter. Hateful conduct policy, 2018. URL https://web.archive.org/web/20181028023901/https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy.

[108] Twitter. Freedom of speech, not reach: An update on our enforcement policy, 2023. URL https://blog.twitter.com/.

[109] Naciones Unidas. ¿qué es el discurso de odio? URL https://www.un.org/es/hate-speech/understanding-hate-speech/what-is-hate-speech.

[110] Gloria del Valle. glorelvalle/socialhaterbert, February 2024. URL https://github.com/glorelvalle/socialhaterbert.

[111] J. Vanian and L. Kolodny. Elon musk and twitter face growing brand-safety concerns after execs depart, 2023. URL https://www.cnbc.com/2023/06/02/elon-musk-twitter-face-brand-safety-concerns-after-executives-depart.html.

[112] Joseph B. Walther. Social media and online hate. *Current Opinion in Psychology*, 45:101298, 2022. ISSN 2352-250X. doi: https://doi.org/10.1016/j.copsyc.2021.12.010. URL https://www.sciencedirect.com/science/article/pii/S2352250X21002505.

[113] X. An update to the twitter transparency center, 2021. URL https://blog.x.com/en_us/topics/company/2021/an-update-to-the-twitter-transparency-center. Accedido: 2023-04-27.

[114] X. Hateful conduct policy, 2023. URL https://help.x.com/en/rules-and-policies/hateful-conduct-policy.

[115] H. B. Zia, J. He, A. Raman, I. Castro, N. Sastry, and G. Tyson. Flocking to mastodon: Tracking the great twitter migration. *arXiv preprint arXiv:230214294*, 2023. URL https://arxiv.org/abs/2302.14294.

[116] Steven Zimmerman, Udo Kruschwitz, and Chris Fox. Improving hate speech detection with deep learning ensembles. In Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL https://aclanthology.org/L18-1404/.

[117] Julia El Zini and Mariette Awad. On the explainability of natural language processing deep models. *ACM Comput. Surv.*, 55(5), December 2022. ISSN 0360-0300. doi: 10.1145/3529755. URL https://doi.org/10.1145/3529755.

Apéndice A

Palabras clave

bisexual	bisexuales	homosexual	homosexuales	cisheteros
cishetero	transfobia	transfobo	transfoba	transfobas
transfobos	transfobico	transfobicos	transfobica	transfobicas
transexual	transexuales	transgenero	transgeneros	trans
transes	maricon	maricones	maricona	mariconas
gays	lesbiana	lesbianas	travesti	travestis
marimacho	marimachos	marimacha	marimachas	maricas
marica	travelo	travelos	travelas	travela
tortillera	tortilleras	homofobo	homofobico	homofobica
homofobicos	homofobicas	homofoba	homofobos	homofobas
homofobia	bollera	bolleras	bollerones	bolleron
cis	queer	queers	drags	drag
cisgenero	cisgeneros	heteros	hetero	lgbt
lgtb	lgtbq	lgtbi	lgtbqia	lgtbia
lgbt+	lgtb+	lgtbq+	lgtbi+	lgtbqia+
lgtbia+	homo	homos	mariquita	mariquitas
arromantico	arromanticos	arromantica	arromanticas	hetero
heteros	heteras	hetera	heterosexual	heterosexuales
pansexuales	pansexual	intersex	interexs	intersexual
asexual	asexuales	demisexual	demisexuales	afeminados
afeminado	no binario	no binarios	no binarias	no binaria
no binarie	no binaries	genero fluido	disforia de genero	gay

Tabla A.1: Términos utilizados para la descarga de tuits por palabras clave (explicado en la subsección 3.2.2) relacionadas con el colectivo LGTBIQ+

transfobia	transfobo	transfoba	transfobas	transfobos
transfobico	transfobicos	transfobica	transfobicas	transexual
transexuales	transgenero	transgeneros	trans	transes
travesti	travestis	travelo	travelos	travela
travelas	drag	drags	no binario	no binarios
no binarias	no binaria	no binarie	no binaries	genero fluido

Tabla A.2: Términos utilizados para el análisis de tuits relacionados con la comunidad trans (explicado en la subsección3.2.5)

Apéndice B

Estudio estadístico "likes" y "retuits"

Comparación	p-valor "likes"	p-valor "retuits"
2015-2019 vs 2020	0.0	<0.01
2015-2019 vs 2021-2022	0.0	<0.01
2015-2019 vs 2023-2024	0.0	<0.01
2020 vs 2021-2022	0,0942	<0.01
2020 vs 2023-2024	<0.01	<0.01
2021-2022 vs 2023-2024	<0.01	<0.01

Tabla B.1: P-valores para "likes" y "retuits" - dataset colectivo

Atributo	2015-2019	2020	2021-2022	2023-2024
Cuartil 25 %	0.00	0.00	0.00	0.00
Mediana (50 %)	0.00	1.00	1.00	1.00
Cuartil 75 %	2.00	3.00	3.00	2.00
Percentil 90 %	6.00	11.00	11.00	9.00
Percentil 95 %	13.00	25.00	24.00	23.00
Percentil 99 %	74.00	166.00	146.00	215.86
Moda	0.00	0.00	0.00	0.00
Media	6.61	14.90	13.58	21.55

Tabla B.2: Resultados de las estadísticas descriptivas de los "likes" para cada periodo - dataset colectivo

Atributo	2015-2019	2020	2021-2022	2023-2024
Cuartil 25 %	0.00	0.00	0.00	0.00
Mediana (50 %)	0.00	0.00	0.00	0.00
Cuartil 75 %	0.00	0.00	0.00	0.00
Percentil 90 %	2.00	2.00	2.00	2.00
Percentil 95 %	5.00	5.00	5.00	4.00
Percentil 99 %	30.00	37.00	30.00	42.00
Moda	0.00	0.00	0.00	0.00
Media	2.48	3.33	2.60	3.84

Tabla B.3: Resultados de las estadísticas descriptivas de los "retuits" para cada periodo - dataset colectivo

Comparación	p-valor "likes"	p-valor "retuits"
2015-2019 vs 2020	<0.01	<0.01
2015-2019 vs 2021-2022	<0.01	<0.01
2015-2019 vs 2023-2024	<0.01	<0.01
2020 vs 2021-2022	0,0102	<0.01
2020 vs 2023-2024	<0.01	<0.01
2021-2022 vs 2023-2024	<0.01	<0.01

Tabla B.4: P-valores para "likes" y "retuits" - dataset aleatorio

Atributo	2015-2019	2020	2021-2022	2023-2024
Cuartil 25 %	0.00	0.00	0.00	0.00
Mediana (50 %)	0.00	1.00	1.00	0.00
Cuartil 75 %	1.00	2.00	2.00	2.00
Percentil 90 %	3.00	6.00	7.00	6.00
Percentil 95 %	7.00	14.00	15.00	16.00
Percentil 99 %	37.00	89.85	99.00	143.00
Moda	0.00	0.00	0.00	0.00
Media	3.78	8.33	10.15	13.54

Tabla B.5: Resultados de las estadísticas descriptivas de los "likes" para cada periodo - dataset aleatorio

Atributo	2015-2019	2020	2021-2022	2023-2024
Cuartil 25 %	0.00	0.00	0.00	0.00
Mediana (50 %)	0.00	0.00	0.00	0.00
Cuartil 75 %	0.00	0.00	0.00	0.00
Percentil 90 %	1.00	1.00	1.00	1.00
Percentil 95 %	3.00	3.00	3.00	2.00
Percentil 99 %	17.00	23.00	19.00	23.00
Moda	0.00	0.00	0.00	0.00
Media	1.63	2.01	1.75	2.42

Tabla B.6: Resultados de las estadísticas descriptivas de los "retuits" para cada periodo - dataset aleatorio

Comparación	p-valor "likes"	p-valor "retuits"
2015-2019 vs 2020	<0.01	0,6692
2015-2019 vs 2021-2022	<0.01	<0.01
2015-2019 vs 2023-2024	<0.01	<0.01
2020 vs 2021-2022	0,6825	0,3317
2020 vs 2023-2024	0,0761	0,0297
2021-2022 vs 2023-2024	0,8460	0,9429

Tabla B.7: P-valores para "likes" y "retuits" - dataset aleatorio tóxico

Atributo	2015-2019	2020	2021-2022	2023-2024
Cuartil 25 %	0.00	0.00	0.00	0.00
Mediana (50 %)	0.00	0.00	0.00	0.00
Cuartil 75 %	1.00	1.00	1.00	1.00
Percentil 90 %	3.00	4.00	4.00	4.90
Percentil 95 %	5.00	9.00	8.00	10.00
Percentil 99 %	32.00	46.32	54.65	60.08
Moda	0.00	0.00	0.00	0.00
Media	4.08	33.10	3.44	6.58

Tabla B.8: Resultados de las estadísticas descriptivas de los "likes" para cada período - dataset aleatorio tóxico

Atributo	2015-2019	2020	2021-2022	2023-2024
Cuartil 25 %	0.00	0.00	0.00	0.00
Mediana (50 %)	0.00	0.00	0.00	0.00
Cuartil 75 %	0.00	0.00	0.00	0.00
Percentil 90 %	1.00	1.00	1.00	0.00
Percentil 95 %	2.00	2.00	1.00	1.00
Percentil 99 %	11.00	12.00	5.65	7.59
Moda	0.00	0.00	0.00	0.00
Media	1.65	9.22	0.52	0.67

Tabla B.9: Resultados de las estadísticas descriptivas de los "retuits" para cada período - dataset aleatorio tóxico

Comparación	p-valor "likes"	p-valor "retuits"
2015-2019 vs 2020	<0.01	<0.01
2015-2019 vs 2021-2022	<0.01	<0.01
2015-2019 vs 2023-2024	<0.01	<0.01
2020 vs 2021-2022	<0.01	<0.01
2020 vs 2023-2024	<0.01	<0.01
2021-2022 vs 2023-2024	<0.01	<0.01

Tabla B.10: P-valores para "likes" y "retuits" - dataset colectivo tóxico

Atributo	2015-2019	2020	2021-2022	2023-2024
Cuartil 25 %	0.00	0.00	0.00	0.00
Mediana (50 %)	0.00	0.00	0.00	0.00
Cuartil 75 %	1.00	2.00	1.00	1.00
Percentil 90 %	3.00	6.00	5.00	4.00
Percentil 95 %	6.00	14.85	11.00	9.00
Percentil 99 %	31.00	126.77	65.00	60.96
Moda	0.00	0.00	0.00	0.00
Media	2.38	12.19	5.01	8.84

Tabla B.11: Resultados de las estadísticas descriptivas de los "likes" para cada período - dataset colectivo tóxico

Atributo	2015-2019	2020	2021-2022	2023-2024
Cuartil 25 %	0.00	0.00	0.00	0.00
Mediana (50 %)	0.00	0.00	0.00	0.00
Cuartil 75 %	0.00	0.00	0.00	0.00
Percentil 90 %	1.00	1.00	1.00	0.00
Percentil 95 %	2.00	2.00	1.00	1.00
Percentil 99 %	9.00	23.00	10.00	8.00
Moda	0.00	0.00	0.00	0.00
Media	0.72	2.60	0.71	0.88

Tabla B.12: Resultados de las estadísticas descriptivas de los "retuits" para cada período - dataset colectivo tóxico

Comparación	p-valor "likes"	p-valor "retuits"
colectivo vs aleatorio	<0.01	<0.01
colectivo tóxico vs colectivo	<0.01	<0.01
aleatorio tóxico vs aleatorio	<0.01	<0.01
colectivo tóxico vs aleatorio tóxico	0.19	0.99

Tabla B.13: P-valores para "likes" y "retuits" entre todos los datasets

Apéndice C

Atributos de los datasets

Cada uno de los tuits de los datasets contiene 34 atributos. Dichos atributos están en inglés, en la tabla C.1 podemos encontrar los nombres de los atributos en inglés, su traducción al español y su explicación en español.

Un ejemplo de tuit contenido en el dataset del colectivo es el de la tabla C.2. Aquellas celdas en blanco indican que no se pudo recolectar el valor del atributo, ya sea por la no existencia del mismo para ese tuit o por otro tipo de fallo.

Como se puede observar en la tabla C.2, el texto es claramente ofensivo, y así lo muestran los atributos de toxicidad. De hecho, dichos atributos poseen valores extremos (> 0.7).

Column Name (English)	Nombre de la Columna (Español)
id	id
createdAt	creadoEn
source	fuente
lang	idioma
retweetCount	numeroRetweets
replyCount	numeroRespuestas
likeCount	numeroLikes
quoteCount	numeroCitas
viewCount	numeroVistas
bookmarkCount	numeroFavoritos
isReply	esRespuesta
conversationId	idConversacion
author_verified	autorVerificado
author_blue_verified	autorVerificadoAzul
author_followers	seguidoresAutor
author_following	siguiendoAutor
author_tweets	tuitsAutor
author_createdAt	creadoEnAutor
media_urls	urlsMedios
hashtags	hashtags
user_mentions	mencionesUsuarios
author_isAutomated	autorEsAutomatizado
author_fastFollowersCount	seguidoresRapidos
author_favouritesCount	favoritosAutor
author_hasCustomTimelines	autorTieneCronologiasPersonalizadas
author_statusesCount	estadosAutor
place_country_code	codigoPaisLugar
texto_analisis	textoAnalisis
toxicity	toxicidad
severe_toxicity	toxicidadSevera
identity_attack	ataqueIdentidad
insult	insulto
profanity	obscenidad
threat	amenaza

Tabla C.1: Atributos de los datasets colectivo y aleatorio

Attribute	Value
id	615261867162902529
createdAt	Sun Jun 28 20:53:54 +0000 2015
source	Twitter for iPhone
lang	es
retweetCount	0
replyCount	1
likeCount	0
quoteCount	0
viewCount	0
bookmarkCount	0
isReply	True
conversationId	615261809734500354
author_verified	False
author_blue_verified	False
author_followers	234
author_following	566
author_tweets	112596
author_createdAt	Sat Mar 29 19:32:04 +0000 2014
media_urls	
hashtags	
user_mentions	morrosporruos
author_isAutomated	False
author_fastFollowersCount	0
author_favouritesCount	79387
author_hasCustomTimelines	True
author_statusesCount	112596
place_country_code	
texto_analisis	a ver si te voy a tener que matar, maricón
toxicity	0.968226
severe_toxicity	0.8219337
identity_attack	0.73072165
insult	0.91372794
profanity	0.8802007
threat	0.92211

Tabla C.2: Ejemplo tuit del dataset del colectivo