CENTRO DE INVESTIGACIÓN EN
CIENCIAS DE INFORMACIÓN GEOESPACIAL, A.C.
CENTROGEO

Centro Público de Investigación CONAHCYT

Detección y monitoreo de discurso de odio hacia la comunidad LGBT en Twitter, un acercamiento desde el aprendizaje computacional.

TESIS

Que para obtener el grado de

Maestro en Ciencias de Información Geoespacial

Presenta

Guillermo Gerardo Vazquez Ferrer

Directora y Co-director de Tesis

Mtra. Karime González Zuccolotto y Dr. Oscar Gerardo Sánchez Siordia

Ciudad de México marzo 2024

CENTRO DE INVESTIGACIÓN EN

CIENCIAS DE INFORMACIÓN GEOESPACIAL, A.C.

CENTROGEO

Centro Público de Investigación CONAHCYT

Detección y monitoreo de discurso de odio hacia la comunidad LGBT en Twitter, un acercamiento desde el aprendizaje computacional.

TESIS

Que para obtener el grado de

Maestro en Ciencias de Información Geoespacial

Presenta

Guillermo Gerardo Vazquez Ferrer

Directora de Tesis

Mtra. Karime González Zuccolotto

Co-director de Tesis

Dr. Oscar Gerardo Sánchez Siordia

Sinodales

Mtra. Karime González Zuccolotto

Dr. Oscar Gerardo Sánchez Siordia

Dr. Gandhi Samuel Hernández Chan

Dr. Hugo Carlos Martínez

Ciudad de México, marzo de 2024

Resumen

La discriminación y el discurso de odio hacia la comunidad LGBT representan lamentables realidades sociales que persisten en diversos contextos. Sin embargo, la comprensión y abordaje de estos fenómenos se ven obstaculizados por la escasez de fuentes de datos oficiales específicas y estructuradas que permitan un análisis exhaustivo desde una perspectiva espacial. La falta de información estandarizada dificulta la identificación de patrones geográficos y la formulación de estrategias efectivas para contrarrestar estas problemáticas. Ante este desafío, la necesidad de recurrir a fuentes de datos alternativas no estructuradas se vuelve imperativa para explorar y entender la discriminación y el discurso de odio contra la comunidad LGBT desde una perspectiva espacial. Estas fuentes alternativas ofrecen la posibilidad de abordar dichas problemáticas desde ángulos novedosos y, al mismo tiempo, constituyen herramientas valiosas para aquellos comprometidos con la promoción de la igualdad y la diversidad.

La presente investigación, se centra en evaluar y comparar el rendimiento de dos modelos de aprendizaje superficial y dos modelos de aprendizaje profundo que por medio de procesamiento del lenguaje natural permitan la identificación y clasificación de discurso de odio dirigido a la comunidad LGBT en textos cortos de Twitter con el objetivo de zonificar la presencia de este tipo de mensajes en el territorio mexicano. Estos modelos corresponden a: 1) Regresión Logística (LR), 2) Máquinas de Soporte Vectorial (SVM), 3) Redes Neuronales Convolucionales (CNN) y 4) Redes Neuronales Recurrentes (RNN).

Para la construcción de los modelos, se implementó un corpus etiquetado de 8,000 tuits escritos en español en los que se hace mención a la comunidad LGBT con una temporalidad entre el 2020 y 2022; este corpus fue dividido 80/20 para la fase de entrenamiento y validación respectivamente, facilitando la identificación y monitoreo de los estados del territorio mexicano con mayor presencia de tuits de discurso de odio LGBT.

Tanto el modelo de LR como SVM fueron construidos con una matriz de características con ayuda de sklearn y un ajuste vectorial por medio del recuento de palabras (TF-IDF), por su parte, CNN y RNN se crearon con un modelo Word2Vec. Los resultados obtenidos señalan que, la utilización de TF-IDF muestra mejor rendimiento en la identificación de discurso de odio LGBT en textos cortos de Twitter respecto a Word2Vec. El modelo basado en LR ha obtenido un Accuracy de 0.9226 y un F1-Score de 0.9223, mientras que SVM obtuvo un Accuracy de 0.9321 y un F1-Score de 0.9302, sin embargo, LR muestra mejor respuesta al sobreajuste (overfitting). En cuanto a los modelos de aprendizaje profundo, CNN obtuvo un Accuracy de 0.8773 y un F1-Score de 0.8782, mientras que RNN obtuvo un 0.8848 y 0.8812 de Accuracy y F1-Score respectivamente.

Durante los 3 años analizados, se observó un comportamiento símil en la presencia de mensajes efectuados desde un discurso de odio hacia la comunidad LGBT. Aunque el 2020 registra una mayor cantidad de discurso de odio, la disminución en años posteriores es poca, además, se reconoció que el porcentaje de tuits con discurso de odio respecto a aquellos que no lo presentan, mantiene fluctuaciones ligeras a lo largo del año, exceptuando al mes de junio donde el porcentaje de discurso de odio decrece considerablemente. Finalmente, se destaca que la Ciudad de México, Estado de México, Nuevo León y Jalisco, son los estados con mayor presencia de discurso de odio respectivamente. Sin embargo, al considerar los valores obtenidos en tasas, sobresalen estados como Tamaulipas o San Luis Potosí al obtener un valor aproximado de 70 tuits identificados como discurso de odio por cada 100 mensajes referentes a la comunidad LGBT.

Índice

1 Marco conceptual

1.1 Discurso de odio, actos de violencia y discriminación en medios masivos de comunicación

Aunque la comunicación agresiva o con un carácter violento muy probablemente ha estado presente desde las primeras interacciones sociales, recientemente con el auge de los medios masivos de comunicación como Twitter y otras redes sociales, la reproducción y propagación de mensajes agresivos, de odio o que inciten a la violencia ha aumentado exponencialmente en la última década (Miró Linares, 2016).

Este hecho se incrementa por la facilidad de búsqueda y difusión de información y contenido tanto escrito como oral y visual sin comprometer el anonimato de los usuarios, permitiendo una expresividad genuina al concebir pocas o ningunas restricciones ante la publicación, difusión o consumo de este tipo de contenido, factor impulsado en gran medida por la dificultad de monitorear las comunicaciones expresadas en este tipo de plataformas (Jacks y Adler, 2015).

La Comisión Nacional de los Derechos Humanos reconoce que, toda persona tiene derecho a la libertad de pensamiento y expresión, es decir, que existe la libertad de buscar, recibir y difundir ideas, opiniones e información, impidiendo la censura previa en cualquier medio de difusión. Este derecho se encuentra sustentado a través del artículo 6° y 7° de la constitución política de los estados unidos mexicanos y el artículo 19 de la Declaración Universal de los Derechos Humanos, sin embargo, esta libertad conlleva deberes y responsabilidades expresadas en la ley (CNDH, 2022 y Gobierno de México, 2016).

La importancia de considerar la libertad de expresión al analizar el discurso de odio, radica en lo conflictivo que resulta la identificación de una línea divisoria que separe claramente lo que es expresado desde el discurso de odio y lo que no lo es. Aunque la libertad de expresión no protege discursos discriminatorios como el racismo, la homofobia, la xenofobia o el antisemitismo, el discurso de odio puede efectuarse de formas sutiles, dificultando la concepción de una definición absoluta de este (MacAvaney et al., 2019).

Para reconocer de manera más clara lo que es el discurso de odio, Miró Linares (2016), habla de la comunicación violenta como cualquier forma de expresión que se conceptualice o reconozca como violenta incluso aunque no conlleve algún motivo discriminatorio, es decir, tanto la incitación, la justificación o la valoración positiva de la violencia física, como la realización o incitación a la discriminación, la humillación o las injurias, entre otras.

Como se observa en la Figura 1.1, la comunicación violenta puede ser clasificada en dos grandes categorías: 1) las relacionadas con algún tipo de violencia moral entendida como la afectación individual o colectiva al dañar y ofender la reputación, los sentimientos, la vida privada, el honor o la percepción de los demás a través de la discriminación, la humillación o las burlas, etcétera; y 2) las relacionadas a la violencia física, entendidas como todas aquellas en donde se incita, defiende o promueve la realización de actos violentos que ejerzan daño físico a una persona o grupo de personas.

Figura 1.1 Categorías básicas de la comunicación violenta. Modificado de Miró Linares (2016)

En cuanto al discurso de odio, algunas definiciones como la del Tribunal Europeo de Derechos Humanos, resaltan que debe existir una incitación directa o indirecta a cometer algún acto de violencia. Sin embargo, algunos autores como López Ulla (2018) reconocen que para hablar de un discurso de odio no es necesario realizar un llamado a la ejecución de un acto violento o criminal, pues, aunque el discurso de odio sí pretende incentivar la violencia hacia los integrantes de un colectivo de personas por motivos de género, etnicidad, nacionalidad, opiniones políticas u orientación sexual, entre otros. El discurso de odio también pretende rechazar, denigrar, intimidar, humillar, acosar y estigmatizar a través de la promoción de prejuicios, la incitación al odio y/o la discriminación (Tabares Higuita, 2018).

En otras palabras, el discurso de odio es toda expresión que incite a cualquier tipo de violencia y al odio, que menosprecie o denigre, y que promueva estereotipos de una colectividad en específico, incluso en formas sutiles como la ironía o el humor. Además, es posible localizar este tipo de discurso en expresiones textuales, verbales, en acciones comunicativas o a través de símbolos, entre otros (Kasianczuk M., 2022).

Así, aunque en la clasificación de la comunicación violenta Miró Linares (2016), posiciona al discurso de odio dentro de la categoría “Incitación a la discriminación/odio” (véase Figura 1.1), de igual manera

reconoce que, tanto la incitación directa o indirecta a la ejecución de violencia física, como las ofensas o daños al honor y dignidad pueden encontrarse inmersas en este tipo de discurso.

Aunado a ello, es posible examinar que dicho tipo de comunicación ha permeado en múltiples y variados medios de comunicación, incluidas las redes sociales que sobresalen como uno de los principales medios de difusión de discurso de odio, en donde la preocupación retoma dos grandes vertientes; por un lado la difusión de discurso de odio y material de odio a miles de personas, y por otro lado la posibilidad de que sean miles de personas quienes comuniquen y compartan odio y violencia a través de redes sociales (MacAvaney et al., 2019) (Miró Linares, 2016).

En este sentido se resalta el trabajo de Fortuna y Nunes (2018), en donde identifican conceptos clave de la comunicación violenta enfocados en los medios masivos de comunicación, así como su relación con el discurso de odio. Como se puede observar en la Tabla 1.1, el discurso de odio puede hacer uso o estar inmerso en la mayoría de estos conceptos, tales como el extremismo, el lenguaje tóxico, o la discriminación, dejando de lado al ciberacoso, el odio, o el flameo por estar dirigidos a un individuo en específico o realizar una comunicación violenta sin un motivo concreto.

Sin embargo, al considerar que el discurso de odio pretende incitar la violencia y/o discriminación de una colectividad o miembros de esta. Es posible mencionar que, dentro del ciberacoso, el flameo y el odio, puede existir discurso de odio mientras la pertenencia a cierto grupo de personas sea un punto clave al ser víctima de este tipo de comunicaciones violentas.

Concepto	Definición	Dentro de discurso de odio
Odio	Expresión de hostilidad sin alguna explicación.	Se centra en estereotipos, y no en expresar odio de manera general.
Ciberacoso	Acto agresivo e intencional, que se realiza repetidamente y a lo largo del tiempo, contra una víctima que no puede defenderse fácilmente por sí misma.	El discurso de odio no se centra necesariamente en una persona en específico.
Discriminación	Proceso a través del cual se identifica una diferencia que es utilizada como base para un trato injusto.	El discurso de odio es una forma de discriminación.
Flameo	Comentarios hostiles, e intimidatorios que pueden perturbar la participación en una comunidad.	El flameo está dirigido a un participante en el contexto específico de una discusión.
Lenguaje abusivo	Lenguaje hiriente, despectivo y blasfemias.	El discurso de odio es un tipo de lenguaje abusivo.
Blasfemias	Palabras o frases ofensivas u obscenas.	El discurso de odio puede usar blasfemias, pero no estrictamente.

Comentarios o lenguaje tóxico	Es un lenguaje o comentarios irrespetuosos que pueden hacer que una persona abandone una discusión.	El discurso de odio puede usar comentarios o lenguaje tóxico.
Extremismo, radicalización	Ideología asociada con extremistas o grupos de odio, donde se promueve la violencia, y a menudo con el objetivo de segmentar la población y reclamar estatus.	Los discursos extremistas utilizan con frecuencia discursos de odio. Sin embargo, estos discursos también se enfocan en otros temas, como el reclutamiento de nuevos miembros.

Tabla 1.1 Conceptos clave de la comunicación violenta en medios masivos de comunicación. Modificado de Fortuna y Nunes (2018)

Respecto a la violencia dentro del discurso de odio, es importante señalar que esta puede subdividirse en dos grandes ramas: 1) la violencia física entendida como toda agresión ya sea intencional o no que cause algún tipo de daño o dolor físico, como golpes, quemaduras o abuso sexual, entre otras; y 2) la violencia psicológica ejercida a través de la alteración de la estabilidad emocional por medio de insultos, humillaciones, faltas de respeto o cualquier otro tipo de agresión verbal (López de Loera, 2019).

Por su parte, la discriminación es entendida como el conjunto de prácticas y acciones que niegan el trato igualitario de una persona o grupo de personas por pertenecer o encajar en grupos sociales sobre los que recaen prejuicios u opiniones sociales negativas, generando desigualdad social o incluso la privación de sus derechos (Solís, 2017).

El factor clave de esta visión del discurso de odio es la intención de ejercer o incentivar a la realización de un daño ya sea físico, psicológico o por discriminación a los miembros de una colectividad o hacia algún integrante de esta solo por el hecho de pertenecer al grupo objetivo. Por lo cual, tanto la comunicación referida a algún tipo de violencia como la comunicación que englobe alguna referencia discriminatoria, se pueden considerar comunicaciones dentro de un discurso de odio (véase Figura 1.2).

Figura 1.2 Discurso de odio, violencia y discriminación. Elaboración propia con información de López de Loera (2019), Solís (2017), Miró Linares (2016) y Arcila, Blanco-Herrero y Valdez (2020)

La principal preocupación con el fácil y rápido esparcimiento de este tipo de mensajes, favorecido por el gran alcance de los medios masivos de comunicación y en específico de las redes sociales, radica en la posible correlación con la prevalencia de la violencia, la discriminación, la negación de derechos, la persecución y agresión contra algún grupo o persona objetivo, esto debido en gran medida a la facilidad de difundir mensajes negativos o de odio mucho más personalizados, así como la posible conexión entre personas que mantengan ideas afines que promuevan o apoyen cierto discurso de odio, enalteciendo este tipo de comunicación (Rodríguez Zepeda, 2018).

Por tal motivo, se reconoce la importancia de identificar la prevalencia de este tipo de mensajes dentro de los medios masivos de comunicación, no sólo por el impacto que puedan generar en las personas afectadas y la sociedad en general, sino también, para poder identificar localizaciones específicas en el espacio, donde este tipo de discursos represente un mayor problema respecto a otros espacios, así como reconocer posibles factores que influyen en la mayor o menor prevalencia de este tipo de comunicación.

1.2 Discurso de odio contra personas pertenecientes a la comunidad LGBT

Bajo el contexto de una sociedad mayormente heteronormativa, toda persona que se aleje de los esquemas binarios del género, de la heterosexualidad o que no cumplan con los estándares de masculinidad y feminidad, y que además dichos estándares vayan acorde al sexo biológico de la persona, muy probablemente experimentarán momentos o procesos de exclusión, segregación y/o violencia (da Silva Anes, 2021).

El caso del discurso de odio no es diferente, las personas que integran la comunidad LGBT se encuentran expuestas a altos niveles de agresión verbal, el cual puede decantar en un considerable daño psicológico. El estudio del discurso de odio hacia esta comunidad resalta al considerar que, en comparación con las personas heterosexuales cisgénero[1], los miembros de la comunidad LGBT son más propensos a sufrir mayor exposición a experimentar amenazas, delitos de odio o vulneración de derechos, impulsando efectos nocivos como depresión, agotamiento, trastornos del sueño, angustia emocional, ataques de pánico, problemas de adicciones o de desenvolvimiento social impulsando el deseo de aislamiento o la afectación a la realización de las actividades diarias (Ștefăniță y Buf, 2021).

Además, la prevalencia de este tipo de discursos, puede promover en las y los miembros de esta comunidad el desarrollo de un sentido de vergüenza, culpabilidad y ansiedad, impulsando el deseo de no pertenecer a esta comunidad o de aislarse como un mecanismo de defensa a través de la eliminación de cuentas personales en redes sociales. De igual manera, el recibimiento constante de mensajes de odio, pueden generar una estigmatización hacia la comunidad LGBT interiorizada, es decir, que los propios integrantes de la comunidad pueden adoptar actitudes negativas hacia esta, incitando también a la negación de su verdadera orientación sexual o identidad de género (Ștefăniță y Buf, 2021).

Esto no es exclusivamente resultado de la existencia del discurso de odio per se, la facilidad de su difusión, o el supuesto anonimato y la fácil creación de perfiles falsos, sino también, de la fácil difusión alrededor del mundo, y de la posibilidad de impulsar o resonar este tipo de discursos a través del apoyo, o acciones inherentes a las redes sociales como la acción "*me gusta*" (da Silva y da Silva, 2021).

Reconocer las diversidades sexo-genéricas como el conjunto de posibilidad que tienen las personas de asumir, expresar y vivir tanto su sexualidad como su identidad de género, muestra un camino de posibilidades que rompen con los estereotipos impuestos socialmente en cuanto a lo aceptado dentro de las preferencias sexuales, identidades de género o expresiones del mismo (Consejo Nacional de Población, 2022).

[1] La palabra cisgénero es utilizada para referirse a una persona cuya identidad de género coincide con la genitalia asignada al nacer.

En este sentido, es necesario reconocer que ni la orientación sexual de las personas determina su expresión de género, y que tampoco, su identidad o expresión de género es determinada por su sexo biológico. La Figura 1.3 muestra de manera general la forma en que se construye esta diversidad sexo-genérica:

- Identidad de género: Es la forma en cómo una persona se percibe de manera interna y cómo interpreta lo que le rodea.
- Orientación sexual: Se refiere a la atracción física o emocional hacia otras personas.
- Sexo asignado al nacer: Es la expresión de los órganos genitales al nacer.
- Expresión de género: Es la forma de expresar la identidad de género a través de acciones como la vestimenta o el comportamiento, etcétera.

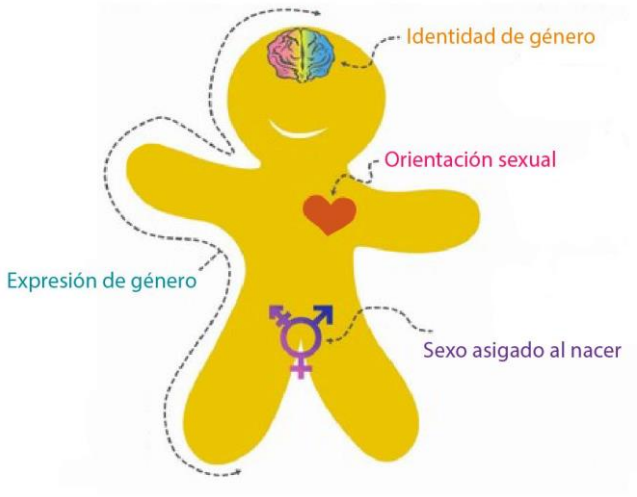


Figura 1.3 Persona de la diversidad sexo-genérica. Modificado de Killermann (2011)

De esta forma, "considerar la diversidad sexo-genérica desde el ejercicio de libertad sexual, como camino para romper la homogeneidad que la sociedad propone hasta ampliar el abanico de elección sexual. Aluden al derecho de todas las personas a poder optar desde la propia voluntad a expresar legítimamente su orientación sexual e identidad de género" (Sánchez y Torrejón, 2021:259).

Pero, ¿quiénes integran esta diversidad sexo-genérica? A pesar de que la heterosexualidad y las identidades de género binarias (hombre, mujer), se consideran parte de la diversidad sexo-genérica, el acrónimo de siglas LGBTTTIQ+ suele hacer referencia a todas aquellas identidades y preferencias sexuales no normativas. Este conjunto de siglas está conformado por las palabras: (L) lesbiana, (G) gay, (B) bisexual, (T) transgénero, (T) transexual, (T) travesti, (I) intersexual y (Q) queer, agregando el símbolo "+" al final del acrónimo a manera de incluir al resto de colectivos que no se encuentran representados con las siglas antes mencionadas (Prince Torres, 2021).

No obstante, aunque este conjunto de siglas engloba un gran número de colectividades de la diversidad sexo-genérica, en muchas ocasiones suele ser acotado a las siglas LGBT, como un mecanismo de delimitar o englobar el resto de colectivos. En este sentido, para el desarrollo de la presente investigación, se ha acotado el acrónimo a las siglas LGBT como una forma de enfocar los esfuerzos en la identificación de

discurso de odio en medios masivos de comunicación hacia personas lesbianas, gays, bisexuales, transexuales, transgénero y travestis (véase Figura 1.4).

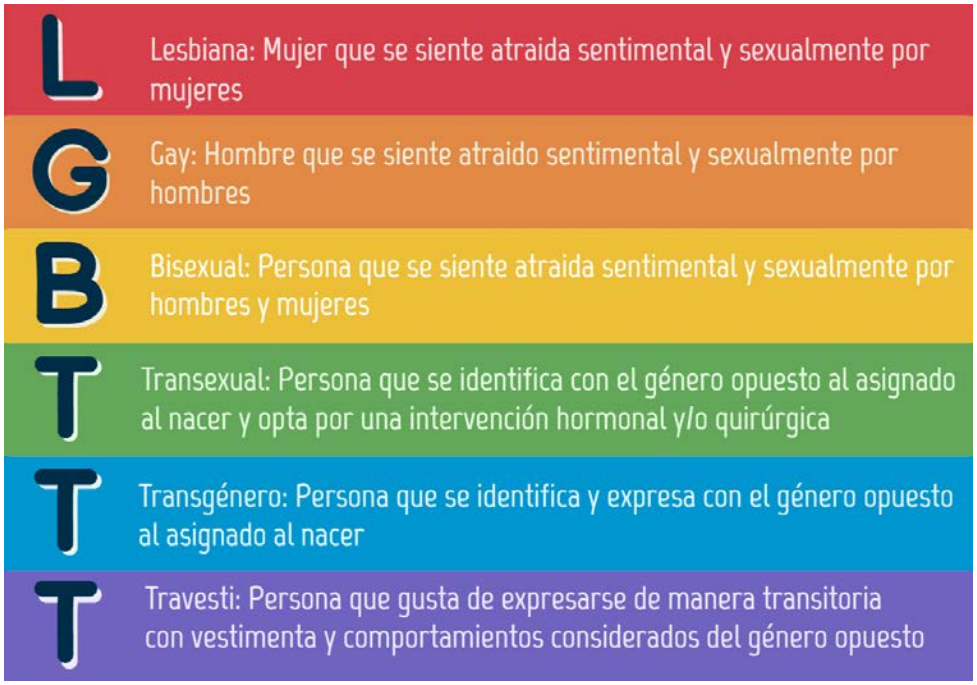


Figura 1.4 Población de estudio "LGBT". Elaboración propia con información de Prince Torres (2021)

El principal motivo de esta delimitación está relacionado a dos grandes factores; inicialmente por la mayor visibilidad y reconocimiento de estos colectivos por parte de los y las agresoras, y, por otra parte, respondiendo a la estereotipación que pueda existir hacia el resto de colectividades, en cuyo caso, se busca encasillar en el acrónimo LGBT al resto de preferencias sexuales, identidades o expresiones de género.

1.2.1 Estigmas, prejuicios y segregación

A lo largo de la historia, la población identificada con una orientación sexual o identidad de género no hegemónica o normativa, ha sido víctima de una constante segregación, discriminación y vulneración de derechos, generando una problemática que atraviesa diferentes contextos, desde lo social, lo económico, lo político, lo familiar o lo educativo (Jugănaru, 2018).

Esta segregación y discriminación ha sido un proceso históricamente sistemático, que encontró un respaldo en los discursos y prácticas dominantes a finales del siglo XIX y principios del siglo XX (Giribuela, 2018). Ejemplo de esto, fueron los discursos médicos, religiosos y del Estado que enmarcaban las disidencias sexuales en el terreno de lo enfermo, lo inmoral, lo anormal o lo delictivo, destacando que toda aquella concepción alejada de la heterosexualidad, necesitaba ser tratada, corregida y curada (Giribuela, 2019).

Esta visión de las diversidades sexo-genéricas ha cambiado con el paso de los años, en gran medida por la lucha social de visibilizar y reconocer a las personas pertenecientes a la comunidad LGBT, con lo que se ha buscado reducir la estigmatización de este colectivo, a través de acciones como la eliminación de la

homosexualidad de la Clasificación Internacional de Enfermedades (CIE) en 1990, o la exclusión de la transexualidad de esta misma lista en el 2018 (de Benito, 2018).

Sin embargo, la discriminación y estigmatización de este grupo poblacional ha continuado hasta la actualidad, pues, aunque hoy en día las personas LGBT se encuentran integradas de mejor manera en diversos ámbitos sociales, aún existen espacios donde el reconocimiento de una preferencia sexual o identidad de género no normativa puede representar un riesgo de discriminación o incluso de actos de violencia (Jugănaru, 2018).

La continuidad de este tipo de discursos prevalece en gran medida a la visión de una sociedad heteronormativa, en donde, las sexualidades no reproductivas debían ser ocultadas, silenciadas e invisibilizadas, a través de una estrategia ejercida desde el poder con el objetivo de eliminar diferencias o minorías, a través de un proceso de estereotipación, violencia simbólica y deslegitimación, con lo cual, se busca consolidar una representación de las diversidades sexuales y de género que desacredite a esta población, buscando la eliminación de estos grupos minoritarios, por lo menos de manera simbólica (Giribuela, 2018).

Como reconoce Ferentinos (2016), la cultura occidental ha mantenido una tendencia hacia lo binario, y el caso de la diversidad sexual y del género no es diferente; social y culturalmente ha sido más fácil englobar en los binarios heterosexual-homosexual y masculino-femenino, dejando de lado el resto de diversidades sexo-genéricas como el colectivo bisexual, transgénero, intersexual o de género fluido, etc. Esta visión binaria de la diversidad, representa nuevamente una segregación para el resto de personas pertenecientes a la comunidad LGBTTTIQ+ que no se reconocen única y exclusivamente como homosexuales.

Aunado a ello, es imperativo recordar que socialmente se han establecido creencias que determinan el comportamiento aceptado de acuerdo al sexo y género, así como los roles y actividades que son aceptados y que deben ser cumplidos, reproduciendo una imagen de lo que es aceptado y lo que no lo es. Esta visión genera prejuicios sobre las disidencias sexo-genéricas al identificar características o comportamientos sobre los que existe un rechazo o desagrado, inspirando una impresión generalmente negativa simplemente por pertenecer a un grupo, suponiendo cualidades, características o condiciones que son mal vistas o poco aceptadas socialmente (Bolaños y Chanrry, 2018).

Es necesario señalar que la comunidad LGBT no sólo se encuentra inmersa en estigmas y prejuicios que se han establecido para esta comunidad, y más bien, reconocer que las y los integrantes pueden ser parte de diversos colectivos segregados y estigmatizados ya sea por el color de piel, la raza, la condición económica o la situación migratoria, entre otras. Por lo que, la violencia y odio percibido puede responder a más de una característica, condición o preferencia (da Silva Anes, 2021).

1.2.2 De la discriminación a la violencia

La intolerancia generalizada hacia las personas auto percibidas o identificadas como parte de la diversidad sexo-genérica, legitima la violencia y discriminación hacia las personas de esta colectividad, resultado de prejuicios sociales y culturales que se han arraigado históricamente en la sociedad, con lo cual, es posible señalar que el trato desigual no es un proceso aleatorio, sino que se relaciona estrechamente a perfiles

construidos socialmente por estereotipos y prejuicios (Comisión Internacional de Derechos Humanos, 2018).

Esta discriminación no sólo funge como un medio de inequidad entre personas al no percibir un trato igualitario de derechos y generando una reproducción social desigual, sino también, como un hecho que genera una desvalorización de las capacidades y aptitudes de los integrantes, en muchas ocasiones obteniendo como resultado planes de vida truncados. Además, dicha discriminación se vive de manera estructural, permeando tanto en la vida pública como privada y personal (Solis, 2017).

De acuerdo a datos del Instituto Nacional de Estadística y Geografía (INEGI), para el año 2021, el 40.7% de la población mexicana consideraba que es poco el respeto que existe a los derechos de las personas lesbianas, gays, bisexuales y transexuales (Encuesta Nacional sobre Diversidad y de Género (ENDISEG, 2021)) (véase Figura 1.5).

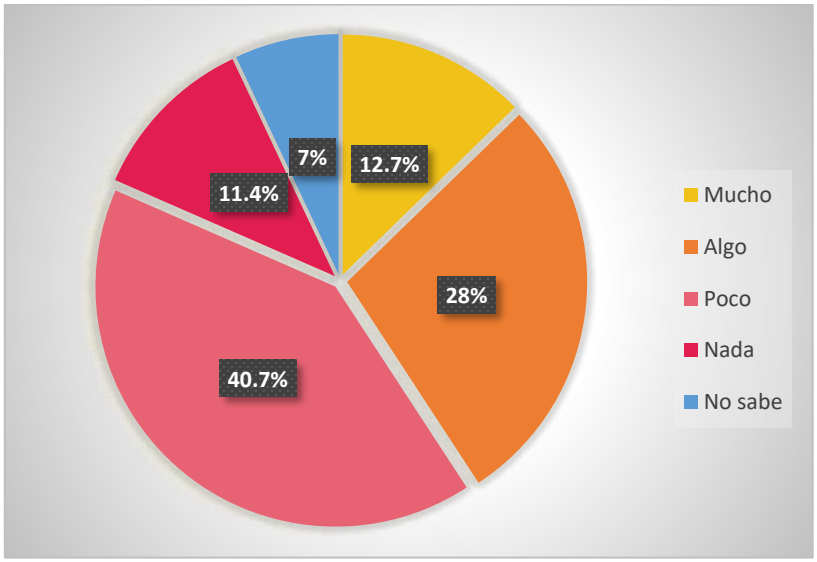


Figura 1.5 Población de 15 años y más, de acuerdo a la opinión que tienen sobre el grado de respeto que existe en México a los derechos de las personas lesbianas, gays, bisexuales y transexuales. Elaboración propia con información de ENDISEG (2021)

Aun cuando dentro de la discriminación, el trato desigual del acceso a derechos es un factor clave, el mantener un trato o actitud que busque anular, ignorar o invisibilizar la existencia de este colectivo es reconocido como un hecho sustancial dentro de esta discriminación. La Comisión Internacional de Derechos Humanos (2018:33), menciona que “no reconocer la existencia de las personas LGBT y privarles de la protección que todas las demás personas tienen, las deja en una situación de absoluta vulnerabilidad a las diversas formas de desigualdad, discriminación, violencia, y exclusión”.

La discriminación ejercida a las personas LGBT en cuanto al escaso reconocimiento, no hace referencia únicamente a la visión pública de su existencia, sino también a la escasa participación en ciertas esferas sociales y públicas, ejemplo de esto, es la poca incidencia política, incrementando la posibilidad de crear barreras de accesibilidad a bienes y servicios como la salud, empleo o enfrentando procesos de impunidad,

y a la vez, manteniendo espacios poco incluyentes que eleven el peligro y exposición a sufrir algún tipo de violencia.

La recolección de información y estadísticos que muestren la realidad de las personas pertenecientes a la comunidad LGBT, es de igual manera un proceso discriminatorio que se ha generado gracias a los estigmas y el poco reconocimiento. Si bien existen esfuerzos como la ENDISEG, aún existe una gran brecha en cuanto al estudio y representación de la comunidad LGBT a través de estadísticas y bases de datos.

Como lo menciona la Comisión Internacional de Derechos Humanos (2018), es de gran importancia la recolección de datos sobre las personas pertenecientes a la comunidad LGBT, y a la vez, generar estadísticas confiables sobre los actos de violencia y discriminación a los que son sujetos, ya que pueden fungir como instrumentos esenciales para visibilizar y zonificar espacios de conflicto o con mayor presencia de discriminación y/o violencia, demandando a los gobiernos el reconocimiento y respeto de sus derechos.

Estos factores impulsan la prevalencia de ambientes hostiles que obstaculicen el desarrollo integral de las personas LGBT objeto de esta discriminación y violencia, y a la vez, facilitan la prevalencia de tal problemática; hecho que genera un estado de vulneración facilitando la ocurrencia de actos violentos tanto físicos como psicológicos.

1.3 Derechos LGBT

En México, tanto la Declaración Universal de los Derechos Humanos como la constitución Política de los Estados Unidos Mexicanos garantizan la igualdad de derechos a toda persona sin importar su identidad de género o preferencia sexual, sin embargo, en diversos espacios ya sean públicos o privados, las leyes locales y los estigmas o prejuicios que ejercen cualquier tipo de discriminación hacia una persona perteneciente a la diversidad sexo-genérica, dificulta el acceso o ejercicio de sus derechos con la misma facilidad que otras personas que no se reconocen como parte de este grupo poblacional. Aun cuando todas las personas tienen derechos y libertades fundamentales por el simple hecho de existir, se crean barreras y limitantes socioculturales que impiden el acceso y ejercicio igualitario de los derechos humanos (García Patiño, 2020).

El marco normativo internacional relacionado a la igualdad, la no discriminación, el acceso a servicios y bienes educativos, de trabajo y de la salud engloban un trato igualitario hacia las personas pertenecientes a la comunidad LGBT. Este marco normativo ha buscado impulsar el cumplimiento efectivo de los derechos humanos y a la vez, buscar el reconocimiento, reivindicación y protección de las personas LGBT al buscar eliminar la estigmatización y violencia contra este grupo poblacional, impulsando un trato igualitario de oportunidades.

Ejemplo de esto es la “Declaración de Montreal: Derechos Humanos LGBT” en el 2006, donde se reconoce el mejoramiento en la aceptación de diferentes aspectos como la raza o el origen étnico, y a la vez, resalta la necesidad de aceptar y respetar otros aspectos de la diversidad humana como la orientación sexual o la identidad de género. Los principios de Yogyakarta en 2007, la resolución general de la asamblea General de la OEA en 2008, o la Convención Interamericana contra toda forma de Discriminación e Intolerancia en 2013, abordan la necesidad de desarrollar un marco normativo que impulse el cumplimiento igualitario de las personas sin importar su orientación sexual o identidad de género, bajo un ambiente social y jurídico que respete dichas particularidades de los individuos, así como resaltar la preocupación en la prevalencia de actos de violencia impulsados por un sentido de odio hacia las preferencias sexuales o identidades de género no normativas (COPRED, 2018) (véase Anexo 1).

El desarrollo de este tipo de acuerdos internacionales ha sido posible en gran medida a la lucha social de la población LGBT, que ha buscado a través de ser visibilizados y reconocidos, la igualdad en el ejercicio de derechos de una colectividad que históricamente ha sido discriminada (Indesol, 2018).

1.3.1 El gobierno mexicano y los derechos LGBT

Aun cuando no se menciona de manera explícita la igualdad de la que goza la comunidad LGBT a las garantías otorgadas por la Constitución Política de los Estados Unidos Mexicanos, el reconocimiento de tal acceso a estas garantías sin distinción, resalta la prohibición de un trato desigual hacia esta comunidad (Constitución Política de los Estados Unidos Mexicanos, 2022).

No obstante, el artículo 149 del Código Penal Federal, estipula el delito y sanciones correspondientes por la negación a bienes y servicios educativos, de salud o de trabajo, entre otros, por motivos de orientación sexual. La ley Federal para Prevenir y Eliminar la Discriminación en 2014, el Protocolo de actuación para quienes imparten justicia en casos que involucren la orientación sexual o la identidad de género (2014), o el protocolo de actuación para quienes imparten justicia para la atención de personas LGBTTI, son solo algunos ejemplos del marco normativo nacional que buscan mejorar la sensibilización de las autoridades en temas relacionados a la diversidad sexo-genérica, impulsando un mejor desarrollo integral de las personas LGBT (véase Anexo 2) (COPRED, 2018).

En México, el movimiento LGBT, ha logrado colocar en la agenda pública la necesidad de impulsar acciones y estrategias que funjan como una base sustancial en la lucha contra la desigualdad y discriminación hacia este grupo poblacional (Indesol, 2018).

Dicho movimiento ha sido impulsado por acciones como la primera marcha lésbico homosexual en 1979, el primer Foro de Diversidad Sexual y Derechos Humanos en la Asamblea Legislativa del Distrito Federal en 1983, o la aprobación de la Ley de Sociedad de Convivencia, publicada en la Gaceta Oficial del Distrito Federal en el 2006 (COPRED, 2018).

Sin embargo, como reconoce la Comisión Internacional de Derechos Humanos (2018), los altos índices de violencia que se registran contra personas LGBT, además de una escasa respuesta estatal efectiva, resaltan la preocupación y necesidad de analizar y desarrollar estrategias efectivas que mejoren las condiciones de vida de este grupo poblacional. De igual forma reconoce que las motivaciones que llevan a la ejecución de tales actos de violencia, es compleja, multicausal, y que no se puede hablar de situaciones aisladas.

Dentro de estas acciones es indispensable resaltar la necesidad de desarrollar mecanismos de recolección de datos e información, pues de nueva cuenta, como resultado de la segregación sistemática, el estudio y análisis de la comunidad LGBT se obstaculiza al no contar con datos confiables invisibilizando y vulnerando a dicho grupo poblacional (Comisión Internacional de Derechos Humanos, 2018).

El acceso y ejercicio de los derechos garantizados en la Constitución Política o el marco jurídico nacional, no representa una garantía al respeto de dichos derechos en cualquier parte del país, pues como reconoce Martínez (2021), tanto los derechos como la tolerancia y aceptación de la población perteneciente a la comunidad LGBT, puede variar en diferentes regiones o estados del país, ya sea por costumbres y prácticas arraigadas o por la misma legislación local.

En este sentido es posible examinar que, aun cuando se han alcanzado ciertos logros para el respeto, reconocimiento y trato digno e igualitario de las personas LGBT como:

- El matrimonio igualitario
- La adopción homoparental
- El reconocimiento de identidad de género
- La prohibición de terapias de conversión
- La protección ante el discurso de odio

El goce de estas libertades y derechos puede no ser homogéneo en el país. Para el caso específico de interés en la presente investigación, es posible resaltar que, en materia de protección ante el discurso de odio, aún hay grandes brechas con las que hay que trabajar para impulsar el complimiento de dicha protección, donde la utilización de fuentes de datos alternativas es sustancial en tal labor.

1.4 La importancia de las fuentes de datos no estructuradas

Aún con el crecimiento exponencial en la generación y recolección de datos, no sólo de la información per se, sino también del carácter espacial y temporal de los mismos, cualquier investigación que haga uso o busque crear bases de datos para estudiar cualquier fenómeno o hecho, se enfrenta a la dificultad de obtener información concreta, completa y que sea significativa para la investigación (Escolano Utrilla, 2017).

Este crecimiento exponencial se ha beneficiado en gran medida al incremento en la accesibilidad a ciertas tecnologías como computadoras o teléfonos celulares, así como de ciertos canales de comunicación como las redes sociales. Para el año 2020 se estimó que la cantidad de datos digitales alcanzó los 64 zetabytes, y se prevé que para el 2025, esta cantidad supere los 180 zetabytes, lo cual representa un crecimiento medio anual de casi el 40% en cinco años, crecimiento que refleja el reto concebido en cuanto a la gestión y manejo de los datos (Mena Roa, 2021).

Sin embargo, la mayoría de los datos digitales suelen carecer de un formato estándar o estructura convencional, lo cual impide su procesamiento a través de modelos relacionales, es decir, que esta información se encuentra en un formato no estructurado (Eberendu, 2016).

A diferencia de los datos no estructurados, los datos estructurados se organizan formalmente tomando como referencia estructuras bien definidas y en lo general se almacenan en bases de datos relacionales en donde es posible recuperar y consultar la información de manera eficiente por medio de un lenguaje de consulta estructurado (SQL); este tipo de información suele organizarse de manera ordenada a través de filas y columnas, utilizando un formato de etiquetado o títulos que facilitan el ordenamiento y acceso a los datos (Sirisha y Babu Reddy, 2017).

Por su parte, en los datos no estructurados, es mayor la dificultad de manejo y organización de datos; esto responde a dos importantes factores, inicialmente debido a que estas bases de datos pueden contener información en diversos formatos, ya sean imágenes, videos, audios o texto, y, por otra parte, la ausencia de una estructura de modelo o esquema de datos fijos, además que esta información no puede ser ajustada en filas y columnas (Sirisha y Babu Reddy, 2017).

En otras palabras, los datos no estructurados son un conjunto de información que carecen de organización, por lo cual, es necesario realizar un preprocesamiento con ayuda de aprendizaje computacional o modelos estadísticos a través de herramientas como la minería de datos o el procesamiento de lenguaje natural (NLP, por sus siglas en inglés) para poder analizar esta información (María Nancy y Maheswari, 2020).

Para el almacenamiento de este tipo de datos, es necesaria la utilización de bases de datos no relacionales o NOSQL ya sea a través de estructuras basadas en gráficos, jerárquicas u orientadas a objetos; este tipo de bases de datos, permite además del almacenamiento de información no estructurada, el manejo de grandes volúmenes de datos, generando un mejor rendimiento en el almacenamiento de la información (Sirisha y Babu Reddy, 2017).

La utilización de bases de datos NOSQL surge precisamente como una solución a la necesidad de trabajar con grandes cantidades de datos que en muchas ocasiones carecen de organización, para los cuales, los sistemas tradicionales son insuficientes (Blanco Rojas, Archila Córdoba y Ballesteros-Ricaurte, 2016).

Aunado a ello, es necesario mencionar que las bases de datos no estructuradas hacen uso de una serie de técnicas como la discretización, la reducción y la cuantificación como un esfuerzo de convertir los datos no estructurados en datos estructurados (Jiang et al., 2022).

De tal forma, hacer uso de las bases de datos no estructuradas, aún con los desafíos inherentes a esta, como la calidad, la actualización o el manejo de grandes volúmenes de datos, permite en primera instancia hacer uso de diferentes fuentes y formatos de información que sean de interés, tales como plataformas de redes sociales, y en segunda instancia, acercarse de mejor manera al comportamiento de los individuos o tema objetivo.

1.4.1 Redes sociales como fuente de información

Como se ha mencionado con anterioridad la actual generación de grandes volúmenes de información ha sido en gran medida gracias a los medios masivos de comunicación; dentro de ellos las redes sociales, reconocidas como aquellas comunidades formadas por usuarios y organizaciones que se relacionan entre sí a través de contenido visual, auditivo y/o textual dentro de una plataforma en línea, las cuales sobresalen como una fuente en la que es posible compartir información de manera rápida, así como de interactuar con otros usuarios por medio de opiniones y comentarios, permitiendo la interacción entre personas desde diferentes ubicaciones espaciales (Peiró, 2023).

Como reconoce De la Cruz (2019), estos medios han ido evolucionando con el paso del tiempo, permitiendo el intercambio de información, comentarios y opiniones en temas diversos, así como un medio de fácil acceso a información relevante en tiempo real para los usuarios de diversas redes sociales como Facebook, Twitter o Instagram. Además, se reconoce la posibilidad de estas plataformas de influir en los usuarios de manera más personal, no sólo al reconocerlos como medios de comunicación, también como plataformas de constante y continua interacción de una gran cantidad de usuarios.

La importancia de las redes sociales como una fuente de información radica en que la mayoría de la información existente suele valerse de herramientas como encuestas, las cuales pueden concebir un sesgo debido a la deseabilidad social, impulsando a los encuestados a responder con cierto sesgo motivados y motivadas por encajar en los estándares aceptados socialmente (Arcila, Blanco-Herrero y Valdez, 2020).

De esta forma, la utilización de fuentes no estructuradas como las redes sociales, han surgido como una importante fuente de información, pues la facilidad de compartir ideas, pensamientos, opiniones, enlaces, noticias o hechos relevantes, en muchas ocasiones desde el anonimato, impulsa a las y los usuarios a expresarse fuera de la deseabilidad social, y a la vez, identificar y visibilizar opiniones dominantes (Arcila, Blanco-Herrero y Valdez, 2020).

Este tipo de información reconoce una ventaja sobre algunos métodos de recolección de datos como las encuestas, entrevistas o censos, sobre todo en la actualización constante, inmediatez, así como la gran cantidad de datos que se generan diariamente, además de las ventajas que trae consigo la recolección automática de contenido (Martí, Serrano-Estrada y Nolasco-Cirugeda, 2019).

No obstante, al utilizar este tipo de información es importante considerar ciertas cuestiones esenciales y desventajas tanto en el manejo como en la recuperación o representatividad de los datos. Algunas de estas consideraciones se muestran en la Tabla 1.2.

Usuarios en redes sociales	El uso de las redes sociales varía entre rangos poblacionales, reconociendo que los adultos jóvenes de entre 18 y 29 años son los usuarios que utilizan mayormente de este tipo de plataformas.
Actividad	Al estudiar y analizar un fenómeno en específico es posible encontrar un sesgo en ciertos usuarios que son más activos respecto a otros, representando un sesgo al considerar una población en general.
Popularidad de la red social utilizada	Es necesario considerar la popularidad de la o las redes sociales utilizadas para la recolección de información, pues la popularidad de estas no es la misma en cualquier parte del mundo. Redes sociales como Twitter o Instagram son más populares en el continente americano y europeo.
Acceso a la información	La recolección de la información de ciertas redes sociales puede ser privada, lo cual imposibilita la utilización de esta información.
Tipo de contenido	El contenido que prevalece en las redes sociales puede ser diferente entre plataformas, ejemplo de esto son aquellas donde prevalece el contenido visual como Instagram o las plataformas donde el contenido consiste principalmente de texto como Twitter.
Velocidad de acumulación	Al hacer uso de las redes sociales como fuente de información es imperativo considerar que constantemente se está generando información y la acumulación de datos puede ser demasiado grande si no se limita una temporalidad de recolección.
Recolección	La utilización de este tipo de información puede efectuarse de manera manual, pero implica la utilización de una gran cantidad de personas que analicen y filtren la información, así como un mayor tiempo de búsqueda. La recolección automatizada requiere un conocimiento técnico previo para efectuar la recolección de manera adecuada.
Veracidad	Al considerar que la mayoría de la información generada en redes sociales se efectúa de manera personal e individual, es necesario considerar la veracidad de la información. Al realizar estudios de sentimientos o prevalencia de opiniones, la veracidad puede no ser un factor de suma relevancia, ya que no se analiza la información publicada como fuente, sino que se analiza la opinión, el carácter u opinión de los usuarios.

Idioma	Al ser accesibles en diferentes territorios del mundo, la diversidad de idiomas encontrados puede ser amplia, por lo que es importante delimitar zonas de estudio o idiomas de estudio. En redes sociales, no suele usarse un lenguaje formal, por lo que es fácil encontrar expresiones coloquiales, abreviaciones, o faltas de ortografía, factores que deben ser considerados al utilizar procesos de automatización. Algunas redes sociales hacen uso de caracteres especiales para categorizar temas o palabras claves, tal es el caso del hashtag en Twitter. De igual manera es importante considerar que una misma publicación puede contener texto escrito en diferentes idiomas.

Tabla 1.2 Consideraciones al utilizar redes sociales como fuente de información. Elaboración propia con información de Toivonen et at., (2019)

Al mismo tiempo, es imperativo considerar la posibilidad de recolectar información dotada de un carácter espacial, en donde a través de información específica como las coordenadas, algunas plataformas como Twitter, facilitan la recolección de información que permitan conocer la ubicación donde fue publicado un texto en específico.

1.4.2 El discurso de odio en redes sociales y su detección automática

En el análisis de la prevalencia de discurso de odio en redes sociales y la caracterización de la información publicada dentro de plataformas como Twitter, el estudio de textos permite discernir el sentido de las publicaciones, posibilitando una diferenciación entre la clasificación de textos que constituyen un discurso de odio y los que no.

Reconociendo que el alcance de las redes sociales y el internet es cada vez mayor, en conjunto a las dificultades de monitorear en tiempo real todo el contenido publicado y replicado, estas plataformas se han convertido en una poderosa herramienta comunicativa que no sólo favorece a la difusión de información relevante, sino también, a la difusión de contenido extremista y de odio. Por este motivo y por la naturaleza misma de las redes sociales, no es de extrañar que son precisamente estas, en donde es posible encontrar mayor cantidad de discurso de odio. Además, este tipo de comunicación puede localizarse no solo en el cuerpo del texto, sino también en las réplicas o comentarios y reacciones de las mismas (Jacks y Adler, 2015).

Lo cierto es que, aun cuando la mayoría de plataformas de redes sociales han desarrollado e implementado políticas de usuarios cada vez más estrictas que prohíben el discurso de odio, el complimiento de las mismas es en suma complejo, no solo por la gran cantidad de datos publicados diariamente, también por la complejidad de identificar toda publicación que expresa discurso de odio, especialmente en todas aquellas que son expresadas de formas sutiles (MacAvaney et al., 2019).

Una de las principales preocupaciones ante la abundancia y prevalencia de este tipo de discursos en las redes sociales, se relaciona estrechamente a la influencia que puede tener en la efectuación de delitos de

odio. Ejemplo de ello, puede observarse en la identificada conexión existente entre ciertos actos terroristas y los perfiles en redes sociales de los sospechosos, en los que se han identificado un extenso historial de publicaciones relacionadas con el discurso de odio. Este hecho subraya la importancia de abordar la presencia de discurso de odio en redes sociales, incluso en plataformas que implementan políticas estrictas sobre el uso y distribución de contenido (MacAvaney et al., 2019).

Sin embargo, al retomar la definición sobre discurso de odio descrita en el capítulo anterior, es necesario resaltar que, aunque uno de los puntos más preocupantes con la existencia del mismo es la incitación y promoción de la violencia o al odio, y su posible relación con la ocurrencia de crímenes de odio, así como las afectaciones nocivas en las personas o grupos de personas que son objeto de este tipo de discurso, también es posible identificar percepciones sociales en torno a ciertos temas. Esto puede impulsar la identificación de poblaciones vulnerables y a la vez, desarrollar herramientas para contrarrestar dicha problemática (Arcila, Blanco-Herrero y Valdez, 2020).

Para el estudio del discurso de odio en redes sociales, se ha visto en el aprendizaje automático una forma eficiente de clasificar la información obtenida a través de la implementación de algoritmos de clasificación que permitan identificar el carácter de la información. Como este tipo de información suele encontrarse en forma textual, algunas redes sociales como Twitter resaltan como plataformas de importancia para realizar dicho estudio. En este sentido, los modelos de aprendizaje automático permiten clasificar la información textual reconociendo el discurso de odio a través de etiquetas prestablecidas facilitando la detección y clasificación de información de interés (MacAvaney et al., 2019).

1.5 Aprendizaje automático y procesamiento de lenguaje natural

El aprendizaje automático es reconocido como una rama de la Inteligencia Artificial (IA), en donde el objetivo principal es proporcionar a las computadoras la capacidad de aprender sin la necesidad de ser programadas directamente; esta rama de la IA hace uso de diversas técnicas computacionales que permiten a las máquinas aprender por si solas, a través de algoritmos específicos, los cuales deben ser alimentados en la mayoría de los casos con grandes volúmenes de datos (Bi et al., 2019).

Una de las principales funcionalidades del aprendizaje automático está enfocada en la clasificación de información, identificando distribuciones desconocidas a través de la utilización de un conjunto de datos observados. Se puede reconocer como una técnica que posibilita la detección de patrones en grandes volúmenes de datos (Alain y Bengio, 2014).

En este sentido, al identificar el crecimiento del Big Data como la gran cantidad de datos digitales que se generan continuamente por la población mundial (Clausen, 2017). El aprendizaje automático ha sobresalido como una opción de suma importancia al analizar grandes volúmenes de información, ante los cuales, ciertos métodos estadísticos pueden no ser lo suficientemente eficientes (Bi et al., 2019).

En la implementación del aprendizaje automático, se reconocen dos grandes vertientes de trabajo: 1) El aprendizaje no supervisado, donde el algoritmo trata de identificar relaciones y agrupaciones de los datos a través de la tipificación de características individuales, sin referencias o información externa a los datos utilizados, y 2) El aprendizaje supervisado, en donde a diferencia del no supervisado, es necesario la generación de un grupo de información previamente etiquetada que funcionará como un modelo utilizado para clasificar información fuera de este grupo etiquetado. En este tipo de aprendizaje, se entrena al algoritmo por medio de características y etiquetas permitiendo la predicción de información externa tomando como referencia este procesamiento. Dichas clasificaciones generadas pueden ser binarias: (si, no), (0,1), (Rojo, Azul) o de múltiples clases: (si, no, tal vez), (0,1,2), (Rojo, Azul, Amarillo) (Bi et al., 2019) (Sandoval, 2018).

En otras palabras, dentro del aprendizaje supervisado, se obtienen relaciones entre los elementos del conjunto de datos caracterizados a través de etiquetas o clases establecidas, para posteriormente utilizar este modelo como referencia para otro conjunto de datos no etiquetado y en la mayoría de los casos de mayor volumen, generando predicciones con base en el modelo de entrenamiento (Riquelme Silva, 2019).

Este proceso de aprendizaje computacional, se divide en cuatro grandes etapas: procesamiento, entrenamiento, validación-prueba y resultados. En este proceso, a través de las características representadas por medio de vectores, es posible obtener características distintivas de los datos, permitiendo el aprendizaje del modelo (véase Figura 1.6).

Figura 1.6 Aprendizaje computacional supervisado. Elaboración propia con información de González Z. (2021)

Posteriormente a la fase de preprocesamiento donde se desarrolla el etiquetado de los datos y la selección del modelo, en la fase de entrenamiento se construye el o los modelos implementados con parámetros específicos, donde clasificadores como Máquinas de Soporte Vectorial (SVM, por sus siglas en inglés) o Regresión Logística (LR, por sus siglas en inglés), se caracterizan por una arquitectura simple en comparación a modelos más complejos como aquellos basados en aprendizaje profundo (Deep Learning), y redes neuronales, en los cuales se busca extraer características útiles de una representación de características de entrada simple (Zhang, Robinson y Tepper, 2018).

Para la fase de entrenamiento, el corpus de datos etiquetado debe ser dividido en porciones específicas para la fase de entrenamiento y validación, estas porciones suelen ser 70/30 u 80/20, utilizando la porción más grande para el entrenamiento del modelo, con la cual será entrenado y aprenderá características relevantes de los datos de entrada. El resto de los datos son utilizados para clasificar un conjunto de datos que el modelo no conoce pero que cuentan con etiqueta previa; este proceso permite obtener las métricas necesarias que serán implementadas en la siguiente fase de validación-prueba, donde se calculan métricas

de evaluación como el Accuracy o el F1-Score con la finalidad de corroborar el buen funcionamiento y clasificación adecuada del modelo.

Sin embargo, este ciclo del aprendizaje supervisado, no es un ejercicio lineal, ya que, al obtener los datos de rendimiento en la fase de validación, es necesario regresar de manera recurrente a la construcción del modelo, permitiendo la modificación de parámetros con la finalidad de obtener los mejore resultados posibles, seleccionando así el modelo con mejor rendimiento.

No obstante, para la implementación de cualquier modelo de aprendizaje computacional, es necesaria la utilización de técnicas como el Procesamiento de Lenguaje Natural (NLP, por sus siglas en inglés), con el cual, es posible analizar información subjetiva de un texto relacionado a un tema en específico, además de poder ejecutar procesos de estandarización y preprocesamiento, facilitando el manejo de estas bases de datos no estructuradas (Riquelme Silva, 2019).

En este sentido, es indispensable reconocer que el Procesamiento de Lenguaje Natural, estudia el proceso computacional de todos los lenguajes naturales, es decir que, el NLP permite generar una comunicación entre los dispositivos computacionales y los humanos a través de la manipulación e interpretación de distintos idiomas, intentando que las computadoras entiendan el lenguaje humano. Este tipo de procesamientos pueden ser utilizados para analizar contenido textual en plataformas de redes sociales como Twitter (Toivonen et al., 2019).

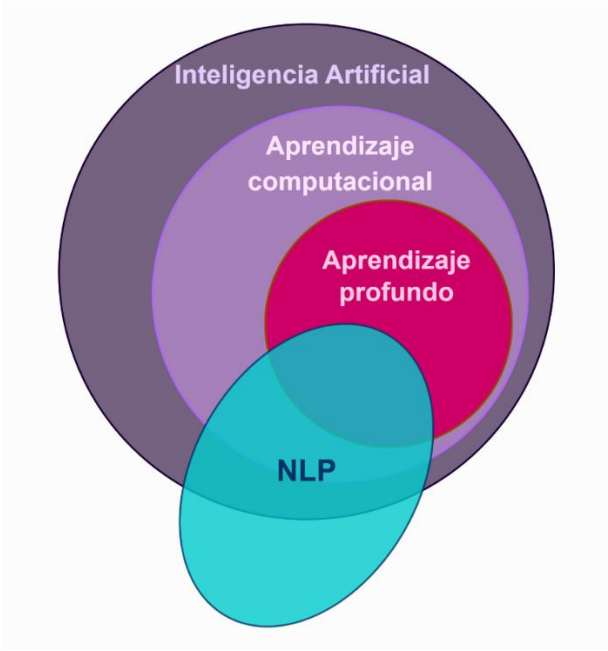


Figura 1.7 Inteligencia artificial, aprendizaje computacional, aprendizaje profundo y procesamiento de lenguaje natural. Elaboración propia

Como se observa en la Figura 1.7, tanto el aprendizaje profundo como el aprendizaje computacional y la misma inteligencia artificial, encuentran de manera adscrita la utilización del NLP como una técnica esencial especialmente en el manejo de información textual, lo cual permite realizar procesamientos necesarios para llegar a la fase de entrenamiento con un conjunto de datos que el modelo pueda procesar de una manera adecuada.

No obstante, para poder implementar este tipo de tareas y modelos, es necesario hacer uso de técnicas y métodos estadísticos, donde con ayuda del NLP, se realiza por ejemplo una división de caracteres en formas simples, ya sean, palabras, signos o números; esta división es realizada como un medio para analizar una cadena de entrada permitiendo la identificación de determinados caracteres o palabras de interés, proceso que es conocido como tokenización (Kibble, 2013).

Esta técnica lingüística es acompañada de muchas otras, tales como la lematización, en donde se busca reducir las palabras a su forma base o raíz, obteniendo lemas. Este proceso es desarrollado con la finalidad de reducir la variabilidad morfológica de las palabras que componen el corpus utilizado, y mejorando la eficiencia de tareas como la clasificación de textos o el análisis de sentimientos. Esta técnica es reconocida por su eficiencia al considerar estructuras gramaticales, en comparación a otras como el Stemming, en donde las palabras únicamente se truncan de manera simple (Miranda-Jiménez S., et al., 2017).

En esta fase de preprocesamiento, es posible implementar algunas técnicas y procesos con la finalidad de estandarizar los datos de entrada; procesos como la transformación del texto a minúsculas, remoción de acentos, caracteres especiales y números, eliminación de menciones como el caso del "#" en Twitter, la sustitución de emojis por textos estándar o la eliminación de palabras vacías o stop-word, entendidas como todas aquellas palabras que solo tienen un carácter de sintaxis gramatical como las preposiciones o los artículos. Este proceso de estandarización permite llegar a la fase de entrenamiento con un corpus adecuado, impulsando el buen aprendizaje de los modelos utilizados (Riquelme Silva, 2019).

El Procesamiento de Lenguaje Natural puede ser utilizado para generar: categorización de contenido, extracción contextual, análisis de sentimientos, conversión de habla a texto o de texto a habla, así como traducciones basadas en máquina, entre otros (Riquelme Silva, 2019).

Por tal motivo, es notorio el gran potencial del Procesamiento de Lenguaje Natural, no obstante, debe ser utilizado considerando ciertos factores que pueden influir fuertemente en la obtención de resultados. Las cuestiones enunciadas anteriormente han sido reconocidas como algunas de las más importantes al utilizar NLP, pero puede haber algunas otras cuestiones como la utilización de palabras coloquiales o jerga popular que puedan cambiar la connotación o sentido del texto, en donde con la finalidad de reducir el impacto de estas consideraciones, se pueden implementar algunas técnicas como la lematización.

1.5.1 Clasificadores y evaluadores

Como se ha mencionado anteriormente, el aprendizaje automático hace uso de ciertos modelos, los cuales pueden ser clasificados en tres grandes categorías: 1) Modelos lineales, en los que se busca ajustar los valores con los que se esté trabajando en una relación lineal; este tipo de modelos suelen ser reconocidos por arquitecturas simples, en donde resaltan algunos modelos como la regresión logística o la regresión de mínimos cuadrados, 2) Modelos de árbol, en los que los datos se dividen en ramas a través de la construcción de reglas de decisión y con los cuales es posible representar relaciones no lineales, como ejemplo de estos modelos se reconocen los árboles de decisiones o random forest, y 3) Redes neuronales, que son reconocidas como un análogo al funcionamiento neuronal del cerebro humano, donde se implementan un conjunto de neuronas artificiales interconectadas en las que ocurre un intercambio de

información con el objetivo de obtener valores de salida. La popularidad de este tipo de modelos ha crecido notablemente gracias a su capacidad predictiva en tareas complejas, sin embargo, suelen requerir mayor capacidad computacional y un tamaño de corpus más elevado (Sandoval, 2018).

Una de las implementaciones más populares en los modelos antes mencionados es la clasificación de información en categorías predefinidas, basándose en características representativas de los datos con los que se entrena el modelo, lo que implica la necesidad de contar con una muestra de datos etiquetados. Modelos como LR o SVM son ampliamente utilizados en tareas como la clasificación binaria, donde se busca predecir si un dato de entrada pertenece a una de dos categorías posibles. Aun cuando este tipo de modelos pueden ser implementados en tareas de clasificación múltiple, suelen mostrar mejor rendimiento en tareas más sencillas como la clasificación binaria. Modelos basados en redes neuronales como las Redes Neuronales Convolucionales (CNN, por sus siglas en inglés) o las Redes neuronales recurrentes (RNN, por sus siglas en inglés), implementan arquitecturas más complejas, con múltiples capas de neuronas, y suelen ser utilizadas en tareas más complejas como la clasificación de multi clase (Bi at. El, 2019) (véase Tabla 1.3).

Las Redes Neuronales Artificiales, se conforman por un conjunto de neuronas interconectadas a través de vías de señalización complejas, a través de axones, buscando obtener resultados por medio del análisis de interacciones entre un grupo de covarianzas medibles (Bi at. El, 2019). En este tipo de modelos, para el procesamiento de textos, suelen implementarse herramientas como las representaciones de palabras distribuidas, popularmente conocidas como incrustaciones de palabras o word embeddings, en donde cada palabra del corpus de textos utilizados, es inducido a una representación vectorial con el objetivo de capturar relaciones semánticas entre palabras, permitiendo que palabras diferentes, pero con semánticas similares, puedan concebirse con significados similares. La utilización de las incrustaciones de palabras, suelen generarse mediante técnicas computacionales como Word2Vec reconocido como un algoritmo utilizado para aprender incrustaciones de palabras, por medio de la estimación de la posición de cada palabra en un espacio vectorial multidimensional tomando como referencia la similitud de entre tokens (Schmidt, A. y Wiegand, M., 2017), (Ganfure G., 2022).

Además, suelen utilizarse funciones de activación, tales como la función sigmoide, relu o tah, las cuales buscan distorsionar los valores de salida de las neuronas por medio de deformaciones no lineales, facilitando la modelación de relaciones complejas. De igual manera, este tipo de modelos pueden hacer uso de optimizadores como el denominado adam, el cual combina las ventajas de otros algoritmos basados en el descenso rápido adaptativo y tiene la finalidad de ajustar las tasas de aprendizaje en modelos basados en redes neuronales (Rojano Aguilar, 2021).

Modelo	Funcionamiento
LR	Utiliza una función logística para estimar la probabilidad de pertenencia a una categoría dada.
SVM	Construyen un límite optimo llamada hiperplano que separa de mejor manera las observaciones de las categorías dadas y maximiza el margen entre estas.
CNN	Utiliza una red neuronal con capas convolucionales interconectadas que a través de filtros buscan extraer características representativas. También suelen implementar capas de activación para trabajar con la no linealidad, además de una capa de agrupamiento o pooling para reducir la dimensionalidad, para finalmente agrupar las características y poder realizar la clasificación.
RNN	Son similares a las CNN, sin embargo, administran un dominio temporal, en donde es posible formar ciclos que fluyan en ambas direcciones, considerando la secuencia de la información.

Tabla 1.3 Modelos de clasificación. Elaboración propia con información de Anezi (2022) y Bi et at., (2019).

En la implementación de los modelos antes mencionados o cualquier otro, es necesario analizar ciertos indicadores o medidas que permiten determinar la eficiencia del modelo a la hora de clasificar la información, facilitando la comparación del rendimiento entre modelos. Dentro de estas métricas, la matriz de confusión es un paso fundamental, ya que, a través de un desglose detallado de las clasificaciones obtenidas, es posible identificar la cantidad de datos clasificados de manera correcta en cada categoría determinada, facilitando la identificación de falsos positivos y falsos negativos (véase Tabla 1.4 y Figura 1.8).

Verdaderos Positivos (VP)	Datos que han sido catalogados como pertenecientes a la clase “A” de manera correcta.
Falsos Positivos (FP)	Datos que han sido catalogados a la clase “A” de manera incorrecta ya que no pertenecen a esta clase.
Verdaderos Negativos (VN)	Datos que no han sido catalogados como pertenecientes a la clase “A”, y realmente no pertenecen a esta clase.
Falsos negativos (FN)	Datos que no han sido catalogados en la clase “A”, pero si pertenecen a ella.

Tabla 1.4 Clasificación de información, matriz de confusión. Elaboración propia con información de (Riquelme Silva, 2019)

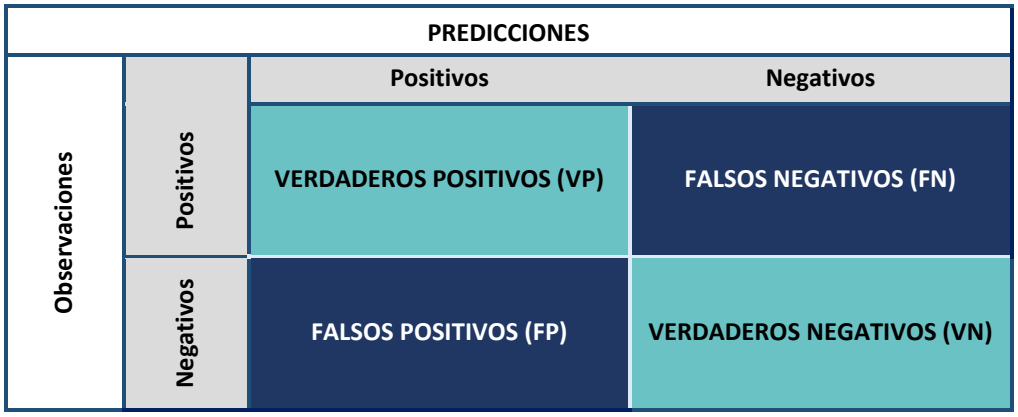


Figura 1.8 Matriz de confusión. Elaboración propia

Aunado a lo anterior, existen métricas estándar que son utilizadas para medir el rendimiento de los modelos según su clasificación de datos realizada. Algunas de estas métricas se desglosan de manera más detallada en la Tabla 1.5, se muestra la fórmula correspondiente en la Tabla 1.6.

Accuracy	Representa el valor obtenido de las predicciones correcta sobre el valor total de predicciones realizadas. Es decir, que muestra la porción de predicciones correctas (VP y VN) entre el total de la muestra.
Precisión	Mide la correspondencia entre la clasificación correcta de la información a cierta clase. Esta métrica hace referencia a la porción de verdaderos positivos sobre el total de predicciones positivas.
Recall	Es la relación entre los datos clasificados correctamente y la suma de todos los datos clasificados correctamente. Mide la porción de datos positivos clasificados correctamente.
F1-Score	Es definida como la media armónica de la precisión y el Recall. Es utilizada para comparar el rendimiento combinado de estas dos métricas.

Tabla 1.5 Métricas de evaluación. Elaboración propia con información de (Riquelme Silva, 2019) y González Z. (2021).

Accuracy	$\frac{VP + VN}{VP + VN + FP + FN}$
Precisión	$\frac{VP}{VP + FP}$
Recall	$\frac{VP}{VP + FN}$

F-Score	$2 \cdot \frac{(Precisión)\ (Recall)}{Precisión + Recall}$

Tabla 1.6 Formulas de métricas de evaluación. Elaboración propia con información de (González, 2021)

En este sentido es imperativo reconocer que la eficiencia de un sistema de aprendizaje automático en la tarea de clasificar elementos que no se encuentran dentro del corpus de la base de datos, depende directamente del algoritmo de clasificación elegido y utilizado, y a la vez, de las características que se han elegido para representar los elementos de ejemplo con los que se entrena el modelo, además del etiquetado de los datos, tarea que será fundamental para que el algoritmo pueda aprender de manera adecuada a través de los parámetros e información de entrada (Riquelme Silva, 2019).

2 Antecedentes, planteamiento del problema, justificación y objetivos

El impacto del Word Wide Web en la comunicación, ha significado una revolución en el intercambio de información; la Web 1.0, mostró una realidad donde tanto el acceso como la difusión de información era más sencilla, situación que se incrementó sustancialmente con la llegada de la Web 2.0, la cual, apertura ciberespacios de intercomunicación personal y social que se popularizaron de manera rápida abriendo oportunidades de relacionarse comunicativamente de una manera más personal incrementando la posibilidad de incidir en el receptor-emisor (Miró Linares, 2016).

Actualmente, el desarrollo de la Web 3.0 examina la importancia de identificar los potenciales efectos que traen consigo este tipo de comunicaciones, en donde, se reconoce al internet como un posible foro de radicalización, así como un medio en el que existen un conjunto de comunicaciones y conductas agresivas y ofensivas que prevalecen especialmente en redes sociales como Twitter.

Sin duda alguna, las redes sociales han revolucionado la comunicación de las personas, la forma en que comparten ideas, opiniones o situaciones cotidianas. Sin embargo, la prevalencia de las comunicaciones violentas y el discurso de odio ha sobresalido como una preocupación latente relacionada a los posibles efectos de este tipo de comunicación, por lo cual, los esfuerzos por detectar y contrarrestar este tipo de discursos ha sido una tarea reconocida por su importancia inmediata, impulsando a la exploración de técnicas que ayuden a esta tarea. (Alshalan y Al-Khalifa, 2020).

El discurso de odio en redes sociales ha sido abordado a través de técnicas y metodologías manuales con ayuda de expertos en la materia, no obstante, esto puede ser sumamente costoso monetaria y temporalmente, por lo cual, las técnicas de aprendizaje automático han sobresalido como herramientas de gran utilidad en el estudio del discurso de odio en medios masivos de comunicación y redes sociales.

Algunos trabajos como el de Clausen et al., (2017) muestran el impacto de generar monitoreos diarios sobre la actividad en redes sociales como Twitter, identificando tuits geolocalizados relacionados a tópicos específicos, permitiendo reconocer anomalías, tendencias, influencias y la perspectiva y opiniones predominantes sobre el tema de análisis. Aunado a ello, también se resalta la carencia de estadísticas en este caso de pruebas de VIH, lo cual resalta nuevamente la importancia de generar estadísticas y bases de datos que impulsen el análisis de ciertas problemáticas y temas en específico a través de datos.

No obstante, en los últimos años, el estudio del discurso de odio ha identificado en los enfoques basados en aprendizaje profundo una oportunidad eficiente y confiable de adentrarse a la identificación y detección del discurso de odio (Alshalan y Al-Khalifa, 2020). Por tal motivo, en este capítulo se abordarán algunos trabajos relevantes que hacen uso de técnicas de aprendizaje automático, ya sea a través de algoritmos de aprendizaje superficial o de aprendizaje profundo, con el objetivo de clasificar información recopilada y detectar el discurso de odio en general o en su caso enfocado a temas específicos.

2.1 Aprendizaje automático en la detección del discurso de odio

El crecimiento de las redes sociales como plataformas de comunicación masiva ha provocado un aumento significativo en la cantidad de información generada día con día. La utilización de plataformas populares como Twitter pueden representar un medio de importancia en la difusión y obtención de información que sea relevante, además de su funcionalidad lúdica. No obstante, la prevalencia del discurso de odio muestra una brecha en la utilización de estas plataformas, pues como se ha mencionado en el capítulo anterior, el impacto que puede traer consigo el discurso de odio a ciertos colectivos o personas puede ser significativamente negativo.

Sin embargo, el monitoreo y análisis de este tipo de comunicación puede ser en suma complicado. La gran cantidad de datos generados diaria y constantemente en redes sociales representa una tarea casi imposible de realizar de manera manual (Anezi, 2022).

Este monitoreo e identificación del discurso de odio, ha sido abordado desde otras perspectivas y técnicas como la elaboración de encuestas e informes, los cuales han hecho grandes esfuerzos por mostrar resultados como el aumento del discurso de odio y delitos de odio, por lo cual, se ha buscado tratar o eliminar contenido ofensivo que se encuentre en línea, así como el incremento de políticas de censura de contenido en diversas plataformas, sin embargo, este proceso también demanda un alto costo monetario, temporal y de mano de obra (Zhang, Robinson y Tepper, 2018).

Además, la utilización de técnicas como las encuestas, las entrevistas ya sean estructuradas o semiestructuradas, así como los grupos focales, permiten identificar factores claves desde la perspectiva y experiencia de la población objetivo, sin embargo, también suelen ser técnicas costosas y tardadas que abordan problemas limitados y obtienen datos a pequeña escala (Karami et al., 2021).

En este sentido, la utilización de técnicas de inteligencia artificial y de manera más específica del aprendizaje automático han sobresalido como herramientas que posibilitan la identificación de información determinada. La necesidad de desarrollar este tipo de metodologías y herramientas se relaciona directamente con la imposibilidad de detectar y prevenir el incremento del discurso de odio únicamente bajo supervisión humana (Baydoğan y Alatas, 2022).

Así, la utilización de técnicas de procesamiento de lenguaje natural y de aprendizaje automático, han ido en aumento en los últimos años, dirigiendo los esfuerzos de los y las investigadoras hacia enfoques centrados en la automatización, buscando proponer soluciones a la tarea de identificar discurso de odio, lenguaje ofensivo, acoso, racismo o misoginia, entre otros (Gómez-Espinosa, V., Muniz-Sanchez, V., y Pastor López-Monroy, A., 2021).

La mayoría de estos estudios enfocados en la detección de discurso de odio a través de técnicas y metodologías de aprendizaje automático, se han centrado principalmente en técnicas de clasificación supervisada, ya sea través de algoritmos más clásicos como las máquinas de soporte vectorial (VSM), y la regresión logística, o bien con técnicas del aprendizaje profundo como las redes neuronales convolucionales (CNN) o las redes neuronales recurrentes (RNN) (Alshalan y Al-Khalifa, 2020).

Como se ha mencionado anteriormente, los modelos de aprendizaje automático dependen directamente de la información con la que son alimentados. Reconociendo la naturaleza de los datos de Twitter como

fuentes de datos no estructuradas, el procesamiento de lenguaje natural es indispensable en el manejo y tratamiento de datos que permitan pasar de una base de datos obtenida con registros de Twitter a un conjunto de datos que sea procesable por el algoritmo de aprendizaje automático en cuestión.

Los estudios desarrollados para la detección de discurso de odio han mostrado eficiencia con la utilización de características lingüísticas simples como las bolsas de palabras o los n-gramas; no obstante, actualmente una de las técnicas más utilizadas sobre todo en el desarrollo de metodologías del aprendizaje profundo, es la incrustación de palabras o Word embeddig, en donde, se busca aprender una representación vectorial de valor real para un vocabulario predefinido y de tamaño fijo, tomando como referencia el corpus de texto. El objetivo de estos modelos reconoce la existencia de cierta importancia semántica entre las palabras de un texto, la cual es reconocida a través de la identificación de probabilidades de co-ocurrencia de palabras (Ganfure G., 2022).

Tal como reconoce Arcila-Calderón et al., (2021) las metodologías basadas en detección automática utilizadas en investigaciones que buscan clasificar discurso de odio pueden dividirse en dos grandes categorías, los basados en aprendizaje superficial que utilizan métodos como las Máquinas de Soporte Vectorial, la Regresión Logística o el Random Forest, y, por otro lado, los métodos basados en aprendizaje profundo donde suelen implementarse técnicas basadas en redes neuronales o algunos métodos más complejos.

Los modelos basados en aprendizaje superficial han demostrado obtener buenos resultados en la detección del discurso de odio especialmente cuando el objetivo de la metodología es obtener una clasificación binaria, es decir, identificar aquellos textos que hacen referencia a discurso de odio diferenciándolos de aquellos que no lo hacen (véase Tabla 2.1).

No obstante, en la búsqueda de identificar y refinar modelos con la finalidad de obtener mejores resultados, se han desarrollado esfuerzos que implementan metodologías basadas en aprendizaje superficial, así como basadas en aprendizaje profundo, esto con la finalidad de identificar y utilizar aquel modelo que obtenga mejores resultados (véase Tabla 2.2).

Ejemplo de esto es el trabajo de Baydoğan y Alatas (2022), quienes han buscado clasificar el discurso de odio relacionado al COVID 19 a través de la utilización de técnicas de aprendizaje superficial y aprendizaje profundo, mediante la utilización de técnicas de procesamiento de lenguaje natural como bolsa de palabras, frecuencia de términos (TF-IDF) y matriz de documentos como herramientas que apoyen al entrenamiento del modelo. Para esto se utilizaron dos conjuntos de datos, uno con 2,319 tuits y 4 clases a clasificar y otro con 20 mil tuits para 5 clases (Baydoğan y Alatas, 2022).

Para realizar esta clasificación se utilizaron 10 metodologías diferentes, 5 basadas en aprendizaje superficial: Máquinas de Soporte Vectorial (SVM), k-Vecinos Cercanos (KNN), Naive Bayes (NB), Random Forest (RF) y Regresión Logística (LR); de los cuales, el algoritmo identificado con mejor rendimiento fue SVM para el primer conjunto de datos con un Accuracy de 0.5528 y F1-Score de 0.5010, seguido por LR con un Accuracy de 0.5308 y F1-Score de 0.4970, mientras que para el segundo conjunto de datos, LR mostró un mejor resultado con un Accuracy de 0.7499 y F1-Score de 0.7130 (Baydoğan y Alatas, 2022).

Por otro lado, se utilizaron los siguientes modelos basados en aprendizaje profundo: Redes Neuronales Artificiales (ANN), Redes Neuronales Convolucionales (CNN), Redes Neuronales Recurrentes (RNN). Long short-termMmemory (LSTM), y Unidades Recurrentes Cerradas (GRUs). De estos modelos, para el primer

conjunto de datos obtuvieron el mejor resultado con RNN al obtener un Accuracy igual a 0.7871 y F1-Score de 0.4511, seguido por CNN con un Accuracy de 0.7655 y F1-Score de 0.3079. Para el segundo conjunto de datos, nuevamente RNN obtuvo el mejor rendimiento con un valor de Accuracy de 0.9030 y F1-Score de 0.7348 (Baydoğan y Alatas, 2022).

Oriola y Kotze (2020) tomaron como objetivo la identificación de discurso de odio y lenguaje ofensivo en inglés y 3 idiomas africanos (afrikáans, IsiZulu y Sesotho) dividiendo la base de datos de 15,702 tuits en 3 muestras diferentes (inglés y afrikaans, inglés e IsiZulu e inglés y Sesotho), obteniendo 14,896 tuits después de eliminar aquellos registros que no concordaban con las anotaciones de los expertos. Los modelos utilizados fueron LR, SVM, RF, y Gradient Boostig (GB), no obstante, con el objetivo de optimizar los modelos de entrenamiento, midieron el valor de verdaderos positivos al utilizar n-gramas de palabras, n-gramas de caracteres, características basadas en la sintaxis y el sentimiento negativo. Con el fin de optimizar su rendimiento, los n-gramas se ponderaron por frecuencia de términos de documento inversa (TF-IDF) que compensan el número de palabras en un documento por la frecuencia de la palabra en un corpus. Como resultado, demostraron que el SVM optimizado con n-gramas de caracteres registró la mejor tasa de verdaderos positivos con un resultado de 0.894 y un Accuracy general de 0.646 para el discurso de odio.

De igual manera con un enfoque basado únicamente en aprendizaje superficial, Arcila Calderón, Blanco-Herrero, y Valdez Apolo (2020) utilizan Naive Bayes original, Naive Bayes para modelos multimodales, Naive Bayes para modelos multivariados Bernoulli, Regresión Logística, Regresión logística con gradiente descendente estocástico y SVM con estimador SVC-T, así como un modelo basado en las votaciones de cada uno de los modelos generados, esto con la finalidad de medir el rechazo en oposición a la neutralidad o aceptación a la migración, para lo cual, se utilizaron bases de datos ya existentes con las etiquetas correspondientes, en donde, se reconoció que todos los modelos empleados obtuvieron valores significativamente por encima de la línea base de 55% obteniendo el mejor resultado con el clasificador basado en la votación de los modelos con un Accuracy de 76.19 y F1-Score de 75.18, seguido del SVM con un Accuracy de 75.36 y un F1-Score de 78.32.

De esta forma, se reconoce la importancia de los algoritmos basados en aprendizaje superficial y el potencial que tienen como metodologías que proporcionan buenos resultados en la detección binaria de discurso de odio, en especial al generar una línea base en el discurso de odio enfocado en algún tema o grupo poblacional específico.

Actualmente, la utilización de técnicas de aprendizaje profundo o Deep Learning han visualizado un mayor auge en la clasificación e identificación de discurso de odio. A través de la implementación de redes neuronales, el aprendizaje profundo se caracteriza por modelos más sofisticados y complejos, buscando que el modelo permita aprender automáticamente características abstractas de los datos (Zhang, Robinson y Tepper, 2018).

Por este motivo, algunos trabajos como el de Alshalan y Al-Khalifa (2020), se enfocan en la implementación de modelos basados en redes neuronales, en donde, si bien también hacen uso de técnicas más tradicionales como el SVM, la investigación se centra en la aplicación de modelos basados en aprendizaje profundo. En este trabajo se han implementado: CNN, Unidades Recurrentes Cerradas (GRU por sus siglas en inglés), CNN + GRU y Representaciones de codificador bidireccional de transformes (BERT); sin embargo, también han utilizado SVM con n-gramas de caracteres y LR con n-gramas de características. El

objetivo de Alshalan y Al-Khalifa ha sido la identificación de discurso de odio en tuits redactados en árabe con una base de datos de 8,964 tuits etiquetados. En todos los modelos utilizados, las características se representaron como incrustaciones de palabras (Word embeding a través de Word2Vec). Para los modelos basados en redes neuronales, utilizaron Keras con TensorFlow como backend. En este trabajo, el modelo que obtiene mejores resultados es CNN con un Accuracy de 0.83 y F1-Score de 0.79 y un AUROC de 0.89, mientras que, para los modelos superficiales, LR obtuvo mejores resultados con un Accuracy de 0.79 y F1-Score de 0.75.

Anezi (2022) identificó discurso de odio en idioma árabe con un conjunto de tuits de 4,203 a través de 7 diferentes categorías: comentarios contra: religión, racista, igualdad de género, contenido violento, contenido ofensivo o contenido de insulto/intimidación, comentarios positivos normales y comentarios negativos. Para permitir al sistema aprender vectores de palabras de acuerdo a su co-ocurrencia, se utilizó Word2Vec y Vector Global (Glove) para la incrustación de palabras, de igual forma se utilizaron varios modelos de aprendizaje superficial: Árboles de Decisión (DT), multicapa perceptrón (MLP), Naive Bayes (NB) y Regresión Logística (LR).

Estos modelos anteriores, se evaluaron para la detección de discurso de odio en árabe en función de las categorías, reconociendo que, para una clasificación binaria, la Regresión Logística obtuvo el mejor resultado con un Accuracy de 63.937 y un F1-Score de 0.65, seguido de NB con un Accuracy de 62.810 y F1-Score de 0.59. Para la detección de tres clases diferentes (textos normales positivos, negativos o de discurso de odio), DT obtuvo el mejor resultado con un Accuracy de 56.192 y un F1-Score de 0.56. Por otra parte, este experimento se realizó nuevamente con estos métodos de aprendizaje superficial, pero para la detección de las 7 diferentes categorías, en donde, los resultados fueron bastante bajos, pues el método que obtuvo las mejores puntuaciones fue nuevamente DT con un Accuracy de 34.761 y un F1-Score de 0.34 (Anezi, 2022).

Para realizar esta detección desde el aprendizaje profundo, Anezi (2022), propone una arquitectura basada en RNNs llamada DRNN-2 las cual consiste en 10 capas con tamaños de lote 32 y 50 iteraciones para la tarea de clasificación, de igual forma propone otro modelo llamado DRNN-1 con 5 capas ocultas. Este último modelo se utilizó únicamente para la detección binaria del discurso de odio, el cual obtuvo un Accuracy en la validación del modelo de 83.2, mientras que el modelo DRNN-2 obtuvo un Accuracy de 90.30. Este mismo modelo fue utilizado para la clasificación del discurso de odio en 3 diferentes clases, en donde con la utilización de DRNN-2 se obtuvo un Accuracy de 86.40, mientras que, para la clasificación de las 7 clases diferentes, se reconoció un valor de Accuracy en la validación de 58.26; de las clasificaciones utilizadas en los modelos que no son binarios, se reportaron únicamente los resultados obtenidos con RNN-2 ya que mostraron un mejor rendimiento que los obtenidos con RNN-1.

Bilal, Khan, Jan y Musa (2022), también evalúan el rendimiento de diferentes técnicas de aprendizaje computacional más tradicionales y otras basadas en aprendizaje profundo para la detección de discurso de odio en urdo romano, sin embargo, los autores hacen grandes aportaciones en los rendimientos de los algoritmos al agregar un factor sensible al contexto basado en Bi-LSTM con una capa de atención y el uso de Word2Vec para la incrustación de palabras; de igual manera, se identificó que los resultados eran mejores tras generar una normalización léxica de las palabras en urdu romano.

Para la recolección de datos, se seleccionaron temas específicos como deportes, religión, política, eventos actuales o perspectivas de género y se utilizaron los siguientes modelos: LR, SVM, SVM lineal, XG Boost,

RF, DT, KNN, LSTM, Bi-LSTM, Bi-LSTM + Atención y CNN. De estos modelos RF obtuvo los mejores resultados con un Accuracy de 0.71, sin embargo, el enfoque basado en Bi-LSTM con capa de atención obtuvo valores mejores pues obtuvo un Accuracy de 0.875, en este sentido, es posible reconocer que aunque los resultados obtenidos con los modelos basados en aprendizaje profundo obtuvieron mejores resultados que los basados en aprendizaje superficial, estos últimos también obtuvieron buenos resultados, además que la consideración de la capa de atención sensible al contexto enriqueció el funcionamiento del modelo, especialmente al considerar la naturaleza de la léxica en el idioma utilizado (Bilal, Khan, Jan y Musa (2022).

La detección del discurso de odio puede centrarse únicamente en algoritmos basados en aprendizaje profundo, tal es el caso de Zhang, Robinson y Tepper (2018) quienes proponen la optimización de un modelo de red neuronal con capas de abandono, agrupación y regularización de red elástica. Zhang, Robinson y Tepper buscan mostrar la eficiencia del método planteado al evaluar las métricas de evaluación registradas para la detección de discurso de odio en bases de datos relacionadas a la religión y refugiados generadas por otros autores, obteniendo como resultado que el modelo supera en 6 de 7 conjuntos de datos utilizados con hasta en un 13% en F1-Score.

Por su parte, Ganfure G. (2022) busca identificar discurso de odio en Twitter y Facebook en idioma Afaan Oromo utilizando CNN, Redes de Memoria a Corto Plazo (LSTM), Redes Bidireccionales de Memoria a Largo Plazo (Bi-LSTM), GRU y CNNLSTM, en donde con CNN y BiLSTM logra el puntaje de F1-Score promedio ponderado más alto del 87%.

La utilización de técnicas de aprendizaje profundo y superficial suelen centrarse en la detección de discurso de odio en general, es decir, sin mantener un enfoque centrado en una población objetivo, o bien en ciertos grupos poblacionales como el caso de Gambäck y Kumar Sadar (2017), quienes buscan generar clasificaciones del discurso de odio entre racismo, sexismo, ambos o no discurso de odio.

No obstante, son pocos los esfuerzos enfocados en la población perteneciente a la comunidad LGBT. Aunque existen algunos trabajos como el de Arcila-Calderón et al., (2021), en donde se busca identificar discurso de odio en Twitter relacionados a la comunidad LGBTI+ o al género, en donde se utilizó la clasificación de Miró Linares para elaborar el etiquetado en una base de datos de 10,855 tuits. Se utilizaron los siguientes modelos Naïve Bayes original, Naïve Bayes para modelos multinomiales, Naïve Bayes para Modelos multivariados de Bernoulli, Regresión Logística, Clasificadores Lineales con Estocástico, Entrenamiento de descenso de gradiente y máquinas de soporte vectorial, de los cuales, el mejor modelo fue la Regresión Logística con un Accuracy de 0.76 y un F1 de 0.57, mientras que, del lado de los modelos de aprendizaje profundo, el mejor resultado obtenido corresponde a RNN con un Accuracy de 0.9030 y un F1-Score de 0.7348.

Fitri, Andreswari y Hasibuan (2019) buscan clasificar tuits relacionados a campañas anti LGBT mediante técnicas de aprendizaje computacional, obteniendo las mejores métricas de evaluación con Naive Bayes con un Accuracy de 86.43.

De igual manera, se reconoce el esfuerzo realizado por Karami et, al. (2021), en donde se busca hacer una clasificación de los perfiles LGBT de cuentas en Twitter, a través de un clasificador de aprendizaje automático basado en el perfil y las características biográficas de las cuentas utilizadas en la base de datos. Se utilizaron NaiveBayes, BayesNet, Random Forest, J48, y SVM, así como redes neuronales

convolucionales (CNN), obteniendo un resultado de 0.7651 para este último en el Accuracy, mientras que se obtuvo un 0.8791 para el Accuracy de BayesNet, identificado como el algoritmo con mejor rendimiento.

	Autor	Año	Tema central	Método utilizado	Lenguaje	Cantidad de datos	Parámetros de evaluación
Aprendizaje superficial	Baydoğan y Alatas	2021	COVID-19	SVM Bag of words, TF-IDF, document matrix)	Inglés	2,319	Accuracy=0.5528 Precission=0.5400 Sensitivity=0.5530 F1=0.5010
	Baydoğan y Alatas	2021	COVID-19	LR (Bag of words, TF-IDF, document matrix)	Inglés	20,000	Accuracy=0.7499 Precission=0.7030 Sensitivity=0.7500 F1=0.7130
	Arcila-Calderón et al.	2021	Comunidad lGBTI/género	LR (Bag of Words)	Español	10,855	Accuracy=0.76 F1=0.86 AUC-ROC=0.57
	Karami et al.	2021	Perfiles LGBT de cuentas en Twitter	Bayes Net	Inglés	16,241	Accuracy=0.8791 AUC=0.913
	Oriola y Kotze	2020	Discurso de odio en inglés y lenguas africanas	SVM n-gram	Inglés Afrikaans IsiZulu Sesotho	14,896 (dividido en 3 muestras)	Hate speech TP=0.894 Offensive speech TP=0.069 Free speech TP=0701 Accuracy=0.646 Precision=0.65 Recall=0.55 F1=0.31

Alshalan y Al-Khalifa	2020	Discurso de odio en idioma árabe	LR	Árabe	8,964	Precision=0.75 Recall=0.74 F1=0.75 Accuracy=0.79 Hate class Recall=0.63 AUROC=0.84
Fitri, Andreswari y Hasibuan	2019	Campañas anti lgbt	NB	No reportado (Indonesia)		Accuracy=86.43% Precision, Recall y F1 en un promedio de entre 56% y 65%
Anezi	2022	Discurso de odio: clasificación binaria	LR	Árabe	4,203	Accuracy=63.937% Precision=0.63 Recall=0.61 F1=0.65
Anezi	2022	Discurso de odio: Positivos, negativos o de discurso de odio	DT	Árabe	4,203	Accuracy=56.192% Precision=0.56 Recall=0.56 F1=0.56
Anezi	2022	Discurso de odio: religión, racista, igualdad de género, comentarios clasificados como insultos/bullying, Violentos/ofensivos, Positivos normales y negativos normales	DT	Árabe	4,203	Accuracy=34.761% Precision=0.35 Recall=0.35 F1=0.34
Bilal, Khan, Jan y Musa	2022	Discurso de odio	RF	Urdu romano	30,000	Accuracy=0.71 F1=0.77

							Precision=0.73 Recall=0.71
	Arcila Calderón, Blanco-Herrero, y Valdez Apolo	2020	Migración	SVC (estimador SVC)	Español	1,469	Accuracy=75.36 Precision=76.11 Recall=80.78 F1=78.32
	Arcila Calderón, Blanco-Herrero, y Valdez Apolo	2020	Migración	Clasificador basado en votación de 6 modelos	Español	1,469	Accuracy=76.19 Precision=73.19 Recall=77.30 F1=75.18

Tabla 2.1 Técnicas de aprendizaje superficial en trabajos similares. Elaboración propia con información de Alshalan y Al-Khalifa (2020), Anezi (2020), Arcila-Calderón et al., (2021), Arcila Calderón, Blanco-Herrero, y Valdez Apolo (2020), Bilal, Khan, Jan y Musa

	Autor	Año	Tema central	Método utilizado	Lenguaje	Cantidad de datos	Parámetros de evaluación
Aprendizaje profundo	Zhang, Robinson y Tepper	2018	Religión/ refugiados	CNN + GRU (Word2Vec)	Inglés	2,435	F1=0.92
	Baydoğan y Alatas	2021	COVID-19	RNN (Bag of words, TF-IDF, document matrix)	Inglés	2,319	Accuracy=0.7871 Precission=0.6038 Sensitivity=0.3718 F1=0.4511
	Baydoğan y Alatas	2021	COVID-19	RNN (Bag of words, TF-IDF, document matrix)	Inglés	20,000	Accuracy=0.9030 Precission=0.8085 Sensitivity=0.6760 F1=0.7348
	Arcila-Calderón et al.	2021	Comunidad lGBTI/género	RNN Word embeddings	Español	10,855	Accuracy=0.84 F1=0.65 AUC-ROC=0.85

	Karami et al.	2021	Perfiles LGBT de cuentas en Twitter	CNN	Inglés	16,241	Accuracy=0.7651 AUC=0.828
	Ganfure G.	2022	Discurso de odio	BiLSTM (n-gramas Word2Vec)	Afaan Oromo	35,200	Clasificación ponderada F1=91%
	Alshalan y Al-Khalifa	2020	Discurso de odio en idioma árabe	CNN Word2Vec Propio Word2Vec con (CBOW)	Árabe	8,964	Precision=0.81 Recall=0.78 F1=0.79 Accuracy=0.83 Hate class Recall=0.67 AUROC=0.89
	Gambäck y Kumar Sadar	2017	Incitación al odio Racismo Sexismo Ambos No discurso de odio	CNN 4-ngrams Word2Vec	Inglés	6,655	F1=78.3%
	Anezi	2022	Discurso de odio: clasificación binaria	RNN Propuesta: DRNN-1 (5 capas profundas y 231,841 parámetros) Word2Vec y GloVe	Árabe	4,203	Train loss=0.0044 Train Accuracy=99.73% Val. Loss=0.9439 Val.Accuracy=83.22%
	Anezi	2022	Discurso de odio: clasificación binaria	RNN Propuesta: DRNN-2 (10 capas profundas con 1,682,787 parámetros) Word2Vec y GloVe	Árabe	4,203	Train loss=0.0578 Train Accuracy=98.18% Val. Loss=0.3574 Val.Accuracy=90.30%
	Anezi	2022	Discurso de odio: religión, racista,	RNN Propuesta:	Árabe	4,203	Train loss=0.4752 Train Accuracy=84.14%

			igualdad de género, comentarios clasificados como insultos/bullying, Violentos/ofensivos, Positivos normales y negativos normales	DRNN-2 (10 capas profundas con 1,682,787 parámetros) Word2Vec y GloVe			Val. Loss=1.6282 Val.Accuracy=58.26%
	Bilal, Khan, Jan y Musa	2022	Discurso de odio	Bi-LSTM con capa de atención Word2Vec	Urdu romano	30,000	Precision=0.875 F1=0.8848 Precision=89.22 Recall=88.36
	Badjatiya et al.	2017	Discurso de odio: Racismo, sexismo o ninguno	LSTM + Random Embedding + GBDT (Gradient Boosted Decision Trees)	No reportado	16,000	Precision=0.930 Recall=0.930 F1=0.930

Tabla 2.2 Técnicas de aprendizaje profundo en trabajos similares. Elaboración propia con información de Alshalan y Al-Khalifa (2020), Anezi (2022), Arcila-Calderón et al., (2021), Badjatiya et al., (2017), Baydoğan y Alatas (2021), Bilal, Khan, Jan y Musa (2022)

Al comparar los resultados obtenidos de los trabajos desglosados en la Tabla 2.1 y la Tabla 2.2, es posible señalar que, inicialmente la utilización de los modelos implementados puede variar de acuerdo al interés de las y los investigadores, así como del objetivo de la investigación y las bases de datos o corpus que serán utilizados para entrenar al modelo. Además, se reconoce que la construcción de la arquitectura de los modelos puede ser tan compleja como el trabajo permita y el objetivo demande, sin embargo, resaltan algunos modelos por la eficiencia observada de manera recurrente en múltiples trabajos, tales como la Regresión Logística y las Máquinas de Soporte Vectorial del lado del aprendizaje superficial o las Redes Neuronales Convolucionales y las Redes Neuronales Recurrentes dentro del aprendizaje profundo.

Sin embargo, es necesario resaltar que el factor gramatical de los idiomas objetivo en las investigaciones funge como un pilar fundamental en el desarrollo metodológico, trabajos como el de Bilal, Khan, Jan y Musa (2022) o Anezi (2022), enfocan gran parte de la investigación en desarrollar una metodología que trabaje con las particularidades gramaticales de los idiomas utilizados.

Aunado a ello, se reconoce la carencia de trabajos enfocados en idioma español, pues, aunque existen trabajos enfocados en este idioma, la mayor parte de corpus desarrollados suelen encontrarse en idioma inglés, esto debido en gran medida al amplio uso de este idioma, así como de la disponibilidad de corpus ya etiquetados en inglés, resaltando la necesidad de generar corpus en idioma español que puedan ser utilizados para desarrollar arquitecturas de modelos más sofisticadas y enfocad en este idioma(Oriola y Kotze, 2020).

Las restricciones idiomáticas en la detección de discurso de odio ha sido una limitante en las tareas de aprendizaje computacional, pues gran parte de los esfuerzos se han centrado en la detección de discurso de odio en un idioma específico. No obstante, existen propuestas que abordan la utilización de modelos multilingües, este es el caso de Tellez et al., (2017), quien proporciona un marco multilingüe de fácil uso e implementación para el análisis de sentimientos, en donde, el objetivo central ha sido proporcionar un enfoque sencillo que consiste en traducir los mensajes al idioma de interés, para posteriormente utilizar un clasificador de opiniones optimizado para el idioma utilizado. Es decir, que el objetivo central de esta investigación es proporcionar una herramienta que facilite su uso para la clasificación de opiniones basadas en aprendizaje supervisado.

Sin embargo, es necesario considerar las particularidades del idioma en demarcaciones geográficas determinadas, pues la utilización de un idioma en diferentes países o regiones del mundo, no determina una misma forma de expresarse por medio de textos, ya que existen expresiones, símbolos o palabras que pueden estar dotadas de significados disímiles de acuerdo al contexto cultural y social, que aunque en un territorio determinado sigue siendo diverso, existe mayor similitud respecto a otro territorio.

Finalmente es necesario resaltar la ausencia de trabajos enfocados única y exclusivamente en la detección de discurso de odio hacia la comunidad LGBT en idioma español y específicamente en el territorio mexicano.

2.2 La necesidad de estudiar con datos a la comunidad LGBT

Históricamente las personas con orientaciones sexuales, identidades o expresiones de género diversas o no normativas han enfrentado contextos de segregación y discriminación, impulsando a esta colectividad a vivir realidades de violencia tanto física como psicológica. Estos contextos pueden generar barreras en el adecuado acceso a bienes y servicios como la salud, la educación, la participación política o el empleo, y en muchas ocasiones enfrentando la negación de derechos.

La discriminación y segregación a la comunidad LGBT responde en gran medida a estereotipos sexuales y pautas sociales que determinan comportamientos y apariencias aceptables de acuerdo al sexo de las personas. Estas percepciones sociales de lo aceptable han generado prejuicios sobre las diversidades sexo-genéricas que no correspondan al binario heterosexual (Bolaños y Chanrry, 2018).

Tales prejuicios impulsan una impresión generalmente negativa sobre la comunidad LGBT, lo cual promueve diversas actitudes hostiles e incluso agresivas. Ejemplo de esto es el discurso de odio que prevalece en redes sociales como Twitter, en donde es posible difundir y acceder a este tipo de discursos de manera sencilla y rápida (da Silva y da Silva, 2021).

Este tipo de mensajes ejercidos a través de insultos, humillaciones, promover o enaltecer a la violencia o incluso en forma de bromas, entre otras, representa una problemática al perpetuar la estigmatización de este grupo poblacional y a la vez, generando posibles afectaciones a la salud mental de los y las afectadas, así como de problemas en el desenvolvimiento social y un posible incremento de actos de violencia hacia esta colectividad (da Silva Anes, 2021) (Rodríguez Zepeda, 2018).

De esta forma, se reconoce la importancia de estudiar la distribución de este tipo de discursos en el territorio mexicano como una forma de reconocer posibles espacios de peligro y vulnerabilidad para las personas pertenecientes a la comunidad LGBT. Aunado a ello, se reconoce la importancia de generar bases de datos y estadísticas que muestren en este caso la importancia de estudiar a esta población desde los datos.

La Comisión Internacional de Derechos Humanos (2018:134), reconoce que, “los Estados no disponen de estadísticas confiables que reflejen la verdadera dimensión de la discriminación sufrida por las personas LGBT en el continente americano, lo que invisibiliza sus necesidades y facilita la subsidencia de estereotipos y prejuicios que contribuyen a perpetuar una situación histórica de estigma y exclusión”.

La identificación de la distribución del discurso de odio hacia la comunidad LGBT puede fungir como un acercamiento a la generación de estadísticas a nivel nacional de la violencia hacia esta población, generando instrumentos que visibilicen posibles desafíos que enfrenta esta minoría, y a la vez, reconociendo posibles puntos de interés en la toma de acciones.

Reconociendo que en los últimos años la creciente propagación del discurso de odio en redes sociales y medios masivos de comunicación ha incrementado la necesidad de analizar este tipo de mensajes, se resalta la importancia de esta investigación, en donde a través de la implementación de modelos de aprendizaje superficial y aprendizaje profundo, la clasificación de discurso de odio hacia la comunidad LGBT, permitirá automatizar una tarea que con la utilización de otras herramientas significaría un alto costo de ejecución.

2.3 Esfuerzos para generar estadísticas LGBT

La Comisión Internacional de los Derechos Humanos señala que, de acuerdo a los "Principios de Yogyakarta +10", reconocidos como principios y obligaciones estatales internacionales sobre la aplicación de legislación de derechos humanos relacionados a la orientación sexual, identidad y expresión de género, que fueron adoptados en noviembre del 2017 y de los cuales México es miembro, hace mención a la obligación de los Estados a recolectar estadísticas, y generar investigaciones que tengan como principio la identificación tanto de la prevalencia, como de la magnitud, causas y efectos de la violencia, discriminación y otros daños a las personas por motivos relacionados a la diversidad sexo-genérica como la orientación sexual, identidad y expresión de género o características sexuales. No solo con la intensión de cuantificar la eficiencia de las medidas diseñadas para prevenir dicha violencia, sino también, como una forma de reconocer la vulnerabilidad a la que se enfrenta esta población e identificar áreas de oportunidad y proponer acciones que mejoren la calidad de vida de las personas (Comisión Internacional de Derechos Humanos, 2018).

Sin embargo, el Consejo para Prevenir y Eliminar la Discriminación de la Ciudad de México (COPRED, 2018) reconoce la escasez de datos oficiales públicos que hagan referencia al peso poblacional de las personas pertenecientes a la comunidad LGBTTTIQ+.

No obstante, al examinar que diversos países como Reino Unido, Canadá, Estados Unidos o Nueva Zelanda entre otros, han considerado a este grupo poblacional dentro de sus planes y programas políticos, el estado mexicano sentó las bases en el país al desarrollar esfuerzos como la Encuesta Nacional sobre Diversidad Sexual y de Género (ENDISEG), reconociendo que aún son pocos los proyectos enfocados a la generación de estimaciones estadísticas acerca de las características de esta población.

ENDISEG, es un proyecto desarrollado por el Instituto Nacional de Estadística y Geografía (INEGI) que tiene como objetivo examinar diferentes aspectos de la población de 15 años y más, acerca de sus características sexuales identidad de género y orientación sexual, así como desarrollar un acercamiento sobre su percepción de la violencia a la que puede estar expuesta este grupo poblacional de la Diversidad Sexo-genérica que salen de la identidad de género y orientación sexual normativa. Esta encuesta tiene como objetivo principal responder a la pregunta ¿cuántos somos en la población LGBTI+?, buscando la obtención de información acerca del volumen de esta población en el país (INEGI, 2023).

ENDISEG, tiene como antecedentes algunas encuestas no probabilísticas, generalmente desarrolladas en línea que han buscado construir información acerca de las características y condiciones de la población LGBTIQ+ en México, tal es el caso de la Encuesta Nacional sobre Discriminación (ENADIS) en 2017, realizada por el INEGI con colaboración del Consejo Nacional para Prevenir la Discriminación (CONAPRED), así como la Encuesta Nacional de Cultura Cívica (ENCUCI) en 2020, desarrolladas por INEGI, en cuyas encuestas se incluyeron un par de preguntas acerca de la autodeclaración de la orientación sexual e identidad de género, sin embargo, estas encuestas no se enfocaron en la población LGBT per se (INEGI, 2023).

Fue en abril y marzo del 2018, que CONAPRED y el Consejo Nacional para Prevenir la Discriminación, desarrollaron la Encuesta sobre Discriminación por motivos de Orientación Sexual e Identidad de Género (ENDOSIG), cuyo objetivo fue conocer las opiniones, expresiones y experiencias de discriminación,

violencia y exclusión a las que se enfrentan las personas de acuerdo a su orientación sexual e identidad o expresión de género (INEGI, 2023).

Tomando como antecedentes los esfuerzos antes mencionados, y tras ajustar distintos aspectos metodológicos, ENDISEG realizó su primer levantamiento para la ENDISEG 2021, a través de audio-entrevistas al reconocer que el tema objetivo representa un alto grado de sensibilidad. Esta encuesta fue aumentada a una plataforma web en el 2022 con la finalidad de ampliar el alcance de la participación de la población LGBTIQ+ (INEGI, 2023).

Aun cuando dicha encuesta no está enfocada estrictamente a la violencia hacia las personas LGBT, se toman algunos temas clave como: características personales, sexualidad, orientación sexual, Identidad de género, salud emocional, rechazo social y opinión, así como apertura social, entre otros. No obstante, uno de los esfuerzos más grandes del INDESIG, es la posibilidad de mostrar una estimación de la población de 15 años auto percibida como parte de la comunidad LGBTIQ+ a nivel estatal (véase Figura 2.1)[2].

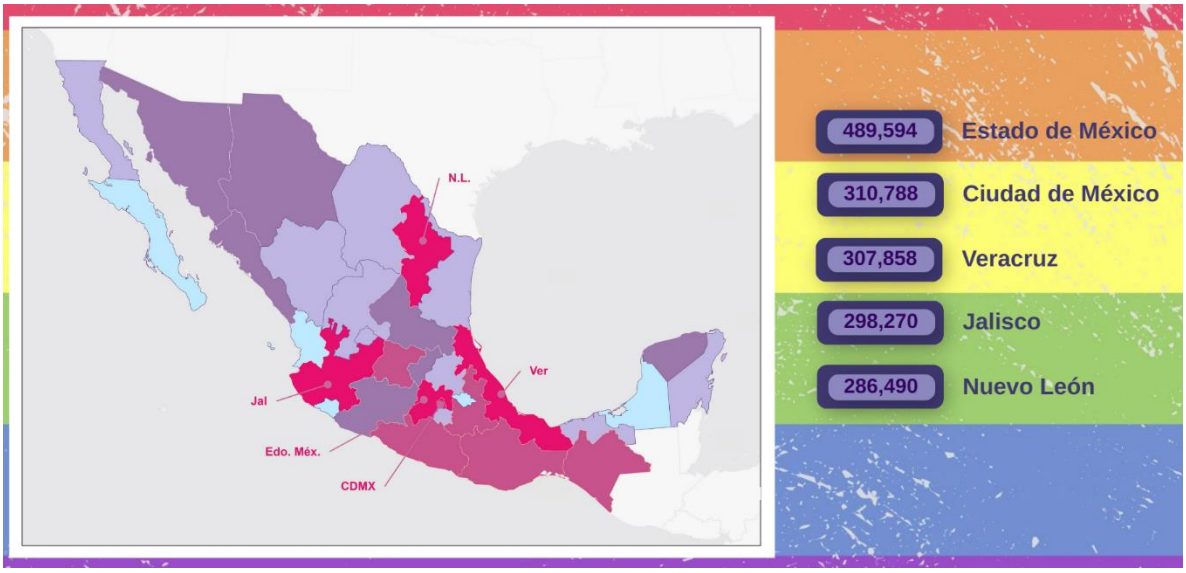


Figura 2.1 Población de 15 años y más que pertenece a la comunidad LGBT por entidad federativa. Elaboración propia con información de INDESIG (2021)

No obstante, como se ha mencionado en el capítulo anterior, una de las problemáticas más importantes al estudiar a la población LGBT, es la violencia e invisibilización a la que puede estar expuesta dicha población, en donde se reconoce la ausencia de bases de datos oficiales.

Por tal motivo, surgieron algunos proyectos como “Visible”, reconocida como la primera plataforma en línea enfocada en la recolección de reportes de violencia y discriminación cometidos hacia las personas LGBT en México, eventos que pueden ser reportados tanto por las propias víctimas como por otras personas que hayan sido testigos del incidente. Bajo el preámbulo de ser una herramienta colaborativa, confidencial y de fácil acceso, Visible reconoce la escasez de información relacionada con personas LGBT

[2] Es imperativo reconocer que, de acuerdo a la naturaleza de la encuesta, los datos mostrados son estimaciones obtenidas de acuerdo a la metodología utilizada por el INEGI, y que algunos de estos datos pueden tener un coeficiente de variación moderado; tal es el caso de algunos estados como: Baja California, Ciudad de México, Estado de México, Puebla o Veracruz, entre otros.

en México, sin la cual es difícil impulsar y desarrollar leyes efectivas que mejoren la realidad actual (Visible, 2021).

En la búsqueda de recolectar información acerca de la violencia y actos de discriminación hacia la población LGBT, Visible no sólo ha recolectado reporte ciudadanos, sino también, ha hecho uso de las redes sociales en la búsqueda manual de actos violentos que hayan ocurrido dentro del territorio mexicano. Al reconocer el carácter colaborativo de la plataforma, no es de sorprender que el alcance de la misma ha incrementado con el paso de los años, ejemplo de esto es lo que se muestra en la Figura 2.3, sin embargo, aun cuando en la plataforma digital se pueden visualizar los datos a partir del año 2017, las bases de datos que proporcionan de manera libre, arrojan registros desde el 2015, en donde es posible reconocer algunas entidades federativas como Veracruz, Chihuahua, Guerrero, Estado de México o la Ciudad de México, como aquellas en donde existen mayores reportes de violencia o discriminación hacia las personas pertenecientes a la comunidad LGBT (véase Figura 2.2).

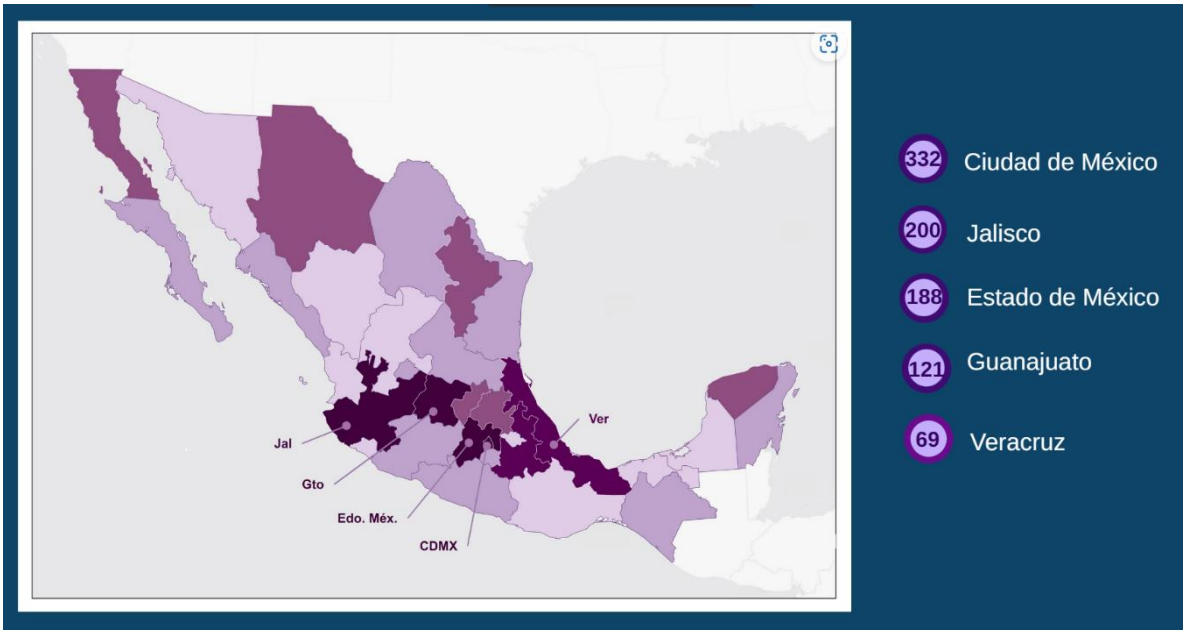


Figura 2.2 Agresiones contra personas LGBT en el 2022. Elaboración propia con información de Visible (2023)

Figura 2.3 Frecuencia de reportes recibidos en plataforma Visible en el año 2022. Visible (Visible,2023)

Visible permite realizar reportes acerca de agresiones verbales, psicológicas, económicas, físicas o sexuales, además de otros aspectos como, negación de derechos, servicios y acceso a diversos establecimientos públicos o privados, así como abuso de autoridad. En donde se observa que gran parte

de las víctimas reportadas se encuentran en un rango de 18 a 35 años y que el tipo de agresión más recurrente es la agresión verbal, seguida de la psicológica, física y sexual. Además, es posible visualizar que las orientaciones sexuales más recurrentes de las víctimas de acuerdo a los datos de Visible son: gay, bisexuales y pansexuales (Visible 2023) (véase Figura 2.4).

Figura 2.4 Rango de edad y orientación sexual de las víctimas reportadas. Visible (2023)

No obstante, existen algunos otros esfuerzos como Letra S, Sida, Cultura y Vida Cotidiana, AC (LetraEse), identificada como una organización civil sin fines de lucro, la cual ha dedicado esfuerzos al estudio y recolección de información acerca de la salud, sexualidad y sociedad, así como a la defensa de los derechos LGBT y de las personas que viven con Virus de Inmunodeficiencia Adquirida (VIH).

Dentro de la línea de trabajo de LetraEse, denominada documentación e investigación, se realiza el informe anual de Crímenes de Odio hacia las personas LGBT, donde a través de un monitoreo de medios digitales se identifican eventos catalogados como crímenes de odio, en donde, al distinguir la falta de registros de datos oficiales sobre la violencia ejercida hacia la comunidad LGBT, LetraEse ha desarrollado esta documentación de muertes violentas contra este grupo poblacional (LetraEse,2023).

Este monitoreo de notas de prensa y notas informativas es revisado y clasificado en una base de datos dividida en 6 secciones y 28 unidades de análisis, con el fin de desglosar de manera amplia la información proporcionada en estos medios. En dicha tarea, solo en el 2022, se registraron 87 homicidios cometidos hacia personas LGBT, los cuales muestran una carga violenta, denotando que muy posiblemente estén impulsados por el odio (LetraEse,2023).

LetraEse reconoce que los homicidios registrados en el último lustro suman al menos 453 muertes violentas (véase Figura 2.5), sin embargo, es imperativo reconocer que la cifra real debe ser mayor, pues debido a la naturaleza de los eventos, el registro, rastreo y reporte de los mismos puede ser complejo. El esfuerzo realizado por esta asociación civil ha permitido identificar algunos estados en donde la cantidad de eventos registrados son altos, tal es el caso de Veracruz, Chihuahua y Guerrero, que entre el 2015 y el 2022, registraron un total de 95, 58 y 52 asesinatos violentos respectivamente (véase Figura 2.6).

Figura 2.5 Homicidios violentos contra personas LGBT. Visible (2022)

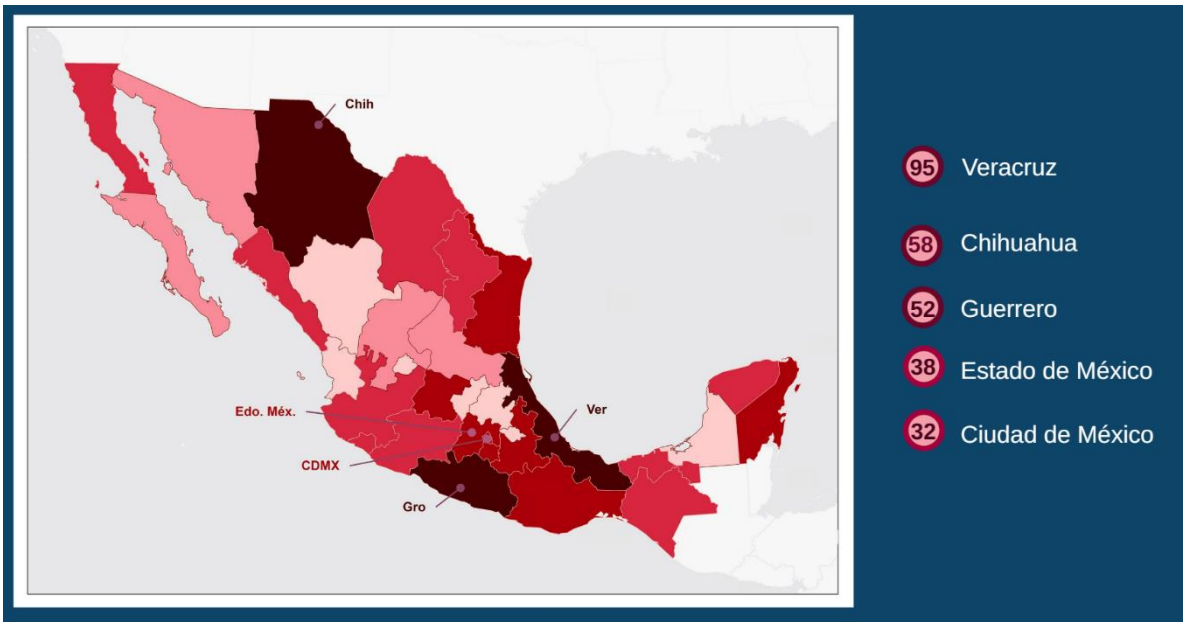


Figura 2.6 Homicidios de personas LGBT en México 2015-2022. Elaboración propia con información de LetraEse (2023)

2.4 Objetivo general:

Zonificar la presencia de discursos de odio en México hacia la comunidad LGBT en Twitter.

2.5 Objetivos particulares:

- Explorar y documentar en Twitter las combinaciones de palabras en español utilizadas para referirse a la comunidad LGBT en México desde el discurso de odio.
- Generar y estructurar una muestra de datos etiquetados, que permita identificar discurso de odio hacia la comunidad LGBT en Twitter a través de un clasificador.
- Entrenar un algoritmo de clasificación usando procesamiento de lenguaje natural para la identificación de tuits asociados a discursos de odio contra la población LGBT.
- Analizar la distribución espacio temporal de los tuits identificados como discurso de odio.

3 Metodología

En la presente investigación, se ha optado por utilizar información de Twitter por las características antes mencionadas en cuanto a la relevancia de la información generada diariamente en redes sociales, así como por la posibilidad de encontrar mensajes mucho más genuinos y que se alejen de la deseabilidad social, además de reconocer Twitter como una plataforma de gran popularidad en México, que aunque no es una plataforma representativa de todos y todas las ciudadanas del país, la naturaleza de la misma, permite una fácil y rápida difusión de contenido diverso, posibilitando una fácil viralización de las publicaciones realizadas, por lo cual, es de suma importancia analizar este tipo de mensajes, especialmente al considerar que dicha plataforma mantiene una apertura de sentimientos y opiniones acerca de temas variados, entre los cuales es posible encontrar discurso de odio o algunas otras comunicaciones violentas (Arcila Calderón, Blanco-Herrero y Valdez Apolo, 2020).

De esta forma, como se ha mencionado anteriormente, la viralización del discurso de odio hacia la comunidad LGBT puede traer consigo diferentes repercusiones negativas no sólo hacia la población objetivo, sino también, hacia el resto de personas que puedan tener acceso a este tipo de comentarios y comunicaciones violentas, impulsando la reproducción de contenido violento, y perpetuando estigmas y prejuicios sobre la comunidad LGBT.

En el presente capítulo se desglosa de manera detallada la metodología seguida para la obtención de los resultados finales. El flujo de trabajo metodológico se muestra en la Figura 3.1, a través de 5 fases clave: la adquisición de datos, el etiquetado de la muestra, el entrenamiento del modelo y la evaluación del mismo, para finalizar con la zonificación de la presencia de discurso de odio hacia la comunidad LGBT con textos cortos de Twitter.

Figura 3.1 Flujo de trabajo

3.1 Delimitación del área de estudio

Las investigaciones centradas en la detección y clasificación de discurso de odio en medios masivos de comunicación encuentran una gran barrera en la delimitación del idioma. Aun cuando existen esfuerzos por desarrollar metodologías que permiten el procesamiento y entrenamiento sin esta barrera, los esfuerzos desarrollados se han mostrado como una opción en el acercamiento a una clasificación más compleja.

Por otra parte, es de gran importancia considerar que existen diversas formas de expresar discurso de odio, así como diferentes formas de entender y darle sentido a una misma expresión. Está divergencia existe notoriamente en la utilización de diferentes idiomas, sin embargo, también es posible distinguir estas diferencias en un mismo territorio, pues el habla en un país determinado responde a particularidades regionales y locales que divergen en las expresiones realizadas.

Considerando la necesidad de desarrollar estadísticas y datos que estudien la realidad de la población LGBT, se resalta la importancia de mantener un enfoque en el territorio mexicano como zona de estudio. Identificar la violencia a la que puede estar expuesta esta población ha sido abordada de forma general por las instituciones gubernamentales, reconociendo el esfuerzo de instituciones no gubernamentales por generar datos enfocados en dicha problemática.

De igual manera, como se mencionó en el capítulo uno, tanto la violencia como la discriminación hacia las diversidades sexo-genéricas no normativas responden a factores socio culturales arraigados en la sociedad perpetuando estigmas y prejuicios sobre esta población. Considerando el desarrollo de la sociedad mexicana sobre un contexto altamente machista, no es de extrañar altos índices de discriminación, estigmatización, invisibilización y desarrollo de prejuicios hacia la población LGBT.

Por estos motivos, y como un esfuerzo de adentrarse a la zonificación del discurso de odio, se reconoce al territorio mexicano como una zona de gran importancia en donde es necesario desarrollar este tipo de estudios, facilitando la identificación de zonas de prevalencia, así como reconociendo el comportamiento temporal en la presencia de este tipo de comunicación agresiva.

Aunado a lo anterior y de acuerdo con datos del INDESIG (2021), en México hay cerca de 5 millones de personas auto percibidas con una orientación sexual y/o identidad de género no normativa; sin embargo, al considerar la delicadeza de este tipo de estadísticas y la dependencia en la auto adscripción y reconocimiento de las personas a este grupo poblacional, se puede indagar que el número real debe ser mucho mayor. Reconociendo así, a la comunidad LGBT como grupo focal de análisis en el estudio del discurso de odio.

3.2 Obtención y descarga de datos

Inicialmente, para la generación de la base de datos utilizada en el entrenamiento del modelo para la detección de discurso de odio hacia la comunidad LGBT, fue necesario partir de una exhaustiva exploración de tuits donde se hiciera mención a la población objetivo. Dicha exploración se efectuó con la intención de identificar palabras y combinaciones de palabras en español que son utilizadas con mayor frecuencia dentro de Twitter para hacer referencia a la comunidad LGBT en México. Esta, fue realizada de manera manual mediante el monitoreo constante de Twitter, y el rastreo de mensajes en los que se tuviera certeza sobre la referencia a la comunidad LGBT o hacia algún miembro de esta; dentro de dicho compendio no se realizó ninguna distinción entre palabras comunes para expresar discurso de odio o mensajes de apoyo (véase Anexo 4).

Tanto el monitoreo como la descarga de datos sigue una temporalidad precisa que parte desde el 01 de enero del 2020 hasta el 31 de diciembre del 2022. Esta temporalidad fue determinada tomando como referencia el inicio de la pandemia de COVID-19 en México, que, si bien tiene como primer registro el 28 de febrero del 2020, se decidió tomar el año completo permitiendo analizar la distribución del discurso de odio a través de los 3 diferentes años de estudio (2020, 2021 y 2022). Tomar como referencia el inicio de la pandemia de COVID-19, ha sido un punto clave al considerar que la utilización de las redes sociales por parte de los usuarios tuvo un aumento considerable con este hecho.

La descarga de tuits fue realizada mediante el Autómata Geointeligente En Internet (AGEI), desarrollado por el Laboratorio Nacional de Geointeligencia (GeoInt), en el cual fue preciso realizar un registro en espera de la autorización de los administradores (véase Figura 3.2).

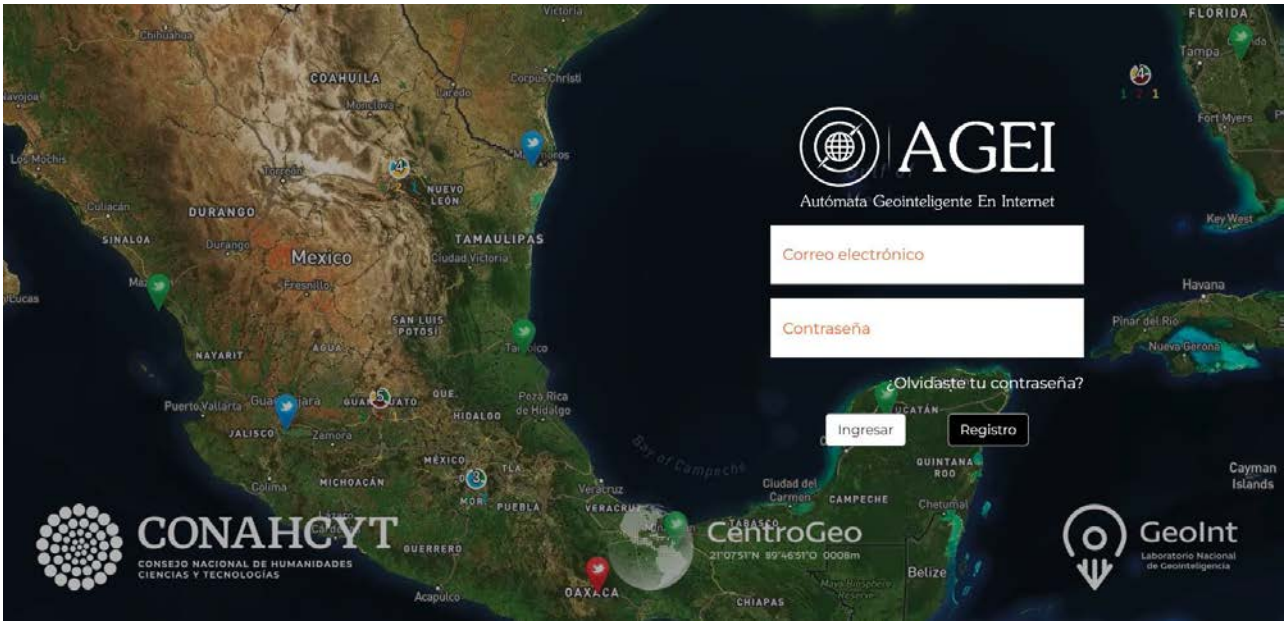


Figura 3.2 Plataforma AGEI. Imagen tomada de: http://agei.geoint.mx/

Esta plataforma cuenta con un repositorio amplio de los mensajes que han sido publicados en Twitter, sin embargo, es imperativo considerar que la descarga de los mismos fue realizada antes de que existiera un cambio en las políticas de Twitter que prohíben la descarga de información sin autorización previa. Además, los registros descargados fueron únicamente aquellos que contaban con una referencia espacial, es decir, latitud y longitud, permitiendo visualizar la localización exacta de la zona en donde el mensaje ha sido publicado[3] (véase Figura *3.3*). Para la descarga de los datos, se trazó un bounding box (véase Figura 3.4) con la intención de delimitar los mensajes al territorio mexicano, no obstante, fue necesario hacer una limpieza de estos datos, en función de descartar aquellos tuits que se localizaran fuera del territorio mexicano, ya sea en la frontera norte o sur del país; este proceso fue ejecutado a través de selecciones espaciales por medio el software de código libre Qgis.

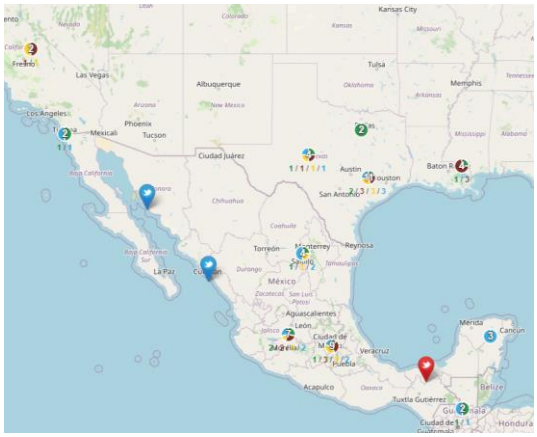


Figura 3.3 Datos geolocalizados

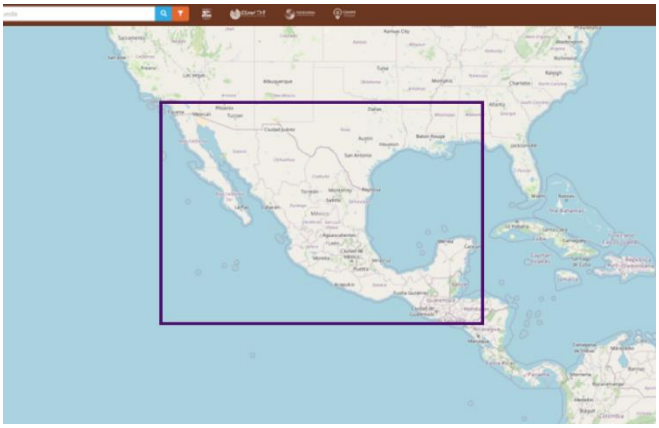


Figura 3.4 Boundin box de búsqueda

En la descarga de datos dentro del AGEI, se implementó una consulta que recopilara todos aquellos mensajes donde se utilizara al menos una de las palabras identificadas durante el monitoreo de Twitter (véase Anexo 3). La construcción de la consulta, debía ser ejecutada por medio del buscador de palabras a través de operadores clave como "|", para indicar la búsqueda de una palabra o conjunto de palabra y otra palabra o conjunto de palabras, además de especificar la temporalidad de búsqueda, para finalmente, solicitar la descarga de datos obtenida con los parámetros establecidos (véase Figura 3.5).

[3] Es esencial reconocer que esta localización de los tuits, corresponde a zonas representativas enmarcadas en pequeñas áreas de bounding boxes, agrupando los datos puntuales aislados en zonas más representativas.

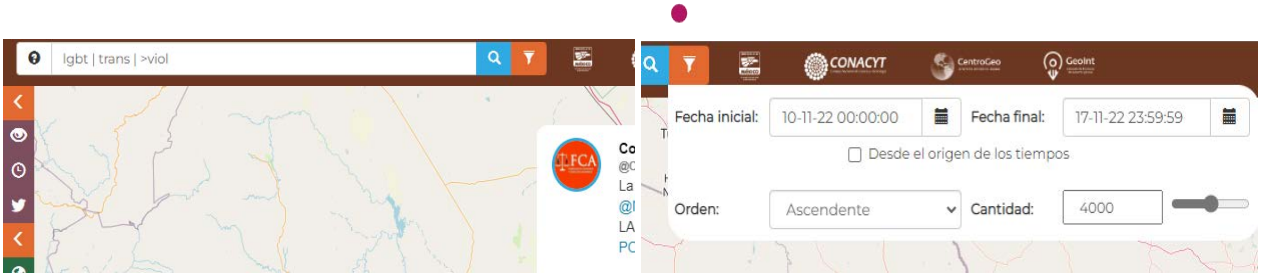


Figura 3.5 Ejemplo de consulta y temporalidad de búsqueda

Sin embargo, la cantidad de datos recolectados con dicho diccionario de palabras ascendía a los 300 mil tuits, en donde, se reconoció que algunos de estos mensajes no hacían referencian a la comunidad LGBT. Por este motivo, y siguiendo el trabajo de Arcila-Calderón et al., (2021), en el que se construyó un conjunto de palabras específicas que hicieran mención sobre discurso de odio por motivos de género y/u orientación sexual; en ésta primer base de datos descargada, fue necesario implementar una serie de consultas para filtrar y depurar de mejor manera los mensajes que conformaban el corpus de entrenamiento, a través de un conjunto de palabras mucho más grande en comparación al primer conjunto utilizado buscando las expresiones más recurrentes identificadas en el monitoreo efectuado, para diferenciar entre expresiones utilizadas para hablar desde el discurso de odio o aquellas utilizadas para apoyar o hacer una mención adecuada de la comunidad LGBT (véase Anexo 4).

Esta depuración fue imprescindible al considerar la gran cantidad de datos obtenida, la cual necesitaría de un esfuerzo abrumador en la tarea de etiquetado, y por otro lado podría representar problemas en la fase de entrenamiento del modelo, necesitando la implementación de técnicas mucho más robustas para desarrollar el entrenamiento.

Por la naturaleza de búsqueda del AGEI, las palabras utilizadas en el código de consulta toman como referencia la raíz del conjunto de palabras utilizadas, por este motivo es que se obtuvieron muchos registros que realmente no aportaban al objetivo de la investigación. Como ejemplo de esto se puede resaltar la implementación de la palabra “vestida” que en muchas ocasiones es utilizada para hacer referencia a una persona travesti de una forma despectiva, sin embargo, por la naturaleza de la plataforma y de la palabra en sí, los resultados obtenidos mostraron algunos registros que carecerían de utilidad para esta investigación, por ejemplo “que linda se ve mi hija con el nuevo vestido que le compré”.

Por tal motivo, a través del compendio de palabras para hacer el filtrado (véase Anexo 4), se construyó una consulta SQL, la cual fue ejecutada desde DBeaver con extensión postgis, no sólo con la finalidad de realizar el filtrado, sino también, como una forma de corroborar la distribución y presencia de tuits descargados dentro del territorio mexicano.

Por el tamaño de las bases de datos, y evitando el sobre esfuerzo computacional, dicho filtrado fue ejecutado de manera anual, obteniendo 3 diferentes bases de datos para cada año las cuales fueron unidas posteriormente para continuar con la fase de etiquetado, entrenamiento y evaluación.

3.3 Etiquetado de la muestra

Como parte esencial en la implementación del algoritmo, el corpus fue etiquetado de manera binaria, en donde cada texto descargado fue clasificado en discurso de odio y no discurso de odio hacia la comunidad LGBT o algún miembro o colectividad de la misma.

Tomando en consideración que las expresiones y palabras utilizadas para referirse a la comunidad LGBT desde el discurso de odio pueden cambiar con el paso del tiempo, se agruparon las bases de datos de los años correspondientes (2020, 2021 y 2022) en una misma base de datos, la cual se conformó de un total de 22,605 tuits. De esta base de datos, se tomó de forma aleatoria una tercera parte del total de registros, donde tras redondear la cantidad se obtuvo un total de 8,000 tuits para el proceso de etiquetado, el cual fue realizado de manera manual.

Tomando como referencia la metodología de Oriola y Kotze (2020), en donde cada una de las 3 muestras divididas de acuerdo al idioma de los textos (inglés y afrikaans, inglés e IsiZulu, e inglés y sesotho) fue clasificada por dos etiquetadores diferentes. En esta investigación, fue necesario la creación de un grupo de etiquetadores y etiquetadoras que validaran el total de la muestra de (8,000 tuits) del corpus original.

La base de datos sometida a etiquetado, inicialmente fue clasificada por una única persona, sin embargo, al realizar esta tarea de manera individual representa un sesgo considerable al tomar únicamente la perspectiva de una persona, por lo tanto, fue necesario realizar un proceso de validación en donde se integraron personas ajenas a la investigación.

Para realizar esta validación, un grupo de personas pertenecientes a la comunidad LGBT y ajenos a la misma, realizaron el etiquetado de manera individual. Al considerar la gran cantidad de datos a clasificar, la muestra se dividió en lotes de 500 tuits y cada lote fue etiquetado por dos personas diferentes, teniendo un total de 32 personas que realizaron dicha tarea.

Posteriormente, estas etiquetas fueron comparadas en conjunto a la clasificación desarrollada previamente, con el objetivo de distinguir concordancias y discrepancias, así como de realizar un doble chequeo en esta clasificación. Con tal comparación, fue posible distinguir expresiones y tipos de mensajes complejos de diferenciar entre discurso de odio o no discurso de odio.

Tanto para el etiquetado previo de la muestra como para el etiquetado desarrollado por las y los etiquetadores, se siguieron los siguientes parámetros:

Una publicación en Twitter fue clasificada como discurso de odio si:

- Incita a las personas a cometer actos de violencia hacia la comunidad LGBT o a algún miembro de esta por el hecho de ser parte de esta población.
- Si incita a la discriminación hacia la comunidad LGBT o a algún miembro de esta por pertenecer a este grupo poblacional.
- Si expresa o promueve estereotipos de la comunidad LGBT.
- Si utiliza lenguaje vulgar y/o agresivo para insultar a la comunidad LGBT o a algún miembro de esta por el simple hecho de ser parte de esta comunidad.

- Si se refiere a la diversidad sexo-genérica o utiliza categorías de la misma como una forma de insultar o denigrar.
- Si utiliza palabras que coloquialmente son empleadas para referirse a algún grupo de la diversidad sexo-genérica de manera despectiva o denigrante para hablar de una persona o grupo de personas de forma negativa o como forma de insultar o denigrar.
- Si apoyan o enaltecen comportamientos o comunicaciones que han sido reconocidas como agresivas o violentas hacia la comunidad LGBT.
- Si invalida la lucha del movimiento LGBT.

La naturaleza y sutileza con la que puede expresarse el discurso de odio, así como la dificultad de distinguir palabras utilizadas coloquialmente sin el deseo de ejercer un discurso de odio o contextualizar ciertas palabras que históricamente demuestran una connotación negativa al texto en cuestión, fueron puntos esenciales en el etiquetado de los datos, lo cual fue abordado en gran parte con la validación del etiquetado.

La forma en que se efectúa la comunicación en las redes sociales y sobre todo el impacto que puede causar en los y las lectoras del contenido que es publicado, responde a diversos factores sociales, culturales, políticos entre otros, teniendo como resultado que el impacto causado en el receptor pueda ser sumamente variante dependiendo de la persona que accede a este tipo de contenido.

El sentirse perteneciente a una comunidad o a un grupo de personas por n características puede permear en la forma en que las personas pueden entender los mensajes, es decir, que el impacto que pueda representar un comentario negativo hacia la comunidad LGBT no es percibido de la misma manera por una persona ajena a esta comunidad que por ejemplo una persona transexual, homosexual o no binaria. Esto no quiere decir que todas las personas LGBT reconozcan un comentario negativo de la misma manera, sin embargo, al considerar que gran parte de estas personas han mantenido historias de vida o vivencias que encuentran similitudes, por lo menos en el rechazo social en relación a su identidad de género u orientación sexual, o bien que pueden mantener un mayor acercamiento a la cultura LGBT, el impacto y distinción puede ser diferenciada entre personas LGBT y no pertenecientes a esta, específicamente en este caso al distinguir discurso de odio hacia el grupo poblacional.

El impacto o repercusión que pueda traer consigo este tipo de mensajes, tiene importancia en cualquier persona que acceda a este tipo de contenido, sin embargo, no es el mismo impacto que tendrá sobre la persona o grupo de personas que son objetivo de comentarios y contenido agresivo o violento, que puede caer por supuesto en el discurso de odio, sin embargo, también representa un impacto significativo en cualquier receptor al apoyar y promover este tipo de comunicaciones.

No obstante, esta consideración fue esencial en el proceso de etiquetado, ya que, al estar bajo la perspectiva de una gran cantidad de personas, fue posible distinguir divergencias en la forma de clasificar entre una persona y otra, es decir que, aunque existía homogeneidad en la forma de etiquetar de una misma persona, existían discrepancias con la forma de etiquetar de otra persona.

Estas consideraciones fueron identificadas especialmente en aquellos tuits que se expresaban desde un discurso de odio hacia la comunidad LGBT de formas sutiles o en aquellos mensajes que parecían no hacer referencia directa a esta población. No se pretende suponer que las personas no pertenecientes a la comunidad LGBT no puedan identificar discurso de odio dirigido a esta población, sin embargo, muy

probablemente la percepción de ciertos comentarios o palabras agresivas no representen el mismo impacto, esto se relaciona sobre todo en el discurso de odio que es efectuado de una manera muy sutil.

Para el proceso de verificación del etiquetado, se consideraron aquellas formas de clasificar que se repetían mayormente en los lotes de los y las etiquetadoras, es decir que, expresiones similares identificadas en diferentes tuits que fueron clasificadas de manera contrastante por parte de los etiquetadores, fueron homologadas tomando en consideración la manera predominante de clasificar. Por otra parte, en aquellos mensajes que se expresaban de formas sutiles, se dio mayor peso al etiquetado realizado por personas pertenecientes a la comunidad LGBT.

Al desarrollar el análisis y comparación entre la clasificación de los diferentes etiquetadores que tuvieron participación en el etiquetado del corpus, se identificaron ciertos patrones. El claro ejemplo de esto ha sido la clasificación de tuits en los que se expresan palabras agresivas o expresiones que estereotipan o denigran a la comunidad LGBT o a algún grupo de esta.

Al contener un contexto en el que la expresión de dichas palabras son empleadas para referirse a sí mismos o como parte de un comportamiento utilizado en las dinámicas sexuales dentro de las redes sociales, algunos etiquetadores clasificaron dichos tuits como no discurso de odio, sin embargo, tomando como referencia el factor de estereotipación del discurso de odio, este tipo de comunicación ha sido clasificada como odio, considerando también que fue así como se clasificó por la mayor parte de las y los etiquetadores.

Censurar completamente algunos tipos de comunicaciones y en específico la utilización de ciertas palabras y expresiones, específicamente en plataformas de redes sociales ¿ayuda a contrarrestar la prevalencia de discursos de odio o de este tipo de comentarios, o más bien impulsan a un rechazo, desagrado y odio más grande al sentirse censurados, bajo el escudo de la libertad de expresión?

Durante el proceso de etiquetado, se identificaron de manera recurrente la presencia de mensajes que mencionaban la frustración de haber sido sancionados en algunas otras plataformas como Facebook por expresar discurso de odio hacia la comunidad LGBT, por lo que encontraban en Twitter una plataforma donde podían expresarse con mayor libertad.

Para esta investigación, tal cuestionamiento se presenta como parteaguas para futuras investigaciones, sin embargo, fue un pilar esencial en el desarrollo del etiquetado. Considerar discurso de odio la utilización de ciertas palabras reconocidas como palabras agresivas que históricamente han sido utilizadas para referirse a la población LGBT de una forma agresiva o denigrante implicaría la censura de una gran cantidad de expresiones.

Lo cierto es que, la utilización de algunas palabras como “puto” han ampliado su significado en relación a la forma de expresarse. Actualmente, el uso de esta palabra es más común de lo que parece, y en muchas ocasiones es utilizada para expresar desagrado, molestia, inconformidad, o incluso para enaltecer un comportamiento o a una persona o grupo de personas, esto es algo que se ha observado por lo menos en el territorio mexicano y en el análisis del corpus desarrollado.

No obstante, la utilización de dicha palabra también es fuertemente utilizada para referirse a la comunidad LGBT o algún integrante de esta de una forma despectiva o denigrante, así como sinónimo de una persona introvertida, temerosa, también como sinónimo de cobarde o sin valor.

Por los motivos anteriormente plateados, a continuación, se muestran algunas particularidades que fueron identificadas durante el proceso de etiquetado y que de la mano a los puntos clave para clasificar un discurso de odio fueron tomados en el proceso de revalidación del primer proceso de etiquetado.

Considerando los diferentes usos que puede tener una palabra, especialmente aquellas que históricamente han sido utilizadas de manera despectiva para ofender denigrar o estigmatizar; un tuit se consideró discurso de odio cuando este tipo de palabras se utiliza como un sinónimo de cobarde, tímido o que alguien carece de valor o validez:

- “No los veo depositando para unas guamas Indio, la light es para putitos.”
- “Un puto maricón seudo influyente ... chale”

Cuando se utiliza para referirse a un grupo de personas o institución de manera despectiva:

- “Putos del Monterrey valen verga.”
- “Chinga tu madre pinche gobierno puto, puro maricón.”

Si reproduce expresiones o se dirige a una persona con palabras que denigran, estigmatizan o son insultantes:

- “Puto el que lo lea”
- “Fea, gorda, nerd, dientona, la que hablaba bien cagado (hasta la fecha) y marimacha que jugaba fútbol. Jiji.”

En aquellos mensajes con connotaciones sexuales que hacen uso de este tipo de palabras impulsando estereotipos negativos de la comunidad LGBT al utilizar expresiones denigrantes, aun cuando se expresan en primera persona:

- “Que rica ensartada de verga le di a este putito.”
- “Soy tv de closet, 26 años y muy putito. Masaje tántrico 40 min.”

También se consideraron aquellas expresiones de odio que no se expresan de forma evidente:

- “Me saca de pedo como la comunidad LGBT se ofende por comentarios simples, damn bro, existe el mame y tú también lo haces. No quiere decir que todos estemos en su contra.”
- “Las mujeres trans son mujeres TRANS”
- “De entrada si es Americanista no es tu novio... Es solo una amiga marimacha”
- “Siempre he pensado que el encargado de poner la música en el Coppel de la Alhóndiga, es miembro de la comunidad LGBT+ Cada que paso por ahí, siento que llueve diamantina.”

Por otra parte, no se consideró discurso de odio cuando la palabras o grupo de palabras no se utiliza para hacer referencia a una persona o grupo de personas:

- “Pinche puto coraje.”
- “Que puto estrés.”
- “Pinche puto perro frío.”
- “QUE PINCHE PUTO ASCO LA CONTAMINACIÓN EN CDMX!!!”
- “No puedo dejar whats por el puto trabajo, chinguen a su madre.”

Cuando se utiliza para enaltecer a una persona, grupo de personas o una situación específica:

- “Eres un puto CRACK.”
- “Eres un puto Dios.”
- “Ay dios, chinguen a su madre, LO AMEEEE me hizo mi PUTO DIAAA!!!”

Cuando este tipo de palabras son utilizadas para hacer referencia al virus COVID-19:

- “Te Odio puto Coronavirus 😢”

Al considerar que el punto objetivo en el etiquetado de los datos era el discurso de odio, tampoco se consideró como tal aquellos reportes sobre discriminación hacia la comunidad LGBT o aquellos que ejemplificaban una situación específica, pero se mostraban en contra de tal situación:

- “Jueza niega amparo a Colectivo de Yucatán, les pide pruebas de que son personas LGBT”
- “Mis papás se están burlando de mujeres trans en un programa de televisión, ¿a dónde vamos a parar?”
- “Hace año y medio, mientras me golpeaba, el agresor decía: te voy a matar pinche puto. Es bien bonito defender esa palabra ojalá, a quien lo hace, se lo digan de la misma forma y sabrá.”

3.4 Entrenamiento y evaluación del modelo

Tal como se mencionó en el apartado de antecedentes, la implementación de arquitecturas complejas en la construcción de modelos enfocados en la clasificación de textos cortos ha ido en aumento en los últimos años. Sin embargo, tomando en consideración el objetivo de la investigación en desarrollar un acercamiento a la utilización de este tipo de modelos en la detección de discurso de odio hacia la comunidad LGBT en Twitter, para la fase de entrenamiento y evaluación del modelo, se seleccionaron 4 algoritmos específicos tomando como referencia los buenos resultados observados en investigaciones previas.

Dentro del aprendizaje superficial se utilizó un modelo basado en Máquinas de Soporte vectorial (SVM) y otro basado en regresión logística (LR); desde el aprendizaje profundo, se utilizaron de igual manera dos modelos diferentes, por una parte, uno basado en Redes Neuronales Convolucionales (CNN) y otro basado en Redes Neuronales Recurrentes (RNN).

Estos modelos fueron elaborados y programados completamente en lenguaje de programación Python a través de un editor de código, en donde fue necesaria la implementación de algunas librerías tanto para el manejo de los datos como para el preprocesamiento, desarrollo y entrenamiento del mismo. Dentro de estás, se reconoce la utilización de pandas, numpy, matplotlib, seaborn, tensorflow con la paquetería de keras, y diversas paqueterías de sklearn. El desarrollo de estos modelos se ejemplifica en la Figura 3.6.

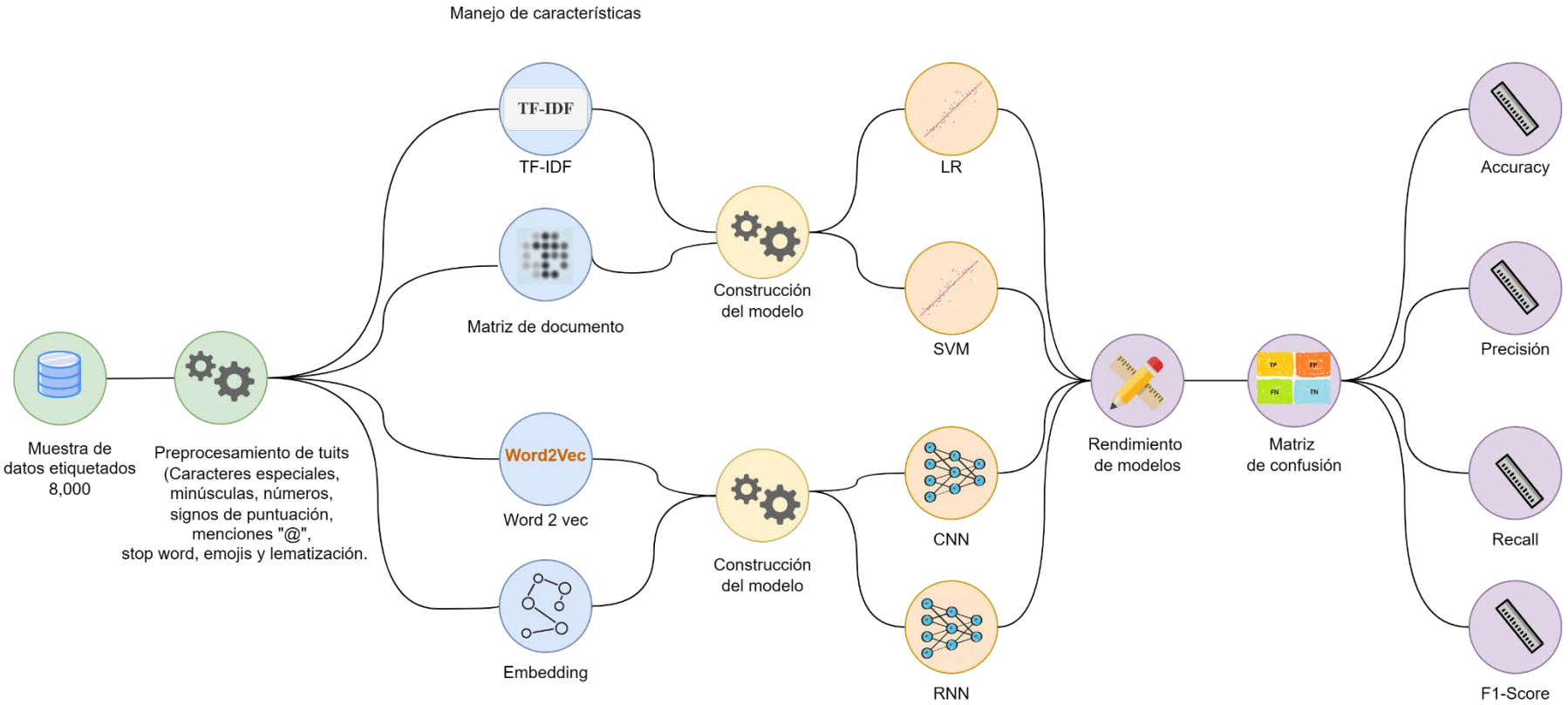


Figura 3.6 Flujo de trabajo para la construcción de los modelos. Elaboración propia con referencia de Baydoğan y Alatas (2022).

De manera anticipada a la construcción de los modelos, fue necesario desarrollar una fase de preprocesamiento, en donde se estandarizó y homologó la estructura de los textos de la base de datos utilizada, a través de la conversión de texto a minúsculas, la eliminación de espacios dobles o la eliminación de acentos. Aunado a ello, se implementaron paqueterías como “re”, para la eliminación de caracteres especiales y números, así como la eliminación de menciones de cuentas identificadas en Twitter con el símbolo “@”; también, se implementaron algunas otras librerías como “nltk”, para el manejo de las stop-word, “spacy”, para la lematización de los textos en español, y, “Unicode emoji” para la sustitución de emojis coloquiales por textos fijos.

Con los textos estandarizados en la fase de preprocesamiento, el corpus utilizado fue dividido 80/20 para la fase de entrenamiento y validación respectivamente. Tanto para el modelo de SVM como para el modelo de LR, se creó una matriz de características tomando los tuits de entrenamiento con ayuda de la paquetería sklearn, y posteriormente se ajustaron los vectores tomando los valores del entrenamiento en representación vectorial basado en el recuento de palabras TF-IDF (Term Freuency-Inverse Document Frecuency), de esta manera, por medio de la matriz de características, los datos pueden ser procesados por el modelo. Ambos modelos fueron creados utilizando la función “LogisticRegression” y “SVM” respectivamente a través de sklearn. Para el modelo LR, se agregó una regularización L1 (LASSO) con optimizador saga, con el objetivo de seleccionar algunas características y exclusión de algunas otras, evitando el sobreajuste u overfitting.

Para los modelos basados en aprendizaje profundo, se creó un modelo Word2Vec de tamaño de dimensión igual a 100, y con conteo mínimo de palabras igual a 3, excluyendo aquellas palabras que en el corpus tengan presencia menor a dicho número. En ambos modelos se ejecutó una tokenización y codificación de los textos con ayuda de “keras”, y estas secuencias se ajustaron nuevamente con ayuda de keras a una longitud fija de 100, que, de acuerdo a Zhang, Robinson y Tepper (2018), es un número idóneo para obtener las características más significativas de un tuit. Además, se creó una matriz de incrustaciones de palabras o embeddings tomando como referencia el modelo de Word2Vec creado con antelación, posteriormente se agregó una capa de convolución 1D.

El modelo CNN, se inició con una capa de incrustación de palabras que toma los valores creados en la matriz de embeding y donde se considera la dimensión de los vectores de palabras, así como la longitud máxima de las secuencias; posteriormente, se creó una capa convolucional 1D con 50 filtros y un tamaño de ventana igual a 4 y una función de activación relu, también se agregó una capa de abandono o “Dropout” igual a 0.2 para regularizar el sobreajuste por medio de la eliminación aleatoria evitando la dependencia de algunas palabras. Para reducir la dimensionalidad y extraer características significativas, se agregó una capa de agrupación MaxPooling, la cual es transformada a un vector unidimensional, para finalmente tener una capa densa de salida, con una función de activación sigmoide, así como una función de regularización L1 con un valor de penalización de 0.01, de nueva cuenta para evitar el sobreajuste. Esta arquitectura fue compilada para crear el modelo de entrenamiento, el cual se estableció con un total de 50 épocas, un tamaño de lote igual a 64 y un optimizador adam (mejorando la pérdida y el estocástico clásico).

El modelo basado en RNN, siguió un procesamiento de los datos similar a CNN, hasta la construcción del modelo, en donde, de nueva cuenta se inicia con una capa de incrustación de palabras, pero en este caso

se agrega una capa recurrente LSTM y una capa de abandono igual a 0.4, para finalmente tener una capa densa con una función de activación sigmoide y una función de regularización L1 igual a 0.01, para posteriormente compilar y entrenar el modelo con los mismos parámetros utilizados en CNN.

Las métricas de evaluación utilizadas para medir el rendimiento de los modelos implementados fueron: Accuracy, Recall y Valor-F1 (véase Tabla 1.5 y Tabla 1.6), además se obtuvo la matriz de confusión para cada uno de los modelos, con la cual fue posible distinguir y cuantificar el total de verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos (véase Figura 1.8). Al obtener estás métricas de evaluación para cada algoritmo implementado, fue posible distinguir el modelo y arquitectura con mejor rendimiento, el cual fue seleccionado para la clasificación del resto de datos que no fueron etiquetados de manera previa.

4 Resultados y discusión

Siguiendo la metodología presentada en el capítulo anterior, en el presente apartado se muestran los resultados obtenidos tras la implementación de los cuatro modelos utilizados (LR, SVM, CNN y RNN).

Por medio de dos grandes secciones, este capítulo parte con el análisis de los resultados y rendimientos obtenidos con los modelos anteriormente mencionados para la detección de discurso de odio hacia la comunidad LGBT en tuits georreferenciados. Posteriormente, se desglosa la fase de espacialización, en la cual tras la implementación del modelo seleccionado para la clasificación del resto de tuits que conforman la base de datos original, se desarrolló la cartografía pertinente. Esto ha permitido observar la distribución espacio-temporal de los datos obtenidos.

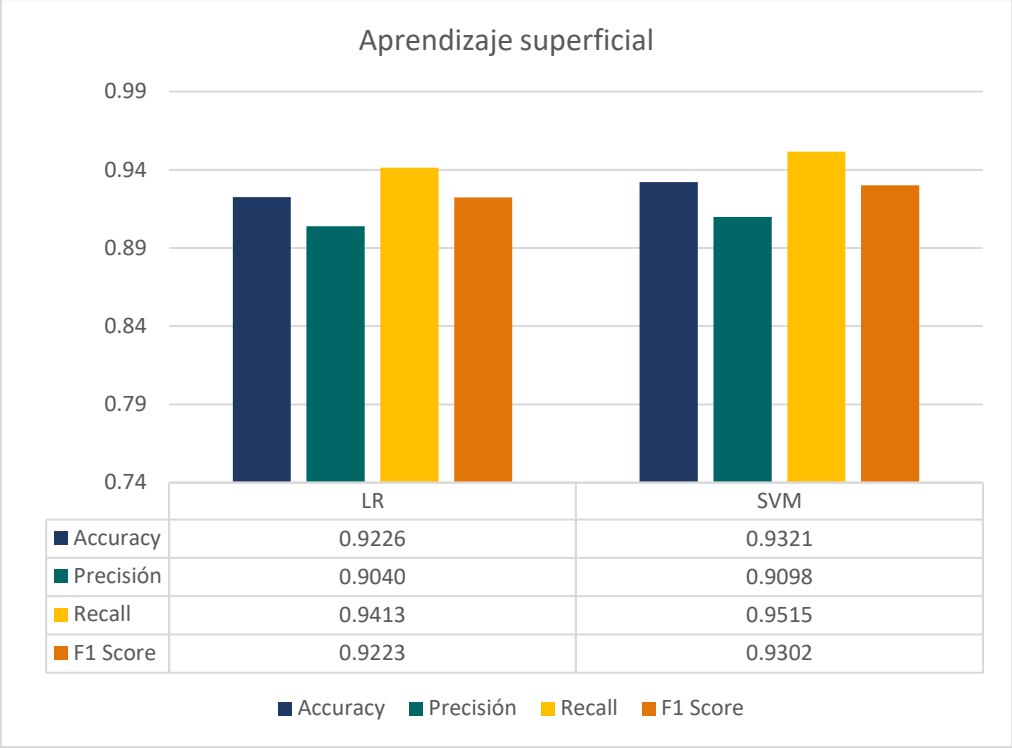
4.1 Modelos de aprendizaje computacional

Para la detección de discurso de odio, los modelos seleccionados fueron entrenados y validados con la muestra de tuits etiquetada. La construcción de estos modelos, siguió el proceso explicado a detalle en el capítulo 3, a través de la recolección de datos, el preprocesamiento de los mismos, el manejo de características, la arquitectura de los modelos y la evaluación de su rendimiento mediante las siguientes métricas: matriz de confusión, Accuracy, Precisión, Recall y F1-Score.

Inicialmente, en la Figura 4.1 y la Figura 4.2, se muestran los valores obtenidos respecto al Accuracy, Precisión, Recall y F1-Score de los modelos de aprendizaje superficial y aprendizaje profundo respectivamente. La Figura 4.1 muestra un mejor rendimiento por parte de SVM respecto a LR en las 4 métricas evaluadas, sin embargo, esta diferencia es sumamente pequeña, pues SVM supera por menos de 0.01 a LR, pues mientras SVM obtuvo un Accuracy de 0.9321 y un F1-Score de 0.9302, LR obtuvo un Accuracy de 0.9321 y un F1-Score de 0.9302.

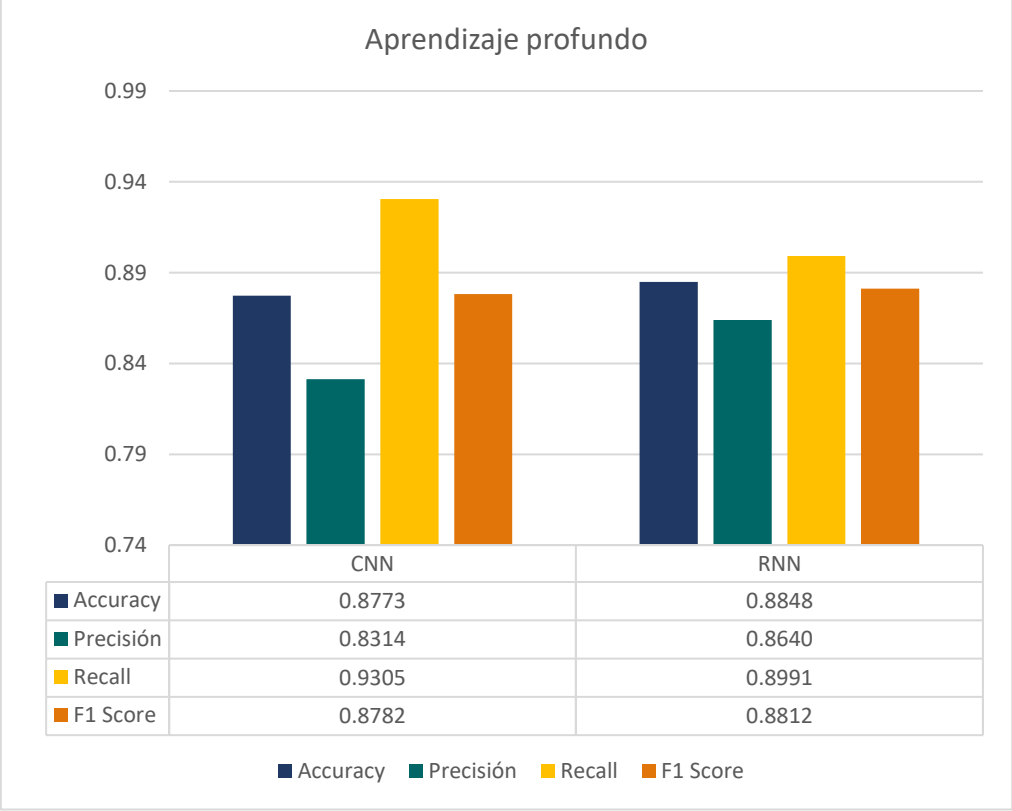
Por otra parte, la Figura 4.2, muestra un mejor desempeño para RNN en cuanto al Accuracy, Precisión y F1-Score respecto a CNN el cual obtuvo un rendimiento ligeramente inferior, sin embargo, se destaca que CNN obtuvo un mejor rendimiento en el Recall respecto a RNN. Para estos modelos, la métrica con puntaje más alto se representa en el Recall, con un valor de 0.9305 y 0.8991 para CNN y RNN respectivamente, en donde el Accuracy y el F1-Score son ligeramente superiores para RNN, mostrando la mayor diferencia en la precisión pues CNN obtuvo un valor de 0.8314, mientras que RNN obtuvo un valor de 0.8640.

De esta forma, es posible reconocer que, aunque los cuatro modelos implementados para la detección de discurso de odio hacia la comunidad LGBT en textos cortos de Twitter han tenido rendimientos considerablemente altos, fueron los modelos con arquitecturas que utilizan la frecuencia de término y frecuencia inversa de documento (TF-IDF), aquellos que obtuvieron mejor rendimiento respecto a los modelos construidos con Word2Vec.



	LR	SVM
Accuracy	0.9226	0.9321
Precisión	0.9040	0.9098
Recall	0.9413	0.9515
F1 Score	0.9223	0.9302

Figura 4.1 Métricas de evaluación para modelos de aprendizaje superficial. Elaboración propia



	CNN	RNN
Accuracy	0.8773	0.8848
Precisión	0.8314	0.8640
Recall	0.9305	0.8991
F1 Score	0.8782	0.8812

Figura 4.2 Métricas de evaluación para modelos de aprendizaje profundo. Elaboración propia

Aunado a lo anterior, la Figura 4.3 y la Figura 4.4 manifiestan los resultados obtenidos en las matrices de confusión tanto para los modelos de aprendizaje superficial como para los modelos de aprendizaje profundo. La Figura 4.3 muestra nuevamente un mejor desempeño en SVM respecto a LR, donde la cantidad de falsos positivos es bastante baja con 37 y 47, mientras que los falsos negativos son ligeramente más altos con 72 y 73 respectivamente, resaltando de igual manera una ligera discrepancia entre ambos modelos.

Para el caso de CNN y RNN, aunque la cantidad de falsos positivos y falsos negativos es más alta, en ambos modelos respecto a SVM y LR, se observa un mejor balance respecto a los datos etiquetados erróneamente para RNN, pues mientras CNN, cuenta con 144 falsos negativos y 53 falsos positivos, RNN muestra un resultado de 108 falsos negativos y 77 falsos positivos.

Con lo cual se reconoce que, aunque todos los modelos muestran mayor dificultad con los falsos negativos, el modelo basado en RNN, tiene un mayor equilibrio en la clasificación errónea de los datos, pues, aunque la clasificación de falsos positivos es mayor que en el resto de los modelos, el porcentaje de falsos negativos es menor respecto a los datos clasificados de manera errónea.

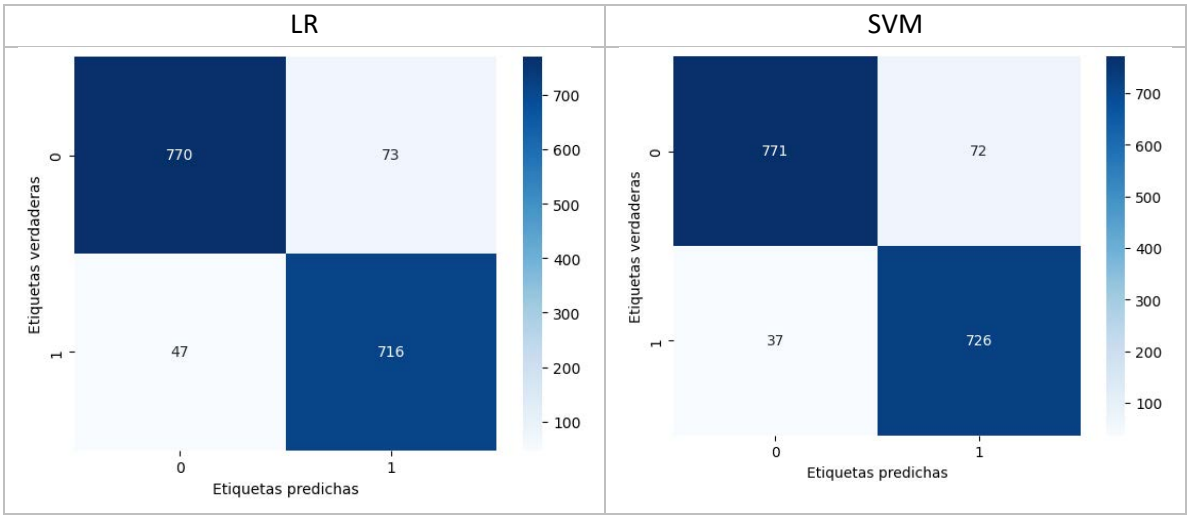


Figura 4.3 Matrices de confusión para modelos de aprendizaje superficial. Elaboración propia.

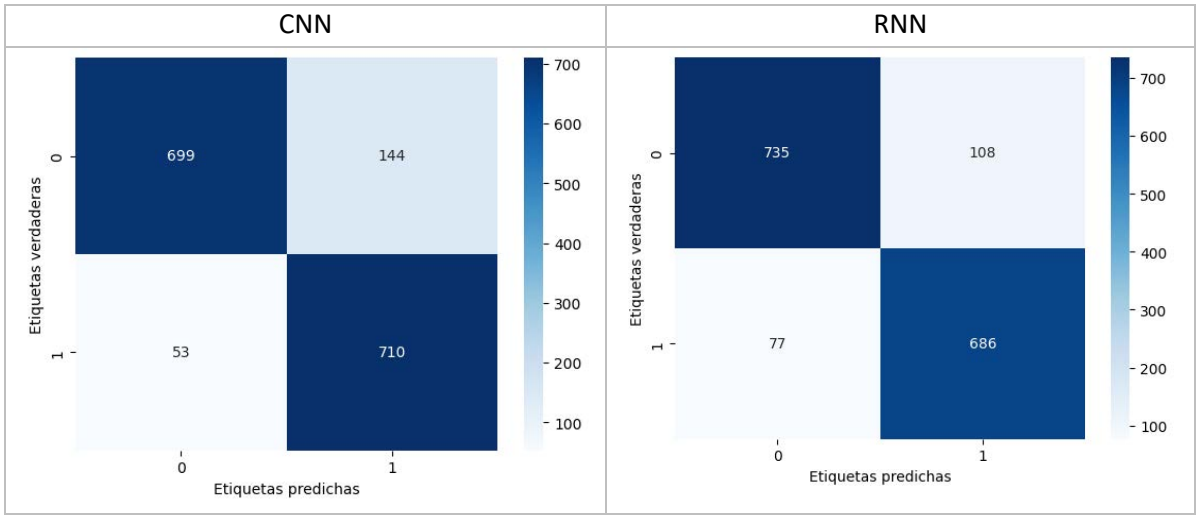


Figura 4.4 Matrices de confusión para modelos de aprendizaje profundo. Elaboración propia.

Al distinguir que las arquitecturas construidas con TF-IDF mostraron rendimientos mayores a las arquitecturas construidas bajo Word2Vec, se visualizó en LR y SVM la mejor opción para desarrollar el etiquetado del resto de tuits que no contaban con una etiqueta previa. Aun cuando SVM mostró el mejor rendimiento, el modelo seleccionado fue LR, esto fue resultado no solo del buen rendimiento del modelo, sino también, al mostrar un mejor comportamiento dé la curva de aprendizaje, entendida como el valor obtenido de una métrica dada (Accuracy en este caso), a lo largo del entrenamiento, mientras el tamaño del conjunto va en incremento.

Como se observa en la Figura 4.5 tanto para SVM como para LR, es posible distinguir que la curva de aprendizaje parte desde una puntuación alta, señalando que los modelos han logrado aprender lo suficiente con una cantidad pequeña de datos, no obstante, esta curva de aprendizaje se va regularizando a la par que se incrementa el tamaño del conjunto de datos.

Al observar la gráfica de curvas de aprendizaje de LR, es notorio el buen funcionamiento del modelo, sin embargo, la importancia de esta grafica radica al identificar que tanto el punto final de la curva de entrenamiento como de validación convergen en un punto cercano, señalando que el modelo logra generalizar de manera adecuada descartando señales de posible overfiting o sobreajuste, posibilitando la clasificación del resto de datos que carecen de etiquetas previas.

Al reconocer que tanto los datos de entrenamiento como de validación tienen un porcentaje alto de aprendizaje desde el primer lote del conjunto de datos, y donde no se observan fluctuaciones o decrecimientos en las curvas de aprendizaje, se reconoce que el modelo ha alcanzado su máximo funcionamiento y la implementación de una mayor cantidad de datos no representaría una ayuda significativa en el rendimiento del mismo.

Para lograr este buen resultado y evitar la presencia de sobreajuste, se implementó en el modelo de LR una regularización L1 (LASSO) con un optimizador “saga”, con el fin de ayudar al modelo con el overfiting que pudiera presentarse antes de la implementación de estos parámetros.

Por su parte, aunque las curvas de aprendizaje del modelo basado en SVM, mostraron un comportamiento similar al de LR, la separación entre ambas curvas de aprendizaje, aunque es poco pronunciada, es evidente que existe una separación entre ambas. Dicha separación se intentó ajustar con diferentes

métricas de regularización como en el caso del modelo de LR, con un parámetro L1 y diferentes optimizadores, sin embargo, al implementar dichas métricas el rendimiento y funcionamiento del modelo se mostraba afectado, por lo que se decidió prescindir de estas.

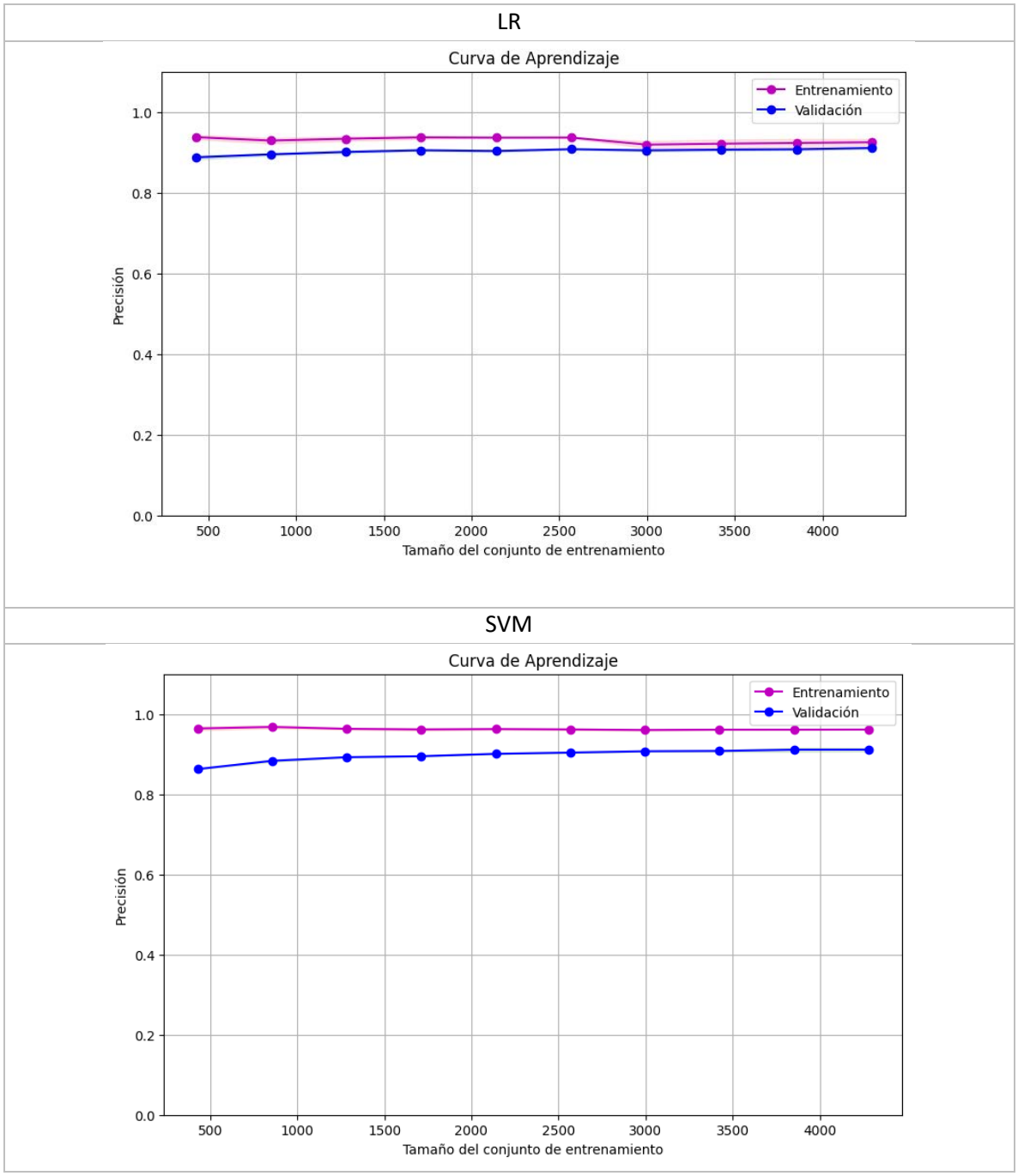


Figura 4.5 Comparación de la curva de aprendizaje en LR y SVM

Como se mencionó anteriormente y observando la Figura 4.2, es posible distinguir que el rendimiento de los modelos basados en redes neuronales (CNN y RNN), son más que satisfactorios, sin embargo, es importante señalar que, en dichas arquitecturas, pueden observarse áreas de mejora en la detección de discurso de odio hacia la comunidad LGBT.

Inicialmente dentro de la arquitectura del modelo, es interesante observar que para este caso en el procesamiento de textos cortos enfocados en la clasificación binaria de discurso de odio hacia un grupo poblacional específico, la utilización de arquitecturas basadas en TF-IDF muestran un mejor desempeño respecto a arquitecturas más complejas y la utilización de Word2Vec, este factor puede deberse a la naturaleza de los datos utilizados y las características de los textos generados en Twitter, donde se concibe la presencia de un léxico y gramática más ligera o simple, resaltando la importancia de las redes neuronales en trabajos que mantienen un punto focal en el análisis gramatical de los textos, en cuyo caso arquitecturas basadas en Word2Vec pueden presentar mayor optimización al considerar este factor.

Aunado a ello, se destaca la necesidad de implementar grandes volúmenes de datos en el corpus de entrenamiento y validación, esto se puede distinguir en algunos de los trabajos mencionados en el capítulo 2, donde se llegan a utilizar corpus que superan los 20,000 registros, en muchos casos tomando bases que han sido construidas y etiquetadas previamente; estos trabajos que suelen mantener enfoques centrados en el mejoramiento de las arquitecturas de los modelos con la implementación de diversas técnicas sensibles al contexto de los textos o algunas otras herramientas como las estructuras basadas en transformadores, entre otras.

Al considerar que esta investigación ha buscado desarrollar un acercamiento a la identificación de discurso de odio hacia la comunidad LGBT de manera binaria y que la cantidad de datos que pudieron ser clasificados y validados por las y los etiquetadores no alcanzó a ser tan alta, se reconocen dos hechos sustanciales: inicialmente, que las arquitecturas de las redes neuronales podrían ser complejas para el enfoque binario que se aborda especialmente al implementar herramientas sensibles al contexto como Word2Vec, en donde enfoques menos complejos pueden desarrollar un mejor manejo de los datos, y por otra parte, la necesidad de contar con una mayor cantidad de registros o datos de entrada para el mejoramiento del modelo.

No obstante, en la implementación de los modelos basados en redes neuronales, se observó que los resultados obtenidos al compilar e implementar el modelo, realmente eran buenos, sin embargo, el problema existente con dichos modelos fue el sobreajuste, el cual fue relativamente alto en un inicio al trabajar con las primeras fases de las estructuras. Este factor del sobreajuste representa un problema considerable al momento de querer clasificar nuevos datos que el modelo no conoce, ya que, al estar sobre ajustado, el modelo está aprendiendo patrones y no generaliza lo suficiente para poder predecir nuevos datos.

En función a ello, se hicieron los ajustes pertinentes en la estructura de ambos modelos buscando el mejor rendimiento de ambos y a la vez reduciendo el sobreajuste, teniendo en mente la necesidad de clasificar nuevos datos. En este sentido, en la fase de re ajuste de parámetros, se implementaron capas de abandono (Dropout) y una regularización L1 para ambos modelos, obteniendo un buen rendimiento y reduciendo el sobreajuste.

Para identificar el funcionamiento de los modelos respecto al sobreajuste, se compararon las perdidas o (losses) para el entrenamiento y la validación (véase Figura 4.6). En este sentido, es posible distinguir que, aunque el rendimiento observado en las métricas de evaluación en el modelo RNN, son mejores respecto a CNN, hay pequeñas fluctuaciones a lo largo de las épocas tanto para el entrenamiento como para la validación, además, próximo a las 40 épocas, las curvas de entrenamiento y validación se comienzan a separar ligeramente, identificando que tal comportamiento de las pérdidas en RNN, puede representar un

conjunto de datos con falta de representatividad, es decir, que el modelo puede requerir la implementación de una mayor cantidad de datos tanto para el entrenamiento como para la validación.

No obstante, aunque el modelo CNN, mostró un rendimiento menor respecto a RNN, el ajuste de las curvas de pérdidas en el entrenamiento y validación reflejan un buen rendimiento, la curva de entrenamiento va decreciendo conforme las épocas incrementan y la línea de pérdidas de validación alcanza un equilibrio a la línea de entrenamiento a partir de las 20 épocas, representando un buen ajuste en las curvas de pérdidas. En este sentido, aunque el rendimiento de RNN es mayor respecto a CNN, el ruido observado en las curvas de pérdidas muestra una mayor probabilidad de estar frente a posibles problemas de sobreajuste, mientras que CNN, obtiene un buen ajuste en las curvas de pérdidas reflejando una mejor generalización de los datos con los que se construyó el modelo.

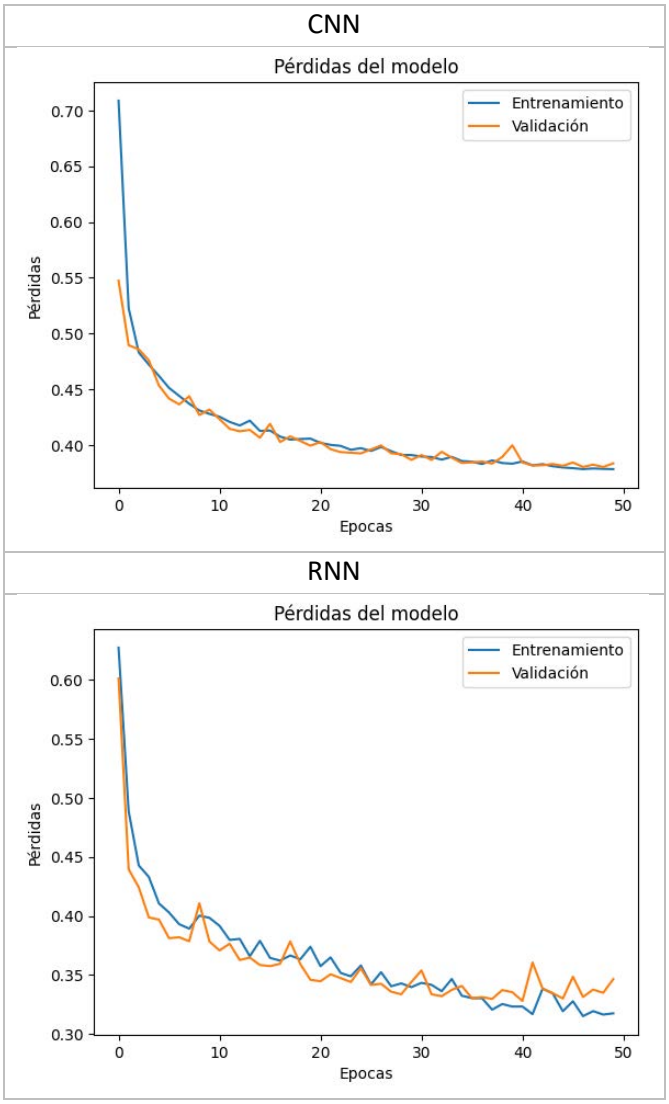


Figura 4.6 Comparación de valores de pérdida (losses) en CNN y RNN

A lo largo del presente capítulo se han analizado particularidades de cada uno de los modelos utilizados, sin embargo, como se mencionó en un inicio para el problema de clasificación de discurso de odio hacia la comunidad LGBT en Twitter, el modelo que implementa Regresión Logística, ha sido el seleccionado como el modelo con mejores rendimientos y buena generalización de los datos permitiendo el etiquetado de datos que el algoritmo desconoce.

No obstante, aunque el modelo seleccionado muestra un rendimiento más que adecuado, es imprescindible no perder de vista que hay datos erróneos que el algoritmo no es capaz de etiquetar de manera adecuada. Para reconocer los problemas que el modelo tuvo durante la clasificación del conjunto de datos de validación, se realizó una comparación de los tuits etiquetados y la clasificación generada por el algoritmo.

Como se muestra en la Figura 4.3, la matriz de confusión en el conjunto de validación, presenta un total de 120 tuits etiquetados de manera errónea, con 73 y 47 falsos negativos y falsos positivos respectivamente. Dentro de los falsos negativos, se identificaron algunos patrones en la mala clasificación de dichos textos; inicialmente se observó un problema en la clasificación de tuits que hacen referencia al coronavirus con palabras que suelen ser utilizadas para referirse de manera despectiva a la comunidad LGBT:

- Extraño ir a los bares, al boliche, al antro, al cine. Los odios a todos, te odio puto coronavirus.
- Cabrón que viernes tan triste aqui cdmx pinche covid putito.

También se identifican algunos mensajes que hacen uso de la palabra *puto,* como una forma de enaltecer a una persona o bien el comportamiento de una persona, así como para hacer referencia a una situación de desagrado o inconformidad, aunque la palabra no es utilizada para referirse a una persona, por ejemplo:

- Que puto ASCOO meter sexo en tu argumento, si tienes libre expresión, pero para mí tu opinión PENDEJA, no cuenta.
- Se jactan de ser bien “chambiadores” y los veo todo el putito tiempo en TW.

Aunque estos son solo algunos ejemplos de los falsos negativos, han sido los patrones más notorios que fueron identificados para esta errónea clasificación.

En este sentido, es imperativo recordar que el algoritmo fue entrenado con el 80% de los datos etiquetados, por lo que ciertas expresiones pueden ser complicadas de identificar para el modelo de manera adecuada. Respecto a los textos que hacen utilización de la palabra *puto* sin hacer referencia a la comunidad LGBT o a una persona de manera negativa, se reconoce la necesidad de incrementar la cantidad de datos con las que el modelo es entrenado, puesto que ciertas expresiones identificadas como no discurso de odio durante la fase de etiquetado relacionadas a este tipo de expresiones fueron clasificadas de manera errónea.

Ejemplo de esto es la capacidad del algoritmo para distinguir que un tuit no es discurso de odio cuando se utiliza la expresión "eres un puto crack" pero no es capaz de distinguir que la expresión "eres un puto Dios" tampoco es reconocida como discurso de odio. Sin embargo, en la base de datos se ha identificado que la presencia de la expresión "eres un puto Dios" es escasa, mientras que "eres un puto crack" presenta mayor frecuencia, registrando que el algoritmo necesita de una base de datos de entrada más extensa impulsando la distinción de este tipo de expresiones.

Para el caso de los falsos positivos (FP), la cantidad de datos etiquetados erróneamente es menor, sin embargo, también se identificaron ciertas características dentro de estos textos clasificados de manera incorrecta, en donde el principal factor se visualiza sobre ciertos textos que expresan discurso de odio de manera no evidente, ejemplo de esto se muestra a continuación:

- Y tu chupandoselo como buen americanista! Levantando en todo momento tu orgullo Gay!!!
- Que cosas En los 70' solo eran la comunidad gay🏳‍🌈 En los 80's LG EN LOS 90'S LGB En el 2000 LGBT 2020 LGBTQI 2100 LGBTQIPHCUJFWNVXZ Y por ahí del 3000 ya habrán terminado con el abecedario y empezaran a agregarle números 😅😅😅.

Este tipo de mensajes pueden ser complejos y difíciles de identificar para el algoritmo, generando una clasificación errónea de este tipo de textos, sin embargo, es importante mencionar que la cantidad de falsos positivos es bastante baja, por lo que el procesamiento y desarrollo del modelo muestran un buen resultado en tal cometido, reconociendo que hay mayor complicación en la diferenciación de los falsos negativos.

No obstante, como se mencionaba anteriormente, la obtención de una clasificación sin errores, es en suma difícil de alcanzar, además que el riesgo de construir un modelo que sufra de sobreajuste se aumenta, impidiendo la posibilidad de clasificar datos desconocidos de manera adecuada.

El análisis del funcionamiento del modelo ha permitido reconocer la posibilidad de su implementación en la clasificación del resto de datos de la base de datos original de 22,605 tuits, excluyendo los datos utilizados para el entrenamiento y validación del modelo, teniendo un total de 15,171 tuits que han sido clasificados con el modelo LR.

En este sentido, es importante mencionar que en cuanto a la distribución y concentración de tuits que hacen referencia a discurso de odio hacia la comunidad LGBT, los resultados obtenidos son una aproximación a los datos reales, en primera instancia por el rendimiento del modelo, que aunque es bastante bueno, existe un porcentaje bajo de datos que pueden ser clasificados de manera errónea, y por otro lado, el hecho que la utilización de tuits georreferenciados deja atrás cierto porcentaje de tuits que carecen de este dato pero que se encuentran en la red.

Sin embargo, como se ha mencionado a lo largo de la investigación esta aproximación a la identificación del discurso de odio hacia la comunidad LGBT, ha pretendido acercarse a tal hecho e identificar espacios prioritarios o de mayor concentración de este tipo de mensajes, fungiendo como parteaguas a la implementación de modelos más complejos y mayor cantidad de datos.

Con la totalidad de datos clasificados, se distinguieron ciertas palabras con mayor prevalencia en los tuits que expresan discurso de odio hacia la comunidad LGBT, identificando que aunque en muchas ocasiones estas palabras no son utilizadas únicamente para referirse a un grupo de personas, sino a todo el colectivo en general sin hacer una distinción adecuada a cada uno de los colectivos que conforman a la población LGBT, las palabras que presentan mayor prevalencia, son aquellas que coloquialmente son utilizadas para hacer referencia de manera negativa o despectiva a la comunidad homosexual (véase, Figura 4.7).

Figura 4.7 Nube de palabras en discurso de odio identificado

La implementación de técnicas de aprendizaje computacional ha permitido generar una clasificación eficiente de la base de datos, impulsando un análisis temporal no sólo de los tuits que expresan discurso de odio (DO), sino también, de los textos que hacen referencia a la comunidad LGBT de manera adecuada.

Como se puede observar en la Figura 4.8, la presencia de tuits referentes a la comunidad LGBT, así como los tuits de discurso de odio, muestran una distribución relativamente análoga a lo largo de la temporalidad de estudio (2020, 2021 y 2022). La cantidad de tuits referentes a DO, muestra una constante aparente a lo largo de los años, sin embargo, es posible identificar ciertas fluctuaciones en cuanto a la cantidad de tuits que hacen referencia a la comunidad LGBT, pero no expresan discurso de odio, lo cual se observa claramente durante el mes de junio, reconocido como el mes del orgullo LGBT.

Sin embargo, aunque durante el 2021 y 2022, la cantidad de tuits y tuits de DO, mantienen un comportamiento muy similar, con un notorio incremento durante el mes de junio y mayo, es posible identificar que el comportamiento durante el 2020, muestra una distribución más errática. En este año, aunque la cantidad de tuits de DO parecen mantener una constante en el primer trimestre del año, se observa un declive abrupto durante el mes de abril y mayo, no sólo del DO sino de tuits en general, además se observa un incremento nuevamente abrupto durante el mes de junio y julio tanto de tuits referidos a comunidad LGBT como tuits con DO, mientras que, en el último trimestre del año este comportamiento se regulariza con el resto de los años estudiados.

No obstante, es imperativo recordar que, siguiendo la metodología planteada, se consideró como preámbulo esencial el inicio de la pandemia de COVID-19 en México durante el primer trimestre del 2020, hecho bajo el cual, las dinámicas sociales presenciaron modificaciones de diferentes maneras. Este hecho pudo influir en el decremento de tuits referentes a la comunidad LGBT, ya que la mente colectiva se encontraba centrada en las implicaciones de la pandemia, por su parte, para el mes de junio de este año, se reconoce la primera marcha celebrada de manera virtual por el orgullo LGBT, hecho que pudo influir en el incremento abrupto de menciones de discurso de odio y de tuits referentes a la comunidad LGBT en general.

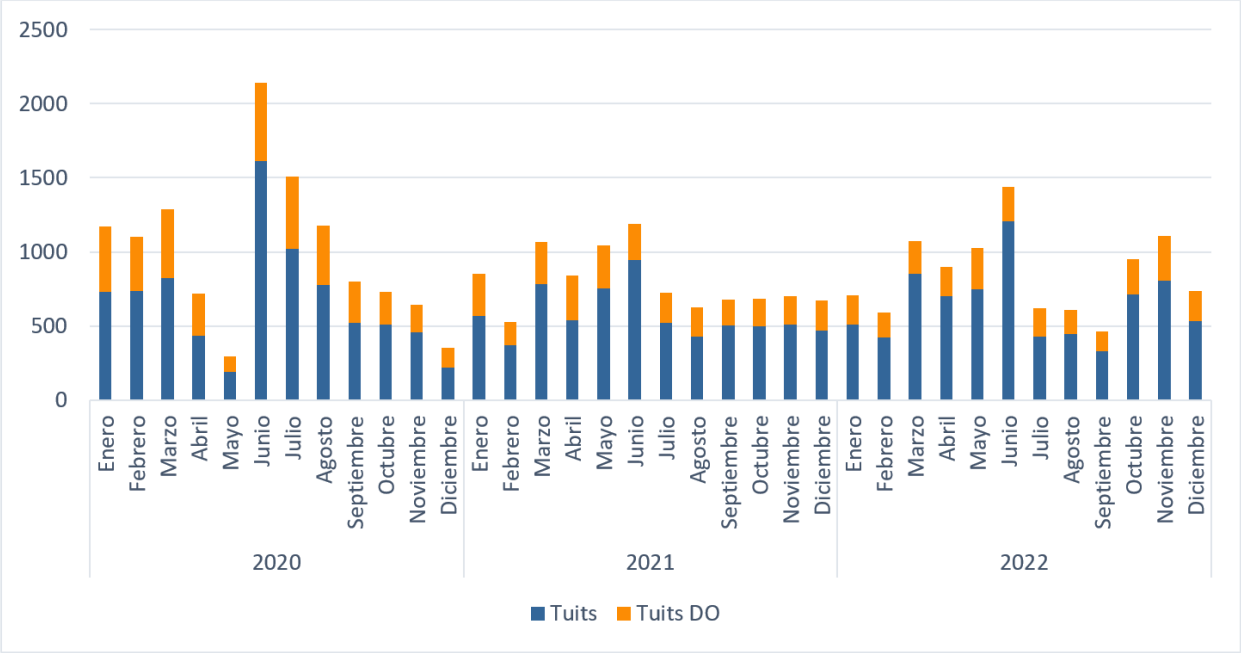


Figura 4.8 Distribución temporal de tuits referentes a la comunidad LGBT y tuits en los que se expresa discurso de odio (DO) hacia la comunidad LGBT

Para poder comparar el comportamiento del discurso de odio hacia la comunidad LGBT respecto a la cantidad de tuits que hacen mención a esta misma sin expresarse desde el odio, la Figura 4.9 muestra la distribución temporal de los tuits que hacen mención o referencia a la comunidad LGBT y los tuits de DO a esta misma en porcentajes.

De esta forma, lo que se ha buscado es observar el comportamiento de la presencia del DO hacia la comunidad LGBT en proporción a la cantidad de mensajes que hacen referencia o mención de la misma comunidad de manera positiva. Como se observa en la figura siguiente, en el primer semestre del 2020, el porcentaje de tuits de DO es mayor en comparación al resto de los años. De nueva cuenta, se identifica en los 3 años que la cantidad de tuits de DO es menor en comparación a los tuits que no se expresan desde el DO en el mes de junio, sin embargo, posterior a este, en el mes de julio, el porcentaje de tuits referidos desde el DO aumente nuevamente, mostrando cierta regularización en el porcentaje de la presencia de este tipo de mensajes.

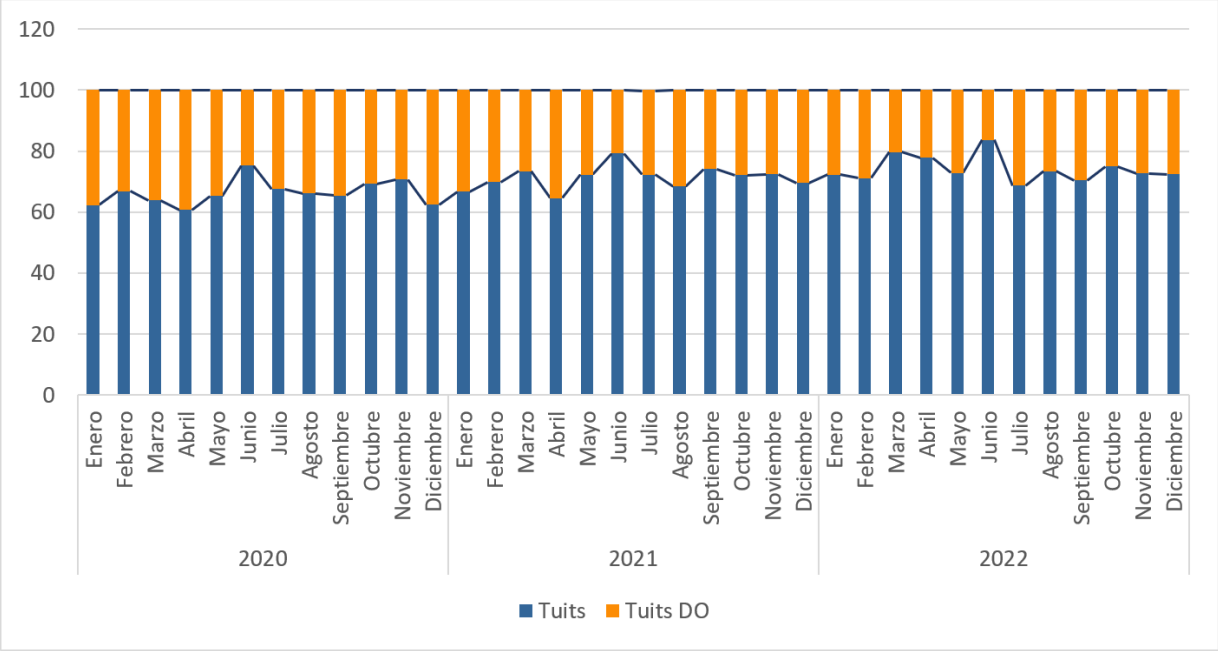


Figura 4.9 Distribución temporal en porcentaje de tuits referentes a la comunidad LGBT y tuits en los que se expresa discurso de odio (DO) hacia la comunidad LGBT

Lo anteriormente expuesto se observa de manera más clara en la Figura 4.10, donde es posible observar mayor presencia de tuits de discurso de odio, respecto a los tuits que no fueron catalogados como DO en el 2020, decreciendo ligeramente en los años posteriores, en donde además se observa un decrecimiento abrupto del porcentaje de tuits de odio en el mes de junio, pero un incremento en el mes de julio, reconociendo el mes de diciembre también como uno de los meses con mayor porcentaje de tuits que se expresan desde el discurso de odio.

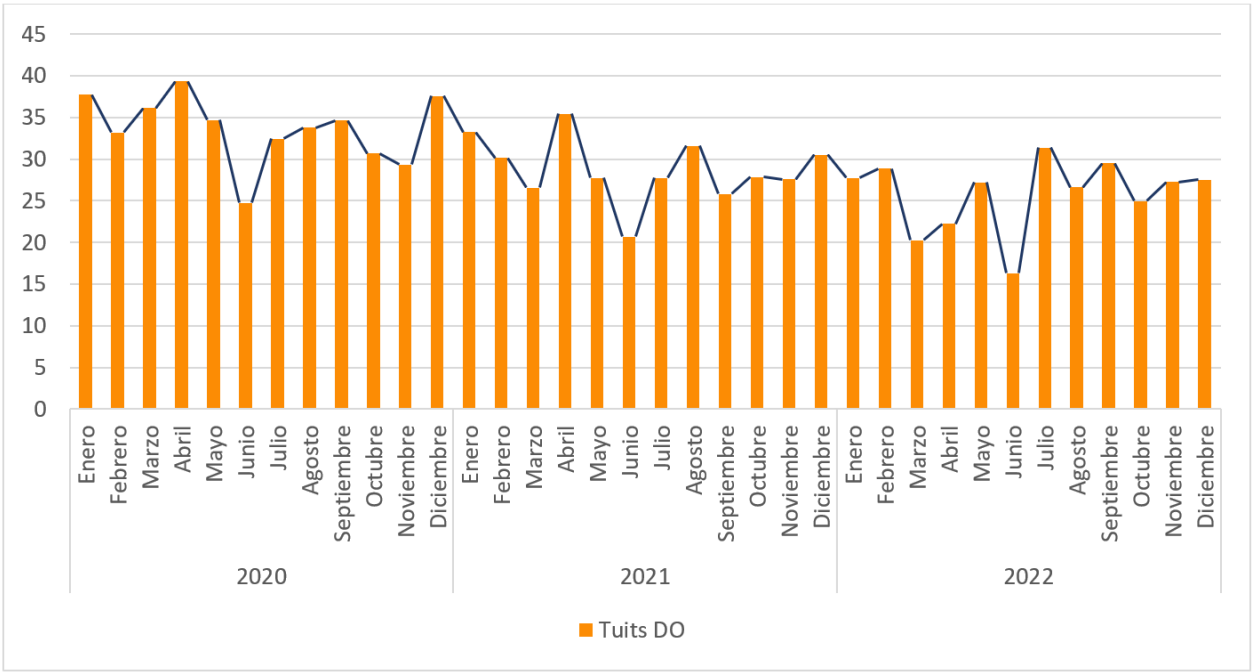


Figura 4.10 Distribución temporal en porcentaje de tuits en los que se expresa discurso de odio (DO) hacia la comunidad LGBT

Estas observaciones, resaltan la prevalencia de discurso de odio dirigido a la comunidad LGBT a lo largo de la temporalidad; si bien se visualiza una ligera disminución para el 2021 y 2022 respecto al discurso de odio observado en el 2020, es importante señalar que la presencia de este tipo de mensajes no se ve afectada de manera absoluta con el incremento de tuits que hacen referencia a la comunidad LGBT, pero no expresan discurso de odio.

4.2 Espacializar la información

La utilización de algoritmos enfocados en el aprendizaje computacional, permiten en este caso la identificación de discurso de odio hacia la comunidad LGBT de manera eficiente a través de la automatización, sin embargo, este tipo de tareas suelen pasar de manera desapercibida la importancia de visualizar la distribución espacial de este tipo de mensajes.

Reconociendo la implementación y utilización de tuits georreferenciados, ha sido posible realizar una distinción tanto temporal como espacial para generar una zonificación del discurso de odio hacia la comunidad LGBT en Twitter en el territorio mexicano.

Para tal cometido fue necesario realizar una segmentación temporal que facilitara la visualización de la distribución de este tipo de mensajes a lo largo de la temporalidad estudiada (2020, 2021 y 2022). La Figura 4.11 muestra dicha zonificación a través de los estados como unidad espacial, identificando que, aunque existen variaciones entre los años de estudio respecto a la cantidad de tuits clasificados como discurso de odio, se perciben ciertos patrones a lo largo de la temporalidad estudiada.

Es importante mencionar que esta distribución se efectuó a través de una generalización de los rangos de clasificación obtenidos de manera individual para cada año de estudio, obtenido por medio de rupturas naturales, sin embargo, como la cantidad de tuits es disímil en los años de estudio, se generó un promedio de los rangos obtenidos y se clasificó nuevamente la presencia de tuits con base en esta generalización, permitiendo una comparación más clara especialmente en las representaciones cartográficas.

Así, se identificaron aquellos estados en los que hay mayor presencia de tuits de discurso de odio hacia la comunidad LGBT a lo largo de los años analizados (véase Tabla 4.1). En la tabla siguiente se muestran de manera explícita la cantidad de tuits clasificados como discurso de odio para los 10 estados con los valores más altos, en donde, aunque se observan fluctuaciones entre las posiciones de los estados respecto a la cantidad de mensajes, también es posible distinguir que, de manera continua los estados con mayor presencia de tuits de discurso de odio hacia la comunidad LGBT son: Ciudad de México, Estado de México, Nuevo León y Jalisco respectivamente.

Estado	Años		
	2020	2021	2022
Ciudad de México	823	642	587
Estado de México	497	338	303
Nuevo León	386	304	210
Jalisco	320	215	211
Quintana Roo	189	78	105
San Luis Potosí	149	103	117
Puebla	155	98	83
Veracruz	156	90	83
Querétaro	129	70	91
Tamaulipas	99	99	85

Tabla 4.1 Distribución de tuits identificados como discurso de odio por estados

Es durante el año 2020, donde se identificó la mayor cantidad de discurso de odio hacia la comunidad LGBT, sin embargo, también es posible reconocer que para este año la presencia de este tipo de mensajes muestra una gran concentración en la Ciudad de México, registrando una diferencia cercana a los 200 y 300 tuits para los años posteriores respectivamente, mientras que dicha diferencia es menor para el resto de los estados en comparación a los registros obtenidos para el 2021 y 2022.

Considerando la totalidad de registros obtenidos con la clasificación de tuits de discurso de odio hacia la comunidad LGBT, la Figura 4.12 engloba los 3 años de estudio, reconociendo a los siguientes estados como aquellos en los que existe mayor presencia de tuits de discurso de odio hacia la comunidad LGBT: Ciudad de México, Estado de México, Nuevo León, Jalisco y Quintana Roo, en donde se identificó una cantidad de registros entre 371 y 1140, seguidos por San Luis Potosí, Puebla, Veracruz, Querétaro, Tamaulipas y Guanajuato con un intervalo de 221 a 370.

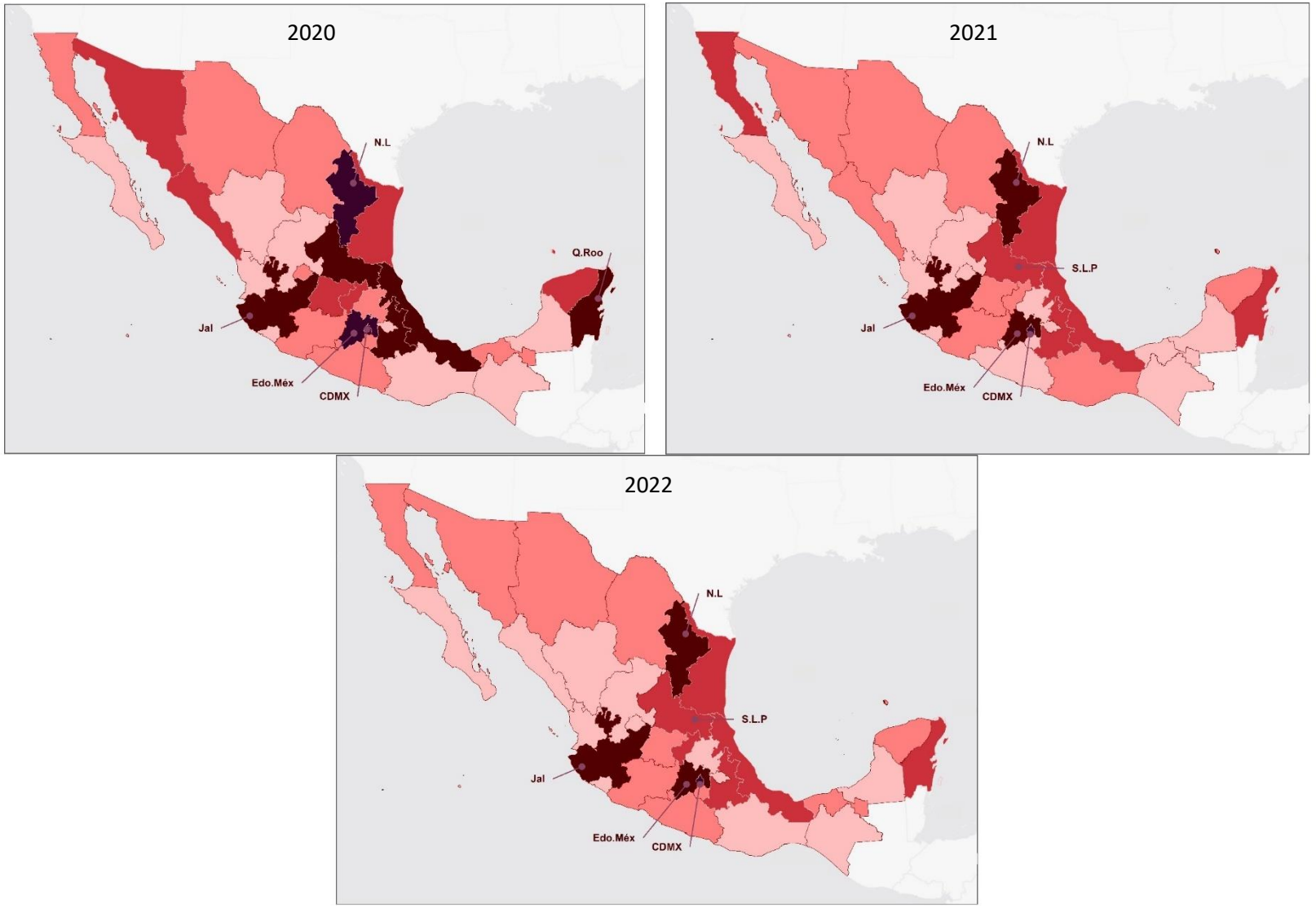


Figura 4.11 Distribución de tuits identificados como discurso de odio hacia la comunidad LGBT en Twitter

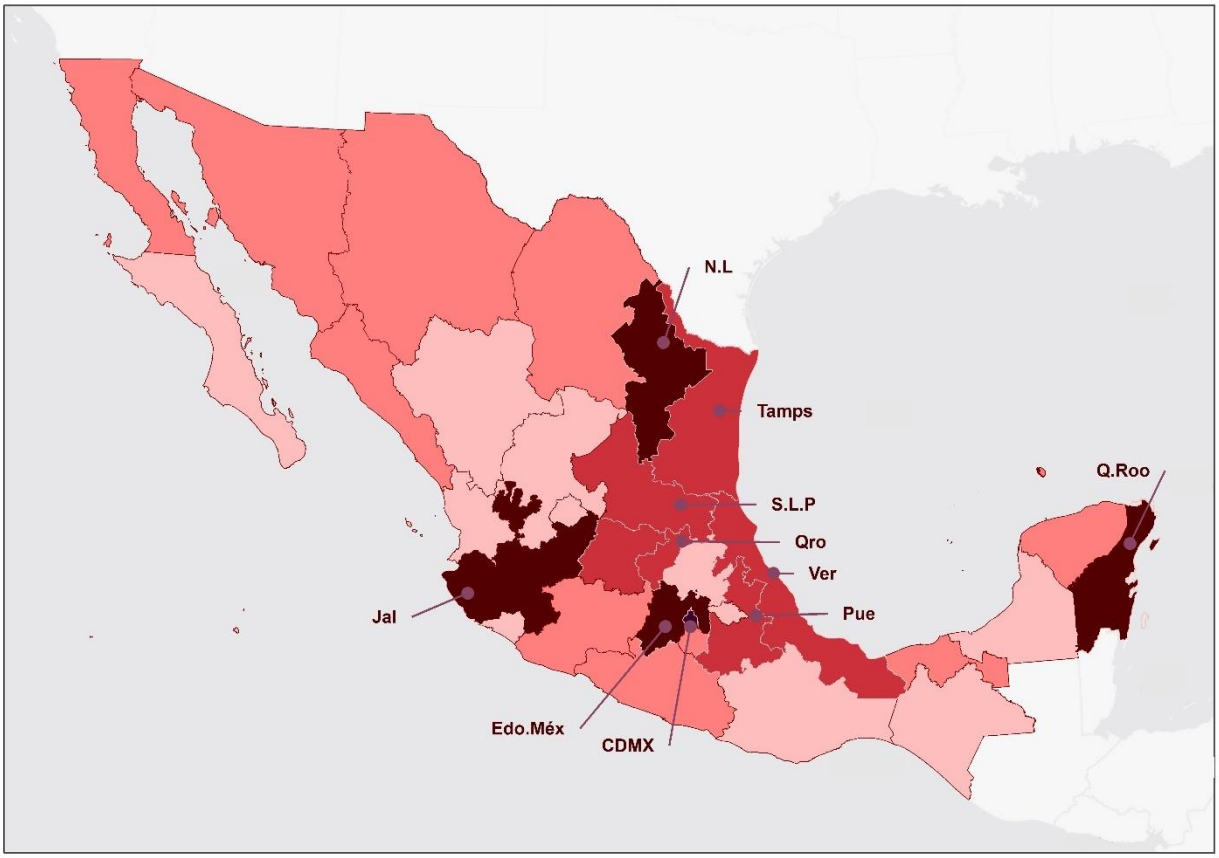


Figura 4.12 Distribución de tuits identificados como discurso de odio hacia la comunidad LGBT en Twitter del 2020 al 2022

Como la mayor concentración de este tipo de mensajes se encuentra dentro de la Ciudad de México y el Estado de México, han sido las entidades en las cuales se ha realizado un esfuerzo mayor por identificar de manera mas local las zonas en donde exista mayor concentración de este tipo de mensajes.

La Figura 4.13 muestra la densidad de registros para estas dos entidades de interés, para lo cual se implementó un proceso de densidad de Kernel, que aunque puede generalizar el espacio de concentración al no contar con zonas en donde existan datos, es posible identificar puntos de calor en donde dicha concentración de mensajes es mayor respecto al resto del territorio.

Para la Ciudad de México, se reconoció a la alcaldía Cuauhtémoc como la zona en donde la concentración de este tipo de mensajes es mayor, seguida por Benito Juérez y Coyoacán; mientras que para el Estado de México su correspondiente es el municipio de Tlanepantla de Baez.

Estado de México

Ciudad de México

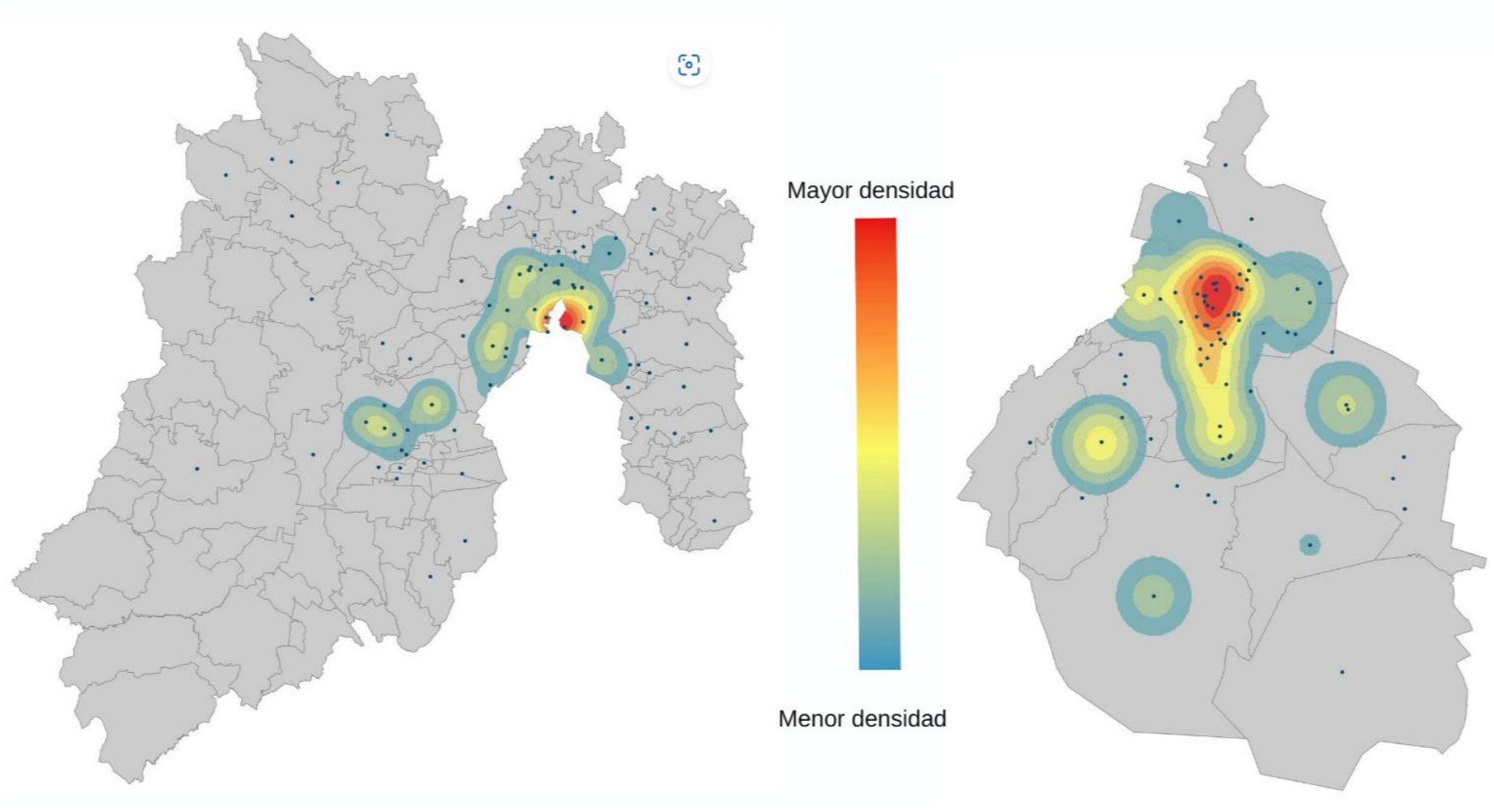


Figura 4.13 Densidad de tuits identificados como discurso de odio hacia la comunidad LGBT en Twitter del 2020 al 2022

No obstante, es imperativo mencionar que los datos mostrados anteriormente representan la totalidad de tuits clasificados como discurso de odio de manera absoluta, sin embargo, como se mencionó en el capítulo 1, está distribución puede estar influenciada por ciertos factores inherentes a las redes sociales como la necesidad de contar con un dispositivo inteligente con acceso a internet, teniendo como resultado una mayor concentración en las ciudades más grandes del país, muy probablemente influido por la mayor concentración de personas en tales espacios.

Por tal motivo, y en busca de los estados donde la presencia de este tipo de mensajes es considerablemente alta en relación a la totalidad de tuits publicados donde se hace referencia a la comunidad LGBT, se realizó un cálculo de tasas[4] por estado. La Tabla 4.2 muestra los 10 estados con los valores más altos respecto a la tasa de tuits clasificados como discurso de odio por año, en donde, sobresalen algunos estados como: Tamaulipas, San Luis Potosí, Quintana Roo, Guerrero o Estado de México (para mayor referencia, véase Anexo 5).

Estado	Años		
	2020	2021	2022
Tamaulipas	70	69	71
San Luis Potosí	71	61	61
Quintana Roo	71	43	53
Guerrero	76	52	46
Estado de México	65	49	49
Tlaxcala	63	52	30
Nuevo León	49	47	40
Hidalgo	55	48	25
Campeche	41	56	27
Colima	54	14	49

Tabla 4.2 Estados como mayor tasa de discurso de odio hacia la comunidad LGBT

Además, se realizó el mismo procedimiento para obtener la cartografía pertinente tomando como datos de entrada los valores calculados en tasas (véase Figura 4.14). En ello, se observa un comportamiento similar en los 3 años de estudio, especialmente en aquellos estados donde la tasa es más alta, como el estado de Tamaulipas y San Luis Potosí, los cuales estuvieron en el rango más alto para los 3 años de estudio, seguidos de Guerrero, Estado de México y Quintana Roo que estuvieron presentes en el rango más altos en dos de los 3 años abordados.

[4] Calculada a través de la cantidad de tuits clasificados como discurso de odio entre la cantidad total de tuits sin distinción por año, multiplicado por 100, con el objetivo de identificar la cantidad de tuits expresados desde el discurso de odio por cada 100 tuits publicados en los que se hace referencia a la comunidad LGBT.

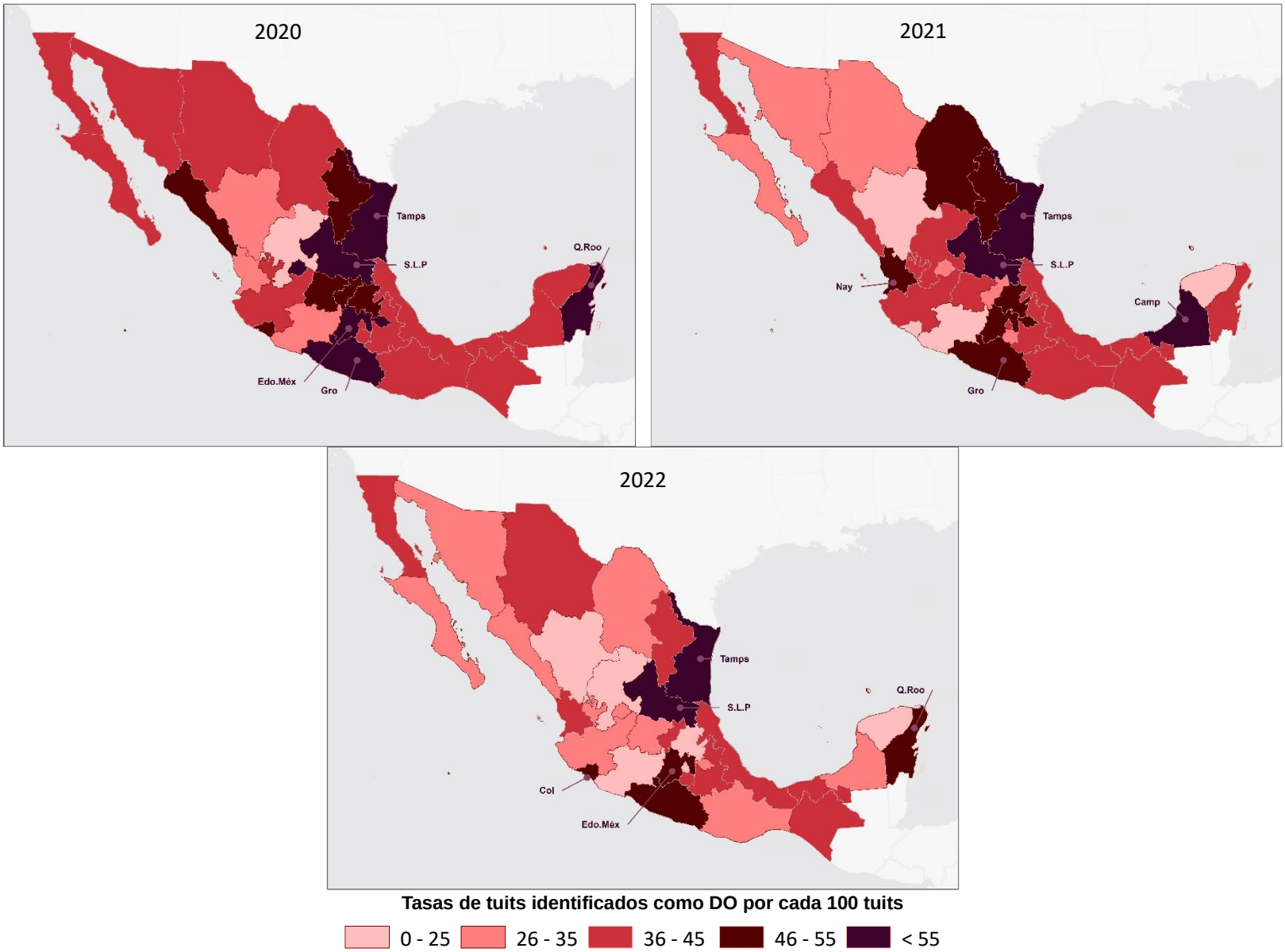


Figura 4.14 Distribución de tuits identificados como discurso de odio hacia la comunidad LGBT por tasas

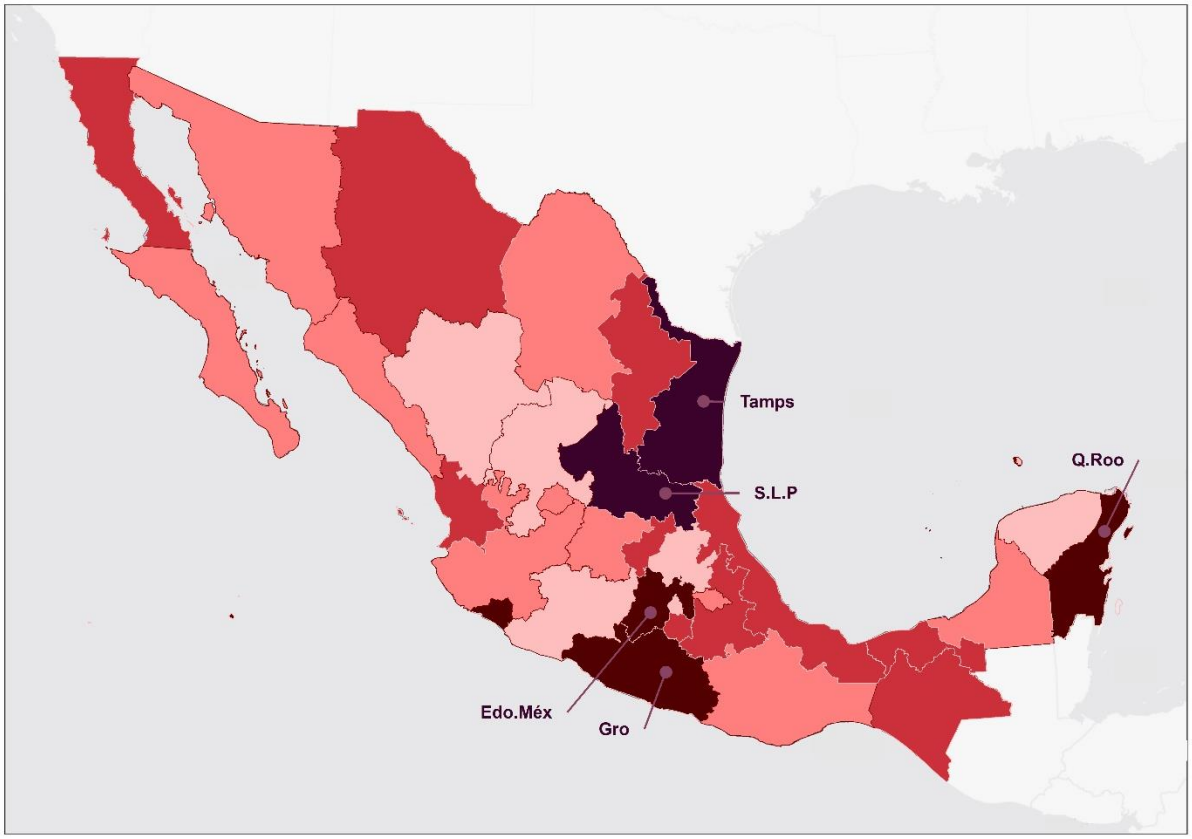


Figura 4.15 Distribución de tuits identificados como discurso de odio hacia la comunidad LGBT por tasas del 2020 al 2022

Además, se calculó la tasa correspondiente al total de tuits para cada estado en la temporalidad de estudio, (véase Figura 4.15), observando que el estado con la tasa más alta es Tamaulipas, seguido de San Luís Potosí, Guerrero y Estado de México respectivamente.

De esta manera, se reconoce la importancia de ciertos estados respecto a la presencia de discurso de odio hacia la comunidad LGBT en textos cortos de Twitter, pues aun cuando en el primer compendio cartográfico sobresalían algunos estados como la Ciudad de México, el cálculo de tasas permite reconocer una alta prevalencia de este tipo de mensajes sobre la totalidad de mensajes referentes a la comunidad LGBT.

El cálculo de las tasas, permite reconocer espacios de importancia como Tamaulipas o San Luis Potosí al considerar que aproximadamente de cada 100 tuits publicados en los que se hace mención a la comunidad LGBT o hacia algún miembro de esta, 70 de ellos se expresan desde un discurso de odio.

5 Conclusiones y trabajo futuro

5.1 Conclusiones

En el presente trabajo se ha desarrollado una exploración de 4 diferentes algoritmos de aprendizaje computacional, 2 de ellos basados en aprendizaje profundo y otros dos basados en aprendizaje superficial; esta implementación de algoritmos ha buscado identificar los modelos con mejor rendimiento en la tarea de clasificar de manera binaria discurso de odio hacia la comunidad LGBT en textos cortos, obtenidos de Twitter en una demarcación acotada al territorio mexicano.

Reconociendo que la tarea de detectar discurso de odio en redes sociales puede llegar a ser un proceso tardado y costoso si es efectuado de manera manual, la implementación del aprendizaje computacional representa una opción de realizar tal tarea de forma eficiente y eficaz, sin embargo, para poder obtener resultados adecuados, tanto la recolección de información, como el procesamiento de los mismos y la construcción del modelo fungen como pasos vitales, donde se debe ser rigurosamente cuidadoso con los procesos desarrollados con la finalidad de obtener los mejores resultados posibles.

Aun cuando en el campo del aprendizaje computacional, la implementación de redes neuronales ha sobresalido en los últimos años como una opción clave en la detección de discurso de odio, al observar que los resultados obtenidos son más que adecuados, algunas técnicas más tradicionales como Regresión Logística, Máquinas de Soporte Vectorial o árboles de decisión siguen siendo utilizadas al mostrar resultados eficientes y eficaces.

Sin embargo, la construcción de los modelos puede ser llevada de diversas maneras de acuerdo al objetivo central; en este caso al haber realizado una exploración en la búsqueda de los modelos con mayor eficacia, es importante reconocer que los modelos construidos desde el aprendizaje superficial fueron desarrollados en espacios símiles, al igual que los modelos de aprendizaje profundo, pero estos últimos fueron modelados en un espacio diferente a los algoritmos de aprendizaje superficial. Desde el aprendizaje superficial se utilizó frecuencia de término y frecuencia inversa de documento (TF-IDF, por sus siglas en inglés), y dentro de los modelos de aprendizaje profundo se implementó Word2Vec.

En este sentido, se reconoce que tanto los modelos utilizados basados en redes neuronales (RNN y CNN) como los modelos de aprendizaje superficial (LR y SVM) obtuvieron buenos resultados en la detección de discurso de odio hacia la comunidad LGBT en tuits escritos en español, sin embargo, la implementación de TF-IDF ha mostrado ligeramente un mejor rendimiento, esto puede deberse en gran parte a la naturaleza de los datos utilizados, pues como se mencionó en el capítulo uno, los textos cortos de Twitter suelen mantener un lenguaje simple y carecer de una semántica compleja.

Considerando que TF-IDF obtiene vectores ponderados que reflejan la relevancia de los términos de acuerdo a la frecuencia de aparición de palabras en los textos utilizados, mientras que Word2Vec se reconoce con un enfoque de representación distribuido al permitir representaciones vectoriales cercanas para palabras con contextos similares facilitando la captura de relaciones semánticas, es posible reconocer que la utilización de TF-IDF puede ser favorable cuando los datos utilizados son textos cortos y

semánticamente simples, no obstante, al dejar de lado el contexto semántico, puede dificultar la identificación de diferencias sutiles en función del contexto y la semántica.

Tal como se muestra en el capítulo cuatro, tanto el modelo LR como SVM, han mostrado un buen rendimiento en la identificación de discurso de odio hacia la comunidad LGBT, sin embargo, muestran dificultades en la diferenciación de aquellos mensajes que se expresan desde el discurso de odio en formas sutiles. Por otra parte, tanto el modelo de CNN como RNN, al implementar Word2Vec aportan un carácter más complejo de clasificación al incluir un carácter semántico, pues muestran mayor eficacia en el reconocimiento de mensajes que expresan discurso de odio de formas sutiles, sin embargo, se identificaron complicaciones en la clasificación de mensajes que expresan discurso de odio de forma evidente o directa.

En ello, es necesario resaltar que el objetivo de la presente investigación ha sido desarrollar un acercamiento a la detección del discurso de odio hacia la comunidad LGBT de manera binaria, motivo por el cual, tanto la implementación de los cuatro modelos como la utilización de TF-IDF y Word2Vec, muestran un contexto general de la aplicación del aprendizaje computacional en la detección de discurso de odio LGBT a través del uso de fuentes de datos no estructuradas.

Por otra parte, es imperativo señalar que modelos como SVM o LR pueden obtener buenos resultados con una pequeña cantidad de datos que alimenten al modelo, mientras que modelos como CNN o RNN al implementar arquitecturas más complejas, requieren la utilización de corpus más grandes. Considerando que la base de datos con la que fueron entrenados los modelos corresponde a un corpus de 8,000 tuits, es posible inferir que los modelos de CNN y RNN podrían necesitar una cantidad mayor de datos para aprender y generalizar de manera adecuada para la clasificación de datos no etiquetados de manera previa.

En cuanto al rendimiento de los modelos, se reconoce que por parte de LR y SVM, es el modelo de Regresión Logística el que ha obtenido los resultados más satisfactorios con un Accuracy de 0.9226 y un F1-Score de 0.9223. Aun cuando SVM ha obtenido un rendimiento ligeramente superior a LR con un Accuracy de 0.9321 y F1-Score de 0.9302, LR fue seleccionado como el ideal para la clasificación del resto de datos que no fueron etiquetados de manera previa por el grupo de clasificadores, en función del nivel de generalización, en donde se reconoció que LR generalizó de mejor manera respecto a SVM, por lo que la clasificación de datos que el modelo no conoce sería más adecuada con LR.

No obstante, es importante reconocer que CNN y RNN han obtenido buenos resultados con un Accuracy de 0.8773 y 0.8848 y un F1-Score de 0.8782 y 0.8812 para CNN y RNN respectivamente. Además, el comportamiento de las pérdidas para estos modelos muestra que tanto el rendimiento como la generalización de los datos es buena, por lo que dicho resultado podría ser más eficiente con la implementación de un corpus integrado por una mayor cantidad de tuits y a la vez permitiendo un análisis más complejo al implementar contextos semánticos, especialmente en la detección de mensajes de odios sutiles.

Con el monitoreo constante de la red social utilizada, el corpus generado y la clasificación obtenida con el modelo de aprendizaje superficial, es posible reconocer que el discurso de odio dirigido hacia la comunidad LGBT en México muestra una predilección por la utilización de ciertas palabras ofensivas utilizadas para denigrar, insultar, estereotipar o menospreciar a una persona o grupo de personas, algunas palabras como

“putito”, “puto”, “joto” o “maricón” se encuentran repetidamente en los textos identificados como discurso de odio. Por lo cual, es posible concluir de manera general que gran parte del discurso de odio dirigido hacia la comunidad LGBT se enfoca en la comunidad gay.

Aun cuando las palabras anteriormente mencionadas son utilizadas comúnmente para referirse de manera negativa a la comunidad gay, es imperativo reconocer que, este tipo de palabras pueden ser utilizadas para referirse a personas que no se perciben única o exclusivamente como homosexuales, sin embargo, con la finalidad de denigrar o insultar a una persona o grupo de personas al expresarse desde un discurso de odio, se puede realizar una generalización, encasillando a la diversidad sexo-genérica únicamente dentro de la comunidad gay.

Además, es importante considerar que este tipo de palabras no son utilizadas única y exclusivamente para referirse a alguna persona auto percibida como LGBT, sino también, como una forma de insultar, denigrar o humillar a cualquier persona o grupo de personas al utilizar estas palabras como sinónimos de cobardía, timidez, o para insultar posturas u opiniones diferentes.

Tomando en consideración la naturaleza de Twitter, es posible encontrar cierta concordancia en la distribución y mayor concentración de tuits clasificados como discurso de odio hacia la comunidad LGBT; la Ciudad de México, Estado de México, Nuevo León y jalisco fueron identificados como los estados en donde hay mayor concentración de este tipo de mensajes, sin embargo, se registró que la mayor presencia de dichos textos se encuentran dentro de las demarcaciones geográficas donde se ubican las ciudades más pobladas de dichos estados, por tal motivo no es sorprendente encontrar que la mayor cantidad de este tipo de mensajes se localice dentro de las ciudades más grandes del país, además de considerar la población flotante que realiza este tipo de publicaciones en dichos estados.

Al calcular las tasas, fue posible visualizar los estados en los que los mensajes referentes a la comunidad LGBT expresados desde el discurso de odio superan por mucho a los mensajes que no lo expresan. Estados como Tamaulipas y San Luis Potosí resaltan al mostrar un valor considerablemente alto.

Sin embargo, es importante tener en cuenta que, estos tuits pueden ser visualizados por cualquier persona que tenga acceso a dicha plataforma sin importar su ubicación, permitiendo la visualización, interacción y reproducción de este tipo de mensajes. Considerando sus implicaciones negativas, se resalta la necesidad de tomar acción en la presencia y reproducción de este tipo de contenido. Al obtener un resultado georreferenciado que muestre las ubicaciones del discurso de odio hacia la comunidad LGBT, permite identificar espacios esenciales que podrían ser clasificados como prioritarios en la búsqueda de reducir este tipo de contenido agresivo y violento.

Tomando en consideración los resultados obtenidos en la detección de discurso de odio, así como los datos recolectados de Visible y Letraese, sobresalen ciertos focos rojos de violencia para la población LGBT. Estados como la Ciudad de México, Estado de México, Jalisco, Veracruz, Nuevo León, Quintana Roo, Puebla Tamaulipas, San Luis Potosí o Chihuahua, resaltan como las entidades con mayor número de registros de violencia o agresiones hacia este grupo poblacional. Sin embargo, algunos otros estados como Guerrero o Guanajuato, sobresalen al considerar los datos recolectados de Visible y Letraese.

La poca existencia de datos relacionados a la comunidad LGBT resaltan la necesidad de incorporar herramientas y metodologías que se centren en estudiar a esta población y obtener datos relevantes que

permitan impulsar su reconocimiento, así como la necesidad de implementar acciones con la finalidad de reducir la violencia a la que puede estar sujeta la comunidad LGBT.

Los resultados obtenidos han mostrado que la utilización de técnicas del aprendizaje computacional puede apoyar en la recolección y generación de datos enfocados en la comunidad LGBT, posibilitando la identificación de espacios prioritarios, además de señalar la necesidad de estudiar a dicha población al considerarla como una población vulnerable mediante la generación y recolección de bases de datos.

5.2 Trabajo futuro

La presente investigación ha mostrado en primera instancia la necesidad de estudiar con datos a la comunidad LGBT, tarea que no es nada fácil al reconocer la ausencia de este tipo de bases de datos, y en segunda instancia la utilidad y posibilidad de implementar modelos basados en aprendizaje computacional para la identificación y clasificación de discurso de odio en bases de datos no estructuradas, en específico en datos textuales.

La implementación de 4 modelos específicos, permitió distinguir el buen rendimiento que representan algunos algoritmos como SVM y LR, para este tipo de tarea y la clasificación binaria de discurso de odio, sin embargo, como se mencionó a lo largo de la investigación, la vanguardia que representan los modelos basados en redes neuronales muestra la oportunidad de realizar tareas más complejas.

En este sentido, se reconoce la importancia de inicialmente enriquecer la base de datos utilizada en el entrenamiento de los modelos, con la finalidad de obtener mejores rendimientos, así como la posibilidad de obtener resultados más complejos, tales como desarrollar una clasificación igual o superior a 3 clases diferentes, permitiendo la distinción de mensajes por grupos objetivo o incluso por el grado de agresividad o violencia efectuada, posibilitando la identificación jerárquica o porcentajes de presencia de discurso de odio hacia cada colectivo perteneciente a la comunidad LGBT.

En este sentido, sería adecuada la implementación de algoritmos más complejos y con estructuras que muestran una sensibilidad notoria al contexto semántico del texto, impulsando la clasificación adecuada y reduciendo los falsos positivos y falsos negativos. Además, se reconoce la importancia de poder implementar datos que carecen de un carácter espacial, en donde con ayuda de ciertas herramientas y metodologías como el geoparsin, podrían ser incorporados.

Finalmente, es imperativo reconocer que las políticas recientemente implementadas por Twitter, dificultan el seguimiento de esta investigación al frenar la descarga indiscriminada y gratuita de los datos publicados, por lo cual, en trabajos futuros y enriqueciendo las fuentes de dato utilizadas, sería pertinente la implementación de otras fuentes de datos alternativas, ya sean redes sociales, medios digitales como periódicos o revistas electrónicas, así como la ampliación en los formatos recolectados, ya sean textuales, visuales o de audio.

Resaltar la necesidad de generar y recolectar bases de datos enfocados a la comunidad LGBT, puede abrir un parteaguas a la creación de herramientas y desarrollo de investigaciones que busquen analizar las problemáticas de violencia, discriminación y segregación que vive esta población objetivo desde enfoques más cuantitativos, abriendo a las esferas públicas y privadas oportunidades de mejoramiento en la calidad de vida de esta comunidad.

Apéndices:

Declaración Universal de los Derechos Humanos (1948)	Toda persona debe tener acceso a los mismos derechos y libertades sin distinción alguna por motivos de raza, religión, nacionalidad o cualquier otra condición
Declaración de Montreal: Derechos Humanos LGBT (2006)	Se reconoce la necesidad de aceptar y respetar ciertos aspectos de la diversidad humana como la orientación sexual o identidad de género, lo cual es identificado como causa de opresión en la vida cotidiana de las personas LGBT en la mayor parte del mundo. Enmarca la necesidad de: 1) Proteger a la población LGBT de la violencia privada y del Estado 2) Reconocer la libertar de expresión, reunión y asociación 3) Libertad de tener relaciones sexuales con personas del mismo sexo (en privado, con consentimiento mutuo y entre adultos)
Principios de Yogyakarta sobre la aplicación del Derecho Internacional de los Derechos Humanos en relación con la orientación sexual y la identidad de género (2007)	Enmarca avances en cuanto a la garantía hacia todas las orientaciones sexuales e identidades de género a vivir con la misma dignidad y respeto. Sin embargo, de igual manera resalta que las violaciones a los derechos humanos hacia esta población continúan, por lo que desarrolla una serie de principios jurídicos internacionales que busquen generar una mayor coherencia en las obligaciones estatales en materia de derechos humanos considerando la orientación sexual e identidad de género
Resolución de la Asamblea General de la OEA sobre Derechos humanos, orientación sexual e identidad de género (2008)	Resalta la necesidad universal de cumplimiento de derechos humanos y a la vez, manifiesta la preocupación a la ejecución y continuación de actos de violencia contra personas debido a su orientación sexual o identidad de género
Convención Interamericana contra toda forma de Discriminación e Intolerancia (2013)	Recalca que una sociedad democrática debe respetar la orientación sexual e identidad de género de toda persona, acentuando que los principios de igualdad y no discriminación impulsan una igualdad jurídica efectiva y que presupone el deber del Estado a adoptar medidas en favor de los grupos poblacionales que son víctimas de discriminación e intolerancia en cualquier esfera ya sea pública o privada
Declaración sobre la violencia contra las mujeres, niñas y	Expone la obligación de los estados a garantizar y respetar los derechos sexuales y reproductivos, reconociendo que la libertad sexual y su desarrollo constituye un bien jurídico

Adolescentes y sus Derechos sexuales y Reproductivos (2014)	respaldado por el reconocimiento internacional de los Derechos Humanos

Anexo 1 Marco normativo internacional a la protección de los derechos de las personas LGBT. Elaboración propia con información de COPRED (2018).

Constitución Política de los Estados Unidos Mexicanos, Reforma del 2011	Estipula la prohibición de discriminación por razones de preferencia sexual. Reconoce el cumplimiento de derechos a toda persona sin distinción alguna
Ley Federal para Prevenir y Eliminar la Discriminación (2014)	Busca sensibilizar sobre la importancia de la diversidad y la garantía de los derechos y el respeto a todas las personas
Protocolo de actuación para quienes imparten justicia en casos que involucren la orientación sexual o la identidad de género (2014)	Destaca la necesidad de auxiliar a los y las juzgadoras en la resolución de asuntos que involucren la afectación de las personas por motivos de orientación sexual o identidad de género
Artículo 149 del Código Penal Federal	Señala el delito y sanciones por la privación o negación a servicios y prestaciones como educación, salud, o trabajo por motivos de la orientación sexual
Protocolo de actuación para quienes imparten justicia para la atención de personas LGBTTI	Establece las reglas que deben cumplir las y los servidores públicos de la PGR que intervengan en la investigación y persecución de los delitos relacionados con las personas LGBT, así como brindar atención a las víctimas que sufran afectaciones físicas o emocionales como resultado de la violencia por motivos de su orientación sexual o identidad de género
Jurisprudencia 43/2015 de la Primera Sala de la Suprema Corte de Justicia de la Nación sobre el matrimonio abierto a las personas del mismo sexo	Reconoce que ninguna norma jurídica limita el matrimonio entre un hombre y una mujer

Anexo 2 Marco normativo nacional a la protección de los derechos de las personas LGBT. Elaboración propia con información de COPRED (2018).

- Diversidad & sexual
- Drag
- Gay
- Homofobico
- Homofobia
- Homosexual
- Identidad & genero
- Joto
- Lencha
- Lesbi
- Lesbiana
- Lesbico gay
- Lgbt
- Lgbtfobia
- Lgbti
- Lgbtti
- Lgbtttiq
- Lgtb
- Machorra
- Marica
- Marico
- Marimach
- Muerde almohada
- Musculoca
- Orientacion & sexual
- Putito
- Puto
- Queer
- sexogenerica
- Tortillera
- Tranfobia
- Trans
- Transexual
- Transfóbico
- Travesti
- Vestida

Anexo 3 Palabras utilizadas para la descarga de tuits. Elaboración propia.

A ese puto	Hombres trans	Pareces joto
A un puto	Homofobia	Pareces lencha
Abiertamente gay	Homofobico	Pareces transexual
Abiertamente gays	Homosexual activista	Pareces travesti
Abiertamente homosexual	Homosexual de cagada	Partir su madre
	Homosexual de mierda	Partir tu madre
Abiertamente lencha	Homosexual imbecil	Partirles la madre
Abiertamente lesbiana	Homosexual imbécil	Paso por los huevos
Abiertamente lesbianas	Homosexual pendejo	Pedazo de puto
Abiertamente lgbt	Homosexuales activistas	Pedo con el joto
Activismo gay	Homosexuales de cagada	Pedo con la lencha
Activista trans	Homosexuales de mierda	Pendeja y lesbiana
Activistas trans	Homosexuales Pervertidos	Pendejadas lgbt
Ademas de joto	Homosexuales putos	Pendejo homosexual
Además de joto	Identidad de género	Pendejo joto
Ademas de puto	Identidad de género	Pendejo y desviado
Además de puto	Identidad de genero mis	Pendejo y gay
Agresion LGBT	huevos	Pendejo y homosexual
Agresión LGBT	Identidad de género mis	Pendejo y joto
Aguanta nada el puto	huevos	Pendejo y maricon

Alto a la violencia	Idioteces jotos	Pendejo y maricón
Amanerado de mierda	Idioteces lgbt	Persona gay
Amanerados de mierda	Inclusion de la ciudadania	Persona lesbiana
Andan de jotos	Inclusión de la ciudadanía	Persona lgbt
Andan de putos	Inclusión de la comunidad	Persona transexual
Andar de puto	lgbt	Persona transgenero
Andas de puto	Inclusión lgbt	Persona transgénero
Apoyar a la comunidad	Jota de mierda	Persona travesti
lgbt	Jotas de mierda	Personas de la diversidad
Apoyo a la comunidad	Joto	Personas del mismo sexo
lgbt	Joto de cagada	Personas del mismo sexo
Apoyo a la diversidad	Joto de mierda	Personas gay
sexual	Joto idiota	Personas gays
Apoyo a la transexualidad	Joto imbecil	Personas lesbianas
Apoyo a las lenchas	Joto imbécil	Personas lgbt
Apoyo a las lesbianas	Joto malparido	Personas travestis
Apoyo a las mujeres trans	Joto pendejo	Perspectiva de género
Apoyo a las trans	Jotos de mierda	Peste homosexual
Apoyo a las transexuales	Jotos pervertidos	Pinche amanerado
Apoyo a las travestis	Jóvenes lgbt	Pinche desviada
Apoyo a los gays	Lencha de cagada	Pinche desviado
Apoyo a los	Lencha de mierda	Pinche gay
homosexuales	lencha estupida	Pinche homosexual
Apoyo a los trans	Lencha malparida	Pinche jota
Apoyo a los travestis	Lencha pendeja	Pinche joto
Apoyo al travestismo	Lenchas de mierda	Pinche lencha
Apoyo lgbt	Lesbaian pendeja	Pinche lesbiana
Asco homosexual	Lesbiana activista	Pinche machorra
Asco homosexuales	Lesbiana de mierda	Pinche marica
Asco joto	lesbiana idiota	Pinche maricon
Asco jotos	Lesbiana malparida	Pinche maricón
Asco lencha	Lesbiana pervertida	Pinche marimach
Asco lenchas	Lesbianas activistas	Pinche muerde almohadas
Asco me dan	Lesbianas de mierda	Pinche putito
Asco trans	Lesbianas pervertidas	Pinche puto
Asco un homosexual	Lesbofobia	Pinche tortillera
Asco un joto	lgbt de cagada	Pinche transexual
Asco una lencha	lgbt de mierda	Pinche travesti
Atención a la diversidad	lgbt mis huevos	Pinche vieja lencha
sexual	Libertad de expresión	Pinche vieja lesbiana
Bifobia	Libre expresión	Pinches amanerados
Bola de putos	Lucha de los derechos	Pinches desviadas
Bueno por puto	homosexuales	Pinches desviados

Cállate puto	Lucha de los derechos	Pinches gays
Callateputo	humanos	Pinches homosexuales
Cara de puto	Lucha de los derechos lgbt	Pinches jotas
Chairo puto	Machorra de mierda	Pinches jotos
Che joto	Machorras de mierda	Pinches lenchas
Che puto	Machorras resentidas	Pinches lesbianas
Chinga tu madre	Maldita desviada	Pinches lgbt
Chingan a su madre	Maldita jota	Pinches machorras
Chinguen a su madre	Maldita lencha	Pinches maricas
Chinguen su madre	Maldita lesbiana	Pinches maricones
Chtpm joto	Maldita machorra	Pinches marimach
Chtpm lencha	Maldita marimach	Pinches muerde almohadas
Chtpm transexual	Maldita tortillera	Pinches putitos
Chtpm travesti	Maldita transexual	Pinches putos
Chupa pito	Maldita travesti	Pinches tortilleras
Chupa pitos	Malditas desviadas	Pinches transexuales
Comunidad gay	Malditas jotas	Pinches travestis
Comunidad homosexual	Malditas lenchas	Pinches viejas lenchas
Comunidad lesbica	Malditas lesbianas	Poblaciones lgbt
Comunidad lgbt	Malditas machorras	PoblacionesLGBT
Comunidad trans	Malditas marimach	Pobre joto
Contra la bifobia	Malditas tortilleras	Pobre marica
Contra la comunidad lgbt	Malditas transexuales	Pobre maricon
Contra la homofobia	Malditas travestis	Pobre marimacha
Contra la lesbofobia	Maldito amanerado	Pobre marimacho
Contra la transfobia	Maldito desviado	Pobre travesti
Cosa de homosexuales	Maldito gay	Por qué no se mueren
Cosa de putos	Maldito homosexual	Porque no se mueren
Cosas de homosexuales	Maldito joto	Preferencia sexual
Cosas de putos	Maldito marica	Preferencias sexuales
Crimen de odio LGBT	Maldito maricon	Prevención de la violencia
Deberían morirse	Maldito maricón	puñales
Deberias morirte	Maldito muerde	Puñetas
Decir nada al puto	almohadas	Puñeton
Defensa de los derechos	Maldito putito	Puñetón
Delitos por diversidad de	Maldito puto	Puta lencha
género	Maldito transexual	Puta lesbiana
Delitos por diversidad	Maldito travesti	Puta machorra
sexual	Malditos amanerados	Puta tortillera
Derecho lgbt	Malditos desviados	Puta travesti
Derechos homosexuales	Malditos gays	Putas lenchas
Derechos humanos	Malditos homosexuales	Putas lesbianas
Derechos lgbt	Malditos jotos	Putas machorras

Derechos lgbt	Malditos lgbt	Putas mariconadas
Derechos lgbtttiq	Malditos maricas	Putas tortilleras
Desviada de mierda	Malditos maricones	Putas travestis
Desviadas de mierda	Malditos muerde	Putito
Desviado de mierda	almohadas	Putito
Desviados de mierda	Malditos putitos	Putito imbécil
Discriminación a la	Malditos putos	Putito pendejo
comunidad lgbt	Malditos transexuales	Putitode mierda
Discriminación a la	Malditos travestis	Puto amanerado
diversidad sexual	Mamadas de lgbt	Puto de mierda
Discriminación hacia la	Mandalos a la verga	Puto desviado
comunidad lgbt	Mandalos alv	Puto el que lo lea
Discriminación hacia la	Maricon de mierda	Puto gay
diversidad sexual	Marica de mierda	Puto homosexual
Discriminación por ser	Maricas de mierda	Puto imbecil
gay	Maricon de cagada	Puto joto
Discriminación por ser	Maricon malparido	Puto malparido
homosexual	Maricón malparido	Puto marica
Discriminación por ser	Mariconadas	Puto maricon
lencha	Maricones de cagada	Puto maricón
Discriminación por ser	Maricones de cagada	Puto muerde almohadas
lesbiana	Maricones de mierda	Puto transexual
Discriminación por ser	Marimach de mierda	Puto travesti
trans	Marimach de mierda	Putos amanerados
Discriminación por ser	Marimacha	Putos de mierda
transexual	Marimachas	Putos desviados
Discriminación por ser	Mariquita	Putos gays
travesti	Me dan asco	Putos homosexuales
Discriminación por ser	Morro puto	Putos jotos
una persona trans	Movimiento de las mujeres	Putos lgbt
Discriminación por ser	lesbianas	Putos maricas
una persona transexual	Movimiento de las	Putos maricones
Discriminada por ser	personas gays	Putos muerde almohadas
bisexual	Muerde almohadas de	Putos transexuales
Discriminada por ser	mierda	Putos travestis
lencha	Mugrosa lesbiana	Que puto eres
Discriminada por ser	Mugrosas lenchas	Que puto es
lesbiana	Mugrosas transexuales	Respeto a la identidad de
Discriminada por ser	Mugrosas travestis	genero
trans	Mugroso gay	Respeto a la identidad de
Discriminada por ser	Mugroso homosexual	género
transexual	Mugroso joto	Respeto a los derechos
	Mugroso putito	

Discriminado por ser bisexual	Mugroso puto	Respeto a los homosexuales
Discriminado por ser homosexual	Mugrosa lesbiana	Se muera puto
Discriminado por ser joto	Mugrosa tortillera	Se muere el puto
Discriminado por ser trans	Mugrosos gays	Se van a ir al infierno
Discurso de odio LGBT	Mugrosos homosexuales	Ser parte de la comunidad
Diversidad sexo genérica	Mugrosos jotos	Ser parte de la comunidad lgbt
Diversidad sexogenerica	Mugrosos putitos	Serofovia
Diversidad sexogenérica	Mugrosos transexuales	Sexual y de género
Diversidad sexual	Mugrosos travestis	Si es puto
Diversidad sexual mis huevos	Mujer trans	Sidoso
Diversidadsexual	Mujeres bisexuales	Sidosos
El muy joto	Mujeres lesbianas	Soy gay y que
El muy marica	Mujeres trans	Soy joto y que
El muy puto	Nada que curar	Soy lencha y que
El puto ese	Nadaquecurar	Soy lesbiana y que
Empezó de gay	No a la bifobia	Soy parte de la comunidad lgbt
Empezó de joto	No a la discriminacion	Soy puto y que
En contra de las lesbianas	No a la discriminación	Soy trans y que
En contra de los homosexuales	No a la discriminacion lgbt	Te odio puto
Enfermas mentales	No a la discriminación lgbt	Te sale lo puto
Enfermos mentales	No a la homofobia	Te vas a ir al infierno
Eres puto	No a la lesbofobia	Todo joto
Eres un joto	No a la transfobia	Todo puto
Eres un marica	No al odio	Tonterias lgbt
Eres un maricon	No discriminación	Tontos homosexuales
Eres un maricón	No eres una mujer de verdad	Tortillera de mierda
Eres un putito	No es una mujer de verdad	tortillera estupida
Eres un puto	No estes de lencha	Tortilleras de mierda
Es bifobia	No estés de lencha	Trans de mierda
Es bifobica	No mas bifobia	Transexual de mierda
Es bifobico	No más bifobia	Transexual malparida
Es homofobia	No mas discriminacion	Transexuales de mierda
Es homofobica	No más discriminación	transexuales pendejos
Es homofobico	No mas discriminacion lgbt	transexuales pendejos
Es lesbofobia	No más discriminación lgbt	Transfeminicidio
Es lesbofobica	No mas homofobia	transfobia
Es lesbofobico	No más homofobia	Transicionar
Es lgbtfobia	No mas lesbofobia	Transmisoginia
	No más lesbofobia	Travesti de mierda
	No mas odio	Travesti imbécil
	No más odio	
	No mas transfobia	

Es lgbtfobica	No más transfobia	Travestis de cagada
Es lgbtfobico	No que no puto	Travestis de mierda
Es transfobia	No sea puto	Tremendo joto
Es transfobica	No seas homosexual	Va empezar de gay
Es transfobico	No seas joto	Valen verga
Es un joto	No seas joto	valen verga putos
Es un marica	No seas lencha	Vales verga
Es un maricon	No seas puto	Vales verga homosexual
Es un maricón	Noalabifobia	Vales verga puto
Es un putito	Noaladiscriminacion	Valga verga pinche puto
Es una machorra	Noalahomofobia	Van a empezar de gays
Esas machorras	Noalalesbofobia	Van a empezar de
Ese we es puto	Noalalesbofobia	homosexuales
Ese wey es puto	Noalatransfobia	Vayanse a la mierda
Eso que puto	Noalodio	Vayanse a la verga
Espacios lgbt	Nomasbifobia	Vayanse alv
Espacios seguros lgbt	Nomasdiscriminacion	Ven al puto
Estupideces lgbt	Nomasdiscriminacionlgbt	Verga puto
Estúpido joto	Nomashomofobia	Ves al puto
Estupidos homosexuales	Nomaslesbofobia	Vete a la verga
Estupidos jotos	Nomasodio	Vete alv
Estupidos los putos	Nomastransfobia	Viejo puto
Estúpidos los putos	Ora puto	Violencia contra la
Estúpidos transexuales	Orgullo bisexual	comunidad lgbt
Estúpidos travestis	Orgullo gay	Violencia contra las
Garantía de los derechos	Orgullo lésbico	lesbianas
Gay activista	Orgullo lgbt	Violencia contra las
Gay de cagada	Orgullo lgbt	personas lgbt
Gay de mierda	Orgullo transexual	Violencia contra las
Gay estúpido	Orgullolgbt	personas trans
Gay imbécil	Orgullosa de ser	Violencia contra las trans
Gay malparido	Orgulloso de ser	Violencia contra los
Gays activistas	Pansexuales	homosexuales
Gays de mierda	Parece transexual	Violencia contra los trans
Hay que matarlos	Parece travesti	Violencia LGBT
Hombres bisexuales	Parecen lenchas	Violentar
Hombre trans		

Anexo 4 Palabras utilizadas para filtrar la base de datos.

Estado	Tuits que hacen referencia a la comunidad LGBT por años			Tuits que expresan discurso de odio a la comunidad LGBT por años			Tasas de Discurso de odio por años			Total de tuits	Total de tuits identificados como discurso de odio	Tasas de Discurso de odio en conjunto
	2020	2021	2022	2020	2021	2022	2020	2021	2022			
Tamaulipas	141	144	120	99	99	85	70	69	71	405	283	70
San Luis Potosí	209	169	192	149	103	117	71	61	61	570	369	65
Quintana Roo	266	182	197	189	78	105	71	43	53	645	372	58
Guerrero	59	56	84	45	29	39	76	52	46	199	113	57
Estado de México	763	691	624	497	338	303	65	49	49	2078	1138	55
Tlaxcala	16	29	20	10	15	6	63	52	30	65	31	48
Nuevo León	781	648	527	386	304	210	49	47	40	1956	900	46
Hidalgo	85	63	63	47	30	16	55	48	25	211	93	44
Campeche	51	52	30	21	29	8	41	56	27	133	58	44
Colima	41	21	39	22	3	19	54	14	49	101	44	44
Aguascalientes	89	81	62	51	27	22	57	33	35	232	100	43
Baja California	153	209	157	68	84	66	44	40	42	519	218	42
Veracruz	347	239	200	156	90	83	45	38	42	786	329	42
Sinaloa	141	84	90	75	31	24	53	37	27	315	130	41
Puebla	343	262	219	155	98	83	45	37	38	824	336	41
Guanajuato	222	148	214	116	57	65	52	39	30	584	238	41
Nayarit	29	21	41	9	11	17	31	52	41	91	37	41
Tabasco	114	72	103	49	26	41	43	36	40	289	116	40
Chiapas	51	48	65	22	19	24	43	40	37	164	65	40
Coahuila	130	103	130	58	47	38	45	46	29	363	143	39
Querétaro	264	234	245	129	70	91	49	30	37	743	290	39
Chihuahua	129	108	93	51	35	42	40	32	45	330	128	39
Jalisco	708	590	668	320	215	211	45	36	32	1966	746	38
Morelos	110	77	98	46	24	35	42	31	36	285	105	37
Oaxaca	71	76	100	30	33	26	42	43	26	247	89	36
Sonora	214	171	172	91	60	46	43	35	27	557	197	35
Baja California Sur	62	47	71	26	13	23	42	28	32	180	62	34
Ciudad de México	1939	1671	2568	823	642	587	42	38	23	6178	2052	33
Yucatán	181	186	190	77	43	40	43	23	21	557	160	29
Zacatecas	81	50	55	14	22	13	17	44	24	186	49	26
Durango	50	36	72	16	3	15	32	8	21	158	34	22
Michoacán	187	318	179	48	44	35	26	14	20	684	127	19

Anexo 5 Total de tuits, tuis identificados como discurso de odio y tasas

Lista de tablas

Lista de figuras

Referencias bibliográficas

Aguilar Pirachicán, M. (2019). El discurso del odio (reseña). *Desde el Jardín de Freud*, [online] (1657-3986), pp.328–333. doi:https://doi.org/10.15446/djf.n19.76731.

Alain, G. y Bengio, Y. (2014). What Regularized Auto-Encoders Learn from the Data-Generating Distribution. *Journal of Machine Learning Research*, 15, pp.3743–3773.

Alshalan, R. y Al-Khalifa, H. (2020). A Deep Learning Approach for Automatic Hate Speech Detection in the Saudi Twittersphere. *Applied Sciences*, 10(23), p.8614. doi:https://doi.org/10.3390/app10238614.

Anezi, F.Y.A. (2022). Arabic Hate Speech Detection Using Deep Recurrent Neural Networks. *Applied Sciences*, 12(12), p.6010. doi:https://doi.org/10.3390/app12126010.

Arcila Calderón, C., Blanco-Herrero, D. y Valdez Apolo, M.B. (2020). Rechazo y discurso de odio en Twitter: análisis de contenido de los tuits sobre migrantes y refugiados en español / Rejection and Hate Speech in Twitter: Content Analysis of Tweets about Migrants and Refugees in Spanish. *Revista Española de Investigaciones Sociológicas*, 172(0210-5233), pp.21–40. doi:https://doi.org/10.5477/cis/reis.172.21.

Arcila-Calderón, C., Amores, J.J., Sánchez-Holgado, P. y Blanco-Herrero, D. (2021). Using Shallow and Deep Learning to Automatically Detect Hate Motivated by Gender and Sexual Orientation on Twitter in Spanish. *Multimodal Technologies and Interaction*, 5(10), p.63. doi:https://doi.org/10.3390/mti5100063.

Badjatiya, P., Gupta, S., Gupta, M. y Varma, V. (2017). Deep Learning for Hate Speech Detection in Tweets. *Proceedings of the 26th International Conference on World Wide Web Companion - WWW '17 Companion*. [online] doi:https://doi.org/10.1145/3041021.3054223.

Baydoğan, C. y Alatas, B.A. (2022). Deep-Cov19-Hate: A Textual-Based Novel Approach for Automatic Detection of Hate Speech in Online Social Networks throughout COVID-19 with Shallow and Deep Learning Models. *Tehnicki vjesnik - Technical Gazette*, 29(1), pp.149–156. doi:https://doi.org/10.17559/tv-20210708143535.

Bi, Q., Goodman, K.E., Kaminsky, J. y Lessler, J. (2019). What is Machine Learning? A Primer for the Epidemiologist. *American Journal of Epidemiology*, 188(12). doi:https://doi.org/10.1093/aje/kwz189.

Bilal, M., Khan, A., Jan, S. y Musa, S. (2022). Context-Aware Deep Learning Model for Detection of Roman Urdu Hate Speech on Social Media Platform. *IEEE Access*, 10, pp.121133–121151. doi:https://doi.org/10.1109/access.2022.3216375.

Blanco Rojas, T., Archila Córdoba, D. y Ballesteros-Ricaurte, J. (2016). Gestión de datos obtenidos desde redes sociales aplicando Business Intelligence Engineering Process. *Revista Virtual Universidad Católica del Norte*, (0124-5821), pp.72–91.

Bolaños Enríquez, T. y Charry Morales, A. (2018). PREJUICIOS Y HOMOSEXUALIDAD, EL LARGO CAMINO HACIA LA ADOPCIÓN HOMOPARENTAL. ESPECIAL ATENCIÓN AL CASO COLOMBIANO. *Estudios*

constitucionales, [online] 16(1), pp.395–424. doi:https://doi.org/10.4067/s0718-52002018000100395.

Bosque González, I. del, Fernández Freire, C., Martín-Forero Morente, L. y Pérez Asensio, E. (2012). *Los Sistemas de Información Geográfica y la Investigación en Ciencias Humanas y Sociales*. [online] *digital.csic.es*. Confederación Española de Centros de Estudios Locales. Available at: http://hdl.handle.net/10261/64940 [Accessed 12 Jan. 2023].

Calderón, J., Tellez, E. y Graff, M. (2021). *DCCD-INFOTEC at MeOffendEs@IberLEF21 Subtask 3: A Transfer Learning Approach Based on EvoMSA's Stacked Generalization*. [online] Available at: https://ceur-ws.org/Vol-2943/meoffendes_paper5.pdf [Accessed 22 Mar. 2023].

Clausen, R., Luengo-Oroz, M., B. Mello, M., Paz, J., Pantin, C. y Erkkola, T. (2017). Social Media Monitoring of Discrimination and HIV Testing in Brazil, 2014–2015. *Springer*, [online] pp.114–120. doi:https://doi.org/10.1007/s10461-017-1753-2.

Comisión Interamericana de Derechos Humanos (CIDH) (2018). *Avances y Desafíos hacia el reconocimiento de los derechos de las personas LGBTI en las Américas*. OEA Más derechos para más gente.

Comisión Nacional de los Derechos Humanos (CNDH) (2022). *Derechos: Libertad de Expresión*. [online] CNDH México defendamos al pueblo. Available at: https://www.cndh.org.mx/pagina/derechos-libertad-de-expresion [Accessed 6 Dec. 2022].

Consejo Nacional de Población (2022). *¿Sabes qué es la diversidad sexual y de género?* [Página web Gobierno de México] *Consejo Nacional de Población, Documentos*. Available at: https://www.gob.mx/conapo/documentos/sabes-que-es-la-diversidad-sexual-y-de-genero?idiom=es [Accessed 4 Dec. 2022].

Consejo para Prevenir y Eliminar la Discriminación de la CDMX (COPRED) (2018). *POBLACIÓN LGBTTTI*. *https://copred.cdmx.gob.mx/storage/app/uploads/public/5b1/ff9/f94/5b1ff9f945326665643161.pdf*, pp.1–33.

Consejo para Prevenir y Eliminar la Discriminación de la Ciudad de México (COPRED) (2021). *Violencia, discriminación y resiliencia en personas jóvenes LGTBI+: antes y durante la pandemia por COVID-19 en la Ciudad de México.* Ciudad de México: COPRED.

CONSTITUCIÓN POLÍTICA DE LOS ESTADOS UNIDOS MEXICANOS.

da Silva Anes, C. (2021). *Os Discursos de Ódio Contra Pessoas LGBT em Contexto Online em*. [Tesis de Maestría] pp.1–51. Available at: https://repositorio-aberto.up.pt/bitstream/10216/138040/3/517322.1.pdf [Accessed 28 Dec. 2022].

da Silva, M.P. y da Silva, L.S. (2021). Disseminação de discursos de ódio em comentários de notícias: uma análise a partir de notícias sobre o universo LGBT em cibermeios sul-mato-grossenses no Facebook. *Intercom: Revista Brasileira de Ciências da Comunicação*, 44(2), pp.137–155. doi:https://doi.org/10.1590/1809-5844202127.

de Benito, E. (2018). La OMS saca la transexualidad de la lista de enfermedades mentales. *El País*. [online] 19 Jun. Available at: https://elpais.com/internacional/2018/06/18/actualidad/1529346704_000097.html [Accessed 6 Apr. 2021].

De la Cruz, I. (2019). Redes Sociales como Fuentes de Información sobre Salud. *Práctica Familiar Rural*, 4(2). doi:https://doi.org/10.23936/pfr.v4i2.95.

de los Cobos Alcalá, P. (2021). *2do Informe de actividades 2020-2021 (visible.lgbt)*. pp.1–100.

Eberendu, A.C. (2016). Unstructured Data: an overview of the data of Big Data. *International Journal of Computer Trends and Technology*, 38(1), pp.46–50. doi:https://doi.org/10.14445/22312803/ijctt-v38p109.

Escolano Utrilla, S. (2017). *Los espacios urbanos: procesos y organización territorial, elementos para su estudio en el grado de 'Geografía y Ordenación del Territorio', Universidad de Zaragoza.* Tesis de grado en Geografía y Ordenamiento del Territorio.

Ferentinos, S. (2016). INTERPRETING LGBTQ HISTORIC SITES. In: M. Springate, ed., *LGBTQ AMERICA a THEME STUDY OF LESBIAN, GAY, BISEXUAL, TRANSGENDER and QUEER HISTORY*. National Park Foundation, pp.1–23.

Fitri, V.A., Andreswari, R. y Hasibuan, M.A. (2019). Sentiment Analysis of Social Media Twitter with Case of Anti-LGBT Campaign in Indonesia using Naïve Bayes, Decision Tree, and Random Forest Algorithm. *Procedia Computer Science*, 161, pp.765–772. doi:https://doi.org/10.1016/j.procs.2019.11.181.

Fortuna, P. y Nunes, S. (2018). A Survey on Automatic Detection of Hate Speech in Text. *ACM Computing Surveys*, 51.

Gambäck, B. y Kumar Sikdar, U. (2017). Using Convolutional Neural Networks to Classify Hate-Speech. In: *Association for Computational Linguistics*. Abusive Language Online. Canada, pp.85–90.

Ganfure, G.O. (2022). Comparative analysis of deep learning based Afaan Oromo hate speech detection. *Journal of Big Data*, 9(1). doi:https://doi.org/10.1186/s40537-022-00628-w.

García Patiño, G. ed., (2020). *Violencia, Derechos Humanos y Sexualidad*. México: Fundación Arcoíris por el Respeto a la Diversidad Sexual, A.C., pp.1–162.

Gieseking, J. (2016). LGBTQ Spaces and Places. In: M. Springate, ed., *LGBTQ AMERICA a THEME STUDY OF LESBIAN, GAY, BISEXUAL, TRANSGENDER and QUEER HISTORY*. National Park Foundation, pp.1–31.

Giribuela, W. (2018). Cuestión social y diversidad sexual Aproximaciones iniciales al análisis de la orientación sexo-genérica disidente como emergente de la cuestión social. *ConCienciaSocial. Revista digital de Trabajo Social*, [online] 2(2591-5339), pp.57–73. Available at: https://revistas.unc.edu.ar/index.php/ConCienciaSocial/ [Accessed 30 Nov. 2022].

Giribuela, W. (2019). Las identidades conformadas a partir de orientaciones sexo-genéricas disidentes.

In: L. Riveiro, ed., *Trabajo Social y feminismos Perspectivas y estrategias en debate*. Argentina: Colegio de Trabajadores Sociales de la Provincia de Buenos Aires, pp.105–129.

Gobierno de México (2016). *Día de la Libertad de Expresión en México*. [online] GOBIERNO DE MÉXICO. Available at: https://www.gob.mx/segob/articulos/dia-de-la-libertad-de-expresion-en-mexico [Accessed 6 Dec. 2022].

Gómez-Espinosa, V., Muúniz-Sanchez, V. y Pastor López-Monroy, A. (2021). *Transformers Pipeline for Offensiveness Detection in Mexican Spanish Social Media*. IberLEF@SEPLN.

González Ortuño, G. (2017). Teorías de la disidencia sexual: de contextos populares a usos elitistas. La teoría queer en América latina frente a las y los pensadores de disidencia sexogenérica. *De Raíz Diversa. Revista Especializada en Estudios Latinoamericanos*, 3(5), p.179. doi:https://doi.org/10.22201/ppela.24487988e.2016.5.58507.

González Zuccolotto, K. (2021). *Evaluación de métodos de visión por computadora para la detección y clasificación automática de formas de techos en imágenes satelitales y datos de altimetría LIDAR*. Tésis de Maestría. pp.1–91.

Goodchild, M.F. (2009). Geographic information systems and science: today and tomorrow. Annals of GIS, 15(1), pp.3–9. doi:https://doi.org/10.1080/19475680903250715.

Gregory, I., Donaldson, C., Murrieta-Flores, P. y Rayson, P. (2015). Geoparsing, GIS, and Textual Analysis: Current Developments in Spatial Humanities Research. *International Journal of Humanities and Arts Computing*, 9(1), pp.1–14. doi:https://doi.org/10.3366/ijhac.2015.0135.

Gritta, M., Pilehvar, M.T. y Collier, N. (2019). A pragmatic guide to geoparsing evaluation. *Language Resources and Evaluation*, 54(3), pp.683–712. doi:https://doi.org/10.1007/s10579-019-09475-3.

Hernández Duran, R. (2022). *Procesos de discriminación a la comunidad LGBTTTIQ+, en la Universidad Veracruzana*. [Tesis para creditar la Experiencia recepcional] pp.1–113. Available at: https://www.uv.mx/personal/jedorantes/files/2022/05/Procesos-de-discriminacion-a-la-comunidad-LGBTTTIQ-en-la-Universidad-Veracruzana.-1.pdf [Accessed 30 Nov. 2022].

https://www.facebook.com/rosa.peiro.79 (2017). *Redes sociales - Definición, qué es y concepto | Economipedia*. [online] Economipedia. Available at: https://economipedia.com/definiciones/redes-sociales.html.

Íñiguez, L. y Pol, E. (1996). *Cognición, representación y apropiación del espacio*. España: Universidad de Barcelona.

Instituto Mexicano de la Juventud (INJUVE) (2017). *¿Qué significa LGBTTTIQ?* [online] Gobierno de México. Available at: https://www.gob.mx/imjuve/articulos/que-significa-lgbtttiq [Accessed 29 Nov. 2022].

Instituto Nacional de Desarrollo Social (Indesol) (2018). *Derechos de las personas LGBTI en la política pública*. México: Indeol, pp.1–158.

Instituto Nacional de Estadística y Geografía (INEGI) (2021). Encuesta Nacional sobre Diversidad Sexual y de Género (ENDISEG) 2021.

Jacks, william y Adler, J. (2015). A Proposed Typology of Online Hate Crime. *Forensic Psychology*, 7(1948-5115), pp.64–89.

Janowicz, K., Gao, S., McKenzie, G., Hu, Y. and Bhaduri, B. (2019). GeoAI: spatially explicit artificial intelligence techniques for geographic knowledge discovery and beyond. International Journal of Geographical Information Science, 34(4), pp.625–636. doi:https://doi.org/10.1080/13658816.2019.1684500.

Jiang, S., Nocera, A., Tatar, C., Yoder, M.M., Chao, J., Wiedemann, K., Finzer, W. y Rosé, C. (2022). An empirical analysis of high school students' practices of modelling with unstructured data. *British Journal of Educational Technology*, 53(5), pp.1114–1133. doi:https://doi.org/10.1111/bjet.13253.

Jugănaru, I. (2018). Creating an identity – safe spaces and events in LGBTQIA+ community: A literature review. *Journal of Comparative Research in Anthropology and Sociology*, [online] 9(2068 – 0317), pp.35–45. Available at: http://compaso.eu/wpd/wp-content/uploads/2019/08/Compaso2018-92-Juganaru.pdf [Accessed 23 Nov. 2022].

Karami, A., Lundy, M., Webb, F., Boyajieff, H.R., Zhu, M. y Lee, D. (2021). Automatic Categorization of LGBT User Profiles on Twitter with Machine Learning. *Electronics*, [online] 10(15), pp.1–15. doi:https://doi.org/10.3390/electronics10151822.

KASIANCZUK, M. (2022). HATE SPEECH TOWARDS LGBT IN UKRAINIAN ONLINE MEDIA. *Sociology: Theory, Methods, Marketing*, (Stmm. 2022 (2)), pp.138–161. doi:https://doi.org/10.15407/sociology2022.02.138.

Kibble, R. (2013). *Introduction to natural language processing*. [online] Reino Unido: University of London, pp.1–113. Available at: https://www.london.ac.uk/sites/default/files/study-guides/introduction-to-natural-language-processing.pdf [Accessed 23 Nov. 2022].

Killermann, S. (2011). *The Genderbread Person by Sam Killermann*. [online] www.samkillermann.com. Available at: https://www.samkillermann.com/work/genderbread-person/ [Accessed 28 Nov. 2022].

Lefebvre, H. (2013). *La producción del espacio*. Madrid: Capitán Swing.

LetraEse (n.d.). *NOSOTROS – LetraEse*. [online] Available at: https://letraese.org.mx/nosotros/ [Accessed 21 May 2023].

Longley, P.A., Goodchild, M.F., Maguire, D.J. and Rhind, D.W. (2015). Geographic Information Science and Systems. [online] Google Books. John Wiley & Sons. Available at: https://books.google.com.mx/books?hl=es&lr=&id=C_EwBgAAQBAJ&oi=fnd&pg=PR10&dq=geographic+information+science&ots=r6PEZv27OK&sig=lXOsY6HKESWF9UT1kRUdF-mQsA8#v=onepage&q=geographic%20information%20science&f=true [Accessed 8 Jan. 2024].

López De Loera, B. (2019). Violencia en las parejas de la comunidad LGBT (Lésbico, Gay, Bisexual,

Transgénero, Transexual). *Revista Electrónica de Psicología Iztacala*, [online] 22(1), pp.106–121. Available at: https://www.iztacala.unam.mx/carreras/psicologia/psiclin/vol22num1/Vol22No1Art7.pdf [Accessed 31 Oct. 2022].

López Ulla, J. (2018). El discurso del odio como concepto: dificultades para definirlo en la práctica. In: Reyes Pérez Alberdi, ed., *Diálogos judiciales en el sistema europeo de protección de derechos: una mirada interdisciplinar*. Valencia: Tirant lo Blanch, pp.399–412.

MacAvaney, S., Yao, H.-R., Yang, E., Russell, K., Goharian, N. y Frieder, O. (2019). Hate speech detection: Challenges and solutions. *PLoS ONE*, [online] 14(8). Available at: https://doi.org/10.1371/journal.pone.0221152 [Accessed 6 Dec. 2022].

Maria Nancy, A. y Maheswari, R. (2020). A REVIEW ON UNSTRUCTURED DATA IN MEDICAL DATA. *Journal of Critical Reviews*, 7(13, 2020), pp.2202–2208.

Martí, P., Serrano-Estrada, L. y Nolasco-Cirugeda, A. (2019). Social Media data: Challenges, opportunities and limitations in urban studies. *Computers, Environment and Urban Systems*, [online] 74, pp.161–174. doi:https://doi.org/10.1016/j.compenvurbsys.2018.11.001.

Mena Roa, M. (2021). *Infografía: El Big Bang del Big Data*. [online] Statista Infografías. Available at: https://es.statista.com/grafico/26031/volumen-estimado-de-datos-digitales-creados-o-replicados-en-todo-el-mundo/ [Accessed 11 Jan. 2023].

Meyer, D. (2016). *Violence against queer people : race, class, gender, and the persistence of anti-LGBT discrimination*. New Brunswick, New Jersey: Rutgers University Press.

Miranda-Jiménez, S., Graff, M., Tellez, E.S. y Moctezuma, D. (2017). *INGEOTEC at SemEval 2017 Task 4: A B4MSA Ensemble based on Genetic Programming for Twitter Sentiment Analysis*. [online] ACLWeb. doi:https://doi.org/10.18653/v1/S17-2130.

Miranda-Jiménez, S., Tellez, E.S., Graff, M. y Moctezuma, D. (2020). *INGEOTEC at SemEval-2020 Task 12: Multilingual Classification of Offensive Text*. [online] ACLWeb. doi:https://doi.org/10.18653/v1/2020.semeval-1.262.

Miró Llinares, F. (2016). Taxonomía de la comunicación violenta y el discurso del odio en Internet. *IDP. Revista de Internet, Derecho y Política*, [online] (22), pp.82–107. doi:https://doi.org/10.7238/idp.v0i22.2975.

Nielsen, R.C., Luengo-Oroz, M., Mello, M.B., Paz, J., Pantin, C. y Erkkola, T. (2017). Social Media Monitoring of Discrimination and HIV Testing in Brazil, 2014-2015. *AIDS and behavior*, [online] 21(Suppl 1), pp.114–120. doi:https://doi.org/10.1007/s10461-017-1753-2.

Oriola, O. y Kotze, E. (2020). Evaluating Machine Learning Techniques for Detecting Offensive and Hate Speech in South African Tweets. *IEEE Access*, 8, pp.21496–21509. doi:https://doi.org/10.1109/access.2020.2968173.

Prince Torres, Á. (2021). El acoso contra la comunidad LGBTIQ+ y el derecho a la paz: implicaciones educativas en Latinoamérica. *Revista Latinoamericana de Estudios Educativos*, [online] 51(2), pp.271–298. doi:https://doi.org/10.48102/rlee.2021.51.2.381.

Riquelme Silva., R. (2019). *Detección de violencia verbal hacia las mujeres en redes sociales mediante técnicas de aprendizaje automático.* Tesis de ingeniero en ejecución en computación e informática. pp.1–82.

Rodríguez Ortiz, A. (2019). El concepto 'minorías'. significados y usos. *Revista de Derecho a las Minorías*, 1, pp.1–17. doi:https://doi.org/10.22529/rdm.2019(1)01.

Rodríguez Zepeda, J. (2018). El peso de las palabras: libre expresión, no discriminación y discursos de odio. In: J. Rodríguez Zepeda and teresa González Luna Corvera, eds., *EL PREJUICIO Y LA PALABRA: LOS DERECHOS A LA LIBRE EXPRESIÓN Y A LA NO DISCRIMINACIÓN EN CONTRASTE*. México: Comisión Nacional para Prevenir la Discriminación, pp.27–75.

Rojano Aguilar, A., Moreno, R., Luis Miranda y Bustamante, W. (2021). Artículo: COMEII-21005. [online] Available at: https://www.riego.mx/congresos/comeii2021/files/ponencias/extenso/COMEII-21005.pdf.

Ruberg, B. y Ruelos, S. (2020). Data for queer lives: How LGBTQ gender and sexuality identities challenge norms of demographics. *Big Data & Society*, 7(1), p.205395172093328. doi:https://doi.org/10.1177/2053951720933286.

Salín, R.J. (2015). La diversidad sexo-genérica: Un punto de vista evolutivo. *Salud mental*, 38(2), pp.147–153. doi:https://doi.org/10.17711/sm.0185-3325.2015.020.

Sánchez Torrejón, B. (2021). La formación del profesorado de Educación Primaria en diversidad sexo-genérica. *Revista Electrónica Interuniversitaria de Formación del Profesorado*, 24(1), pp.253–266. doi:https://doi.org/10.6018/reifop.393781.

Sánchez, A., Cardona, L., Monroy, L., Arévalo, A. and Fulchiron, A. (2020). *Violencia, Derechos Humanos y Sexualidad*. Fundación Arcoíris por el Respeto a la Diversidad Sexual, A.C., pp.1–161.

Sandoval, J. (2019). ALGORITMOS DE APRENDIZAJE AUTOMÁTICO PARA ANÁLISIS Y PREDICCIÓN DE DATOS MACHINE LEARNING ALGORITHMS FOR DATA ANALYSIS AND PREDICTION A A Palabras clave Keyword. *Revista tecnológica ITCA FEPADE*, 11.

Schmidt, A. y Wiegand, M. (2017). A Survey on Hate Speech Detection using Natural Language Processing. In: *Fifth International Workshop on Natural Language Processing for Social Media*. Association for Computational Linguistics, pp.1–10.

Sirisha, J. y Babu Reddy, M. (2017). Unstructured Data: Various approaches for Storage, Extraction and Analysis. *International Journal of Computer Science and Information Security (IJCSIS)*, 15, pp.101–106.

Solís, P. (2017). *Discriminación estructural y desigualdad social. Con casos ilustrativos para jóvenes indígenas, mujeres y personas con discapacidad*. Primera edición ed. [online] Ciudad de México, México:

Consejo Nacional para Prevenir la Discriminación (conapred). Available at: https://repositorio.dpe.gob.ec/bitstream/39000/2084/1/CONAPRED-066.pdf [Accessed 31 Oct. 2022].

Ștefăniță, O. y Buf, D.-M. (2021). v. *Romanian Journal of Communication and Public Relations*, 23(1), p.47. doi:https://doi.org/10.21018/rjcpr.2021.1.322.

Tabares Higuita, L. (2018). ANÁLISIS DEL DISCURSO VIOLENTO Y DE ODIO EN DOS GRUPOS DE FACEBOOK CONTRA LA CANDIDATURA DE RODRIGO LONDOÑO 'TIMOCHENKO' A LA PRESIDENCIA DE COLOMBIA. *index.comunicación*, 8(2444-3239), pp.158–183.

Tellez, E.S., Miranda-Jiménez, S., Graff, M., Moctezuma, D., Suárez, R.R. y Siordia, O.S. (2017). A simple approach to multilingual polarity classification in Twitter. *Pattern Recognition Letters*, 94, pp.68–74. doi:https://doi.org/10.1016/j.patrec.2017.05.024.

Toivonen, T., Heikinheimo, V., Fink, C., Hausmann, A., Hiippala, T., Järv, O., Tenkanen, H. y Di Minin, E. (2019). Social media data for conservation science: A methodological overview. *Biological Conservation*, [online] 233, pp.298–315. doi:https://doi.org/10.1016/j.biocon.2019.01.023.

Visible (n.d.). *Visible | Reporta incidentes de violencia contra las personas LGBTQ+ en México.* [online] Visible. Available at: https://visible.lgbt/?y=2022#estadisticas [Accessed 21 May 2023].

Zhang, Z., Robinson, D. y Tepper, J. (2018). Detecting Hate Speech on Twitter Using a Convolution-GRU Based Deep Neural Network. *The Semantic Web*, [online] 10843, pp.745–760. doi:https://doi.org/10.1007/978-3-319-93417-4_48.

Zimmerman, S., Kruschwitz, U. y Fox, C. (2018). Improving hate speech detection with deep learning ensembles. In: *n Proceedings of the Eleventh International Conference on Language Resources and Evaluation*. Japón.