UNIVERSIDAD AUTÓNOMA DE QUERÉTARO
FACULTAD DE LENGUAS Y LETRAS

Discursos mexicanos de migración en redes sociales. Una visión desde la Lingüística de corpus y Computacional

TESIS

Que como parte de los requisitos para obtener el Grado de
Doctora en Lingüística

Presenta
Ana Ruth Sánchez Barrera

Dirigido por
Ignacio Rodríguez Sánchez
Antonio Rico Sulayes

Querétaro, Qro a. 4 de octubre de 2024

UNIVERSIDAD AUTÓNOMA DE QUERÉTARO
FACULTAD DE LENGUAS Y LETRAS

Discursos mexicanos de migración en redes sociales. Una visión desde la Lingüística de corpus y Computacional

TESIS

Que como parte de los requisitos para obtener el Grado de

Doctora en Lingüística

Presenta

Ana Ruth Sánchez Barrera

Dirigido por

Ignacio Rodríguez Sánchez

Antonio Rico Sulayes

Ignacio Rodríguez Sánchez, Presidente

Antonio Rico Sulayes, Secretario

Eva P. Velásquez Upegui, Vocal

Antonio Reyes Pérez, Suplente

Rafael Saldívar Arreola, Suplente

Centro Universitario, Querétaro, Qro.
Fecha de aprobación por el Consejo Universitario (mes y año)
México

.

Ya no me sorprende que la lucha contra el terrorismo se fije tanto en las redes sociales; las redes radicalizan como una visión del infierno, como el pentotal sódico o como la telepatía, que es el superpoder que más destruye a los superhéroes porque nadie soporta la crudeza de un monólogo interior, lo que tu vecino piensa de ti, lo que tu país piensa de los inmigrantes…

Cambiar de idea, Aixa de la Cruz

En palabras de Jaques Laccarrière, Herodóto se esforzó por derribar los prejuicios de sus compatriotas griegos, enseñándoles que la línea divisoria entre la barbarie y la civilización nunca es una frontera geográfica entre diferentes países, sino una frontera moral dentro de cada pueblo; es más, dentro de cada individuo.

El infinito en un junco, Irene Vallejo

Agradecimientos

Al Consejo Nacional de Humanidades, Ciencias y Tecnologías (Conahcyt) y la Universidad Autónoma de Querétaro (UAQ) por el apoyo económico que me permitió dedicarme de lleno al desarrollo de mi tesis doctoral.

A mis directores de tesis quiero agradecerles la generosidad con la que me compartieron tiempo y conocimientos para que este proyecto saliera adelante. Al Dr. Ignacio Rodríguez Sánchez le agradezco su constante apoyo en mi desarrollo académico; termino esta etapa comprobando su calidad humana y su compromiso por el bienestar de sus estudiantes. Al Dr. Antonio Rico Sulayes le agradezo su rigor académico y su alta exigencia; su constante presión por la excelencia, acompañada siempre con su dedicación y compromiso, es ahora una enseñanza para mi labor como investigadora.

A mis sinodales, la Dra. Eva P. Velásquez Upegui, el Dr. Antonio Reyes Pérez, y el Dr. Rafael Saldívar Arreola, les expreso mi más sincero agradecimiento por su tiempo y dedicación en las revisiones semestrales y a la lectura de mi tesis. Sus comentarios y observaciones fueron de gran utilidad para mejorar la calidad de mi trabajo y me permitieron tener una perspectiva más amplia de mi investigación.

A Nimsi Arroyo Flores que generosamente me orientó sobre las migraciones centroamericanas y caribeñas en México.

A David que me apoyó y sacó fuerzas aun en su momento más difícil.

Tabla de contenido

Índice de Ilustraciones

Índice de Figuras

Índice de Tablas

Resumen

Este proyecto abordó, desde una perspectiva crítica y multidisciplinar, el discurso en redes sociales sobre migraciones centroamericanas y caribeñas en su ingreso, transcurso o estadía en México. El análisis se centró en el discurso relfejado en las dos redes sociales (Twitter y YouTube) entre los años 2018 y 2021, particularmente, a partir de las llamadas caravanas migrantes que crecieron en densidad a finales de 2018. El objetivo principal fue realizar un estudio guiado por datos en torno a los comentarios generados en estas redes sociales sobre estas migraciones.

De Twitter se elaboraron dos corpus: uno para un análisis de sentimientos y otro, junto con el corpus de YouTube, para un análisis crítico del discurso desde la lingüística de corpus.

El análisis de sentimientos tenía el doble objetivo de detectar mensajes de odio contra migrantes y elementos lingüísticos representativos de estos discursos, particularmente n-gramas. Como resultado, se obtuvo un detector de mensajes competitivo ante otros trabajos similares en habla hispana).

El segundo análisis implicó la clasificación de elementos en campos semánticos y la selección de tres elementos del campo de LUGAR para una revisión de sus patrones de ocurrencia en corpus de redes sociales y en un corpus de referencia. Se buscaron asociaciones semánticas que dieran cuenta de activación léxica y evidencia del refuerzo de patrones lingüísticos asociados a una postura ideológica.

Finalmente, se revisaron desde la semántica cognitiva las propiedades deícticas de los n-gramas seleccionados, con el objetivo de responder cómo el léxico de los n-gramas está preparado para ocurrir en asociaciones semánticas con un sentido ideológico nacionalista.

1 Introducción

Durante el fin del siglo XX y las tres décadas transcurridas de este, buena parte de los conflictos sociales se ha centrado en una lucha por el reconocimiento en que distintas identidades de género, raza, etnicidad y nacionalidad –por mencionar algunas– toman protagonismo frente a cuestiones de desigualdad económica todavía irresolutas (Butler & Fraser, 2016; Fraser, 2000). Es este un momento en el que el reclamo de justicia redistributiva y de reconocimiento, en términos de Nancy Fraser, no generan la misma movilidad política, aunque los problemas sociales actualmente experimentados estén compuestos por ambas dimensiones. En este contexto, la migración provocada por desequilibrios redistributivos es recibida, en diferentes países de destino, en medio de discursos polémicos que, a muy grandes rasgos, se debaten entre la aceptación o el rechazo de la población migrante (Taylor, 2019; van Dijk, 2016) e influyen en su reconocimiento.

La investigación presente consta del discurso sobre migraciones provenientes de Centroamérica y el Caribe que tuvieron lugar en los años 2018 a 2021. Precisamente estos años marcan un parteaguas en la forma de ingreso pues hubo un notable surgimiento de caravanas de migrantes como método a través del cual se elige entrar a México. Esta forma particular de movimiento migratorio se distingue por su intrincado proceso de coordinar la movilización de grandes grupos de personas de diversos sectores de la sociedad civil y/o grupos de interés específicos. Un punto focal esencial de estas caravanas es establecer un mayor nivel de visibilidad y, por lo tanto, ofrecer cierto grado de protección a las personas que emprenden el viaje migratorio dentro de ellas (Reyes Vázquez & Barrios de la O, 2019). A pesar de que este método de movimiento de personas organizadas no es del todo nuevo, ya que los casos de tales movimientos se remontan a 1999, la escala y el alcance sin precedentes que había alcanzado en noviembre de 2018 marcan un cambio significativo en su evolución, (Huerta & McLean, 2019; Reyes Vázquez & Barrios de la O, 2019). Varela y McLean (2019) describen dos diferencias importantes de las caravanas sucedidas a partir del 2018 con respecto a las de sus antecesoras. En primer lugar, su gestación. Si bien en un inicio nacían una vez que las personas ingresaban a México por su frontera sur y por

iniciativa de grupos de la sociedad civil laicos y religiosos con el doble propósito tanto de proteger a los migrantes, como de llamar la atención de la sociedad mexicana de la necesidad de seguridad para tales caravanas, a partir del 2018 nacen desde el momento de la organización del viaje, en los lugares de origen de los migrantes, y por iniciativa de ellos mismos. En segundo lugar, a la diferencia del volumen que ya se nombraron se sumaba el hecho de su composición, pues estas comenzaron a ser constituidas por familias, particularmente con un gran número de mujeres y menores de edad.

Estas nuevas características han sido un hecho que ha impactado la vida social, en su sentido más amplio, al punto de que dichos discursos son parte del entorno académico, jurídico, periodístico, y civil. Por un lado, actores desde la organización civil y desde la academia nombran a estas formas de organización como una expresión de lucha migrante; por otro, han sido señaladas como una forma de criminalizar el hecho de migrar, a saber "la idea de caravanizar la transmigración para estos actores alude a una forma concreta de flujo que hay desarticular" (Huerta & McLean, 2019, p. 174).

Ante este escenarioel, el propósito de esta tesis fue el de analizar el interés que la migración centroamericana y caribeña ha suscitado en redes sociales a raíz de estas formas de migración. En este estudio, que se emprendió con la intención de mantener una perspectiva crítica, se recogieron datos del entorno digital en tanto este contexto implica formas masivas de comunicarse que pueden acercar al discurso de personas fuera de las élites. Los intercambios comunicativos en estas redes tienen características que las sitúan a medio camino entre lo público y lo privado (Claridge, 2007); o como un híbrido entre lo personal y la comunicación de masas (Marcoccia, 2004) de modo que pueden jugar un rol en establecer, estabilizar o cambiar la opinión pública (Tanner, 2001). En ese sentido, el entorno digital del que son partes las redes sociales ha sido llamado ágora electrónica (Claridge 2007), ágora digital (Damiris & Wild, 1997) o aldea global (Foster, 1996).

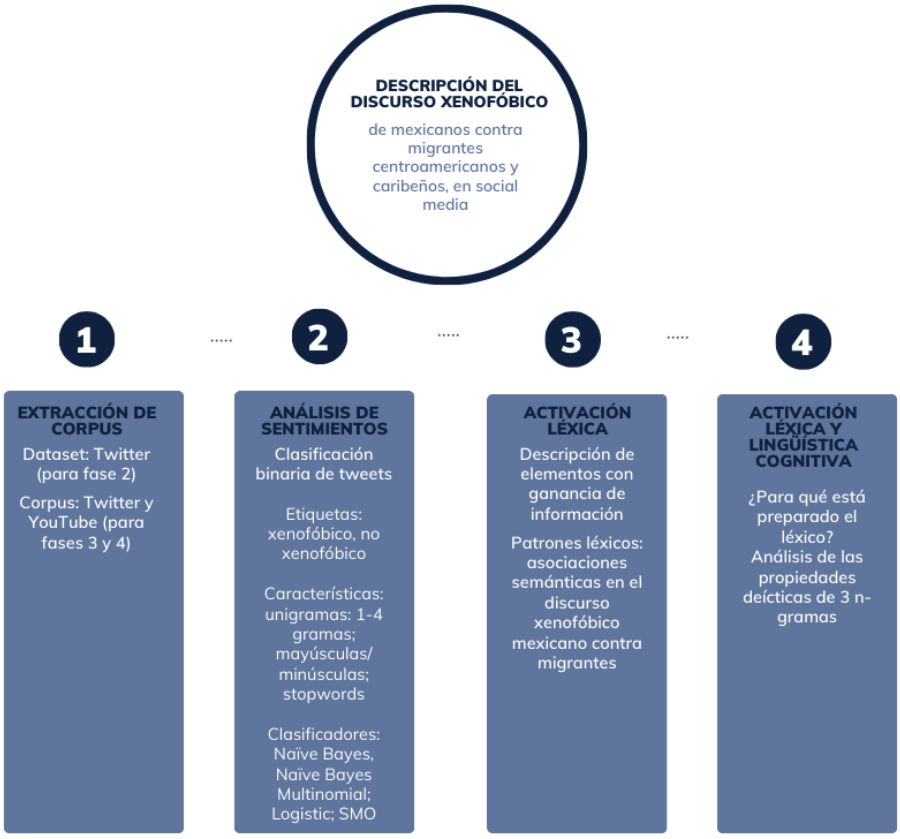
Así es que se conceptualiza a las redes sociales como espacios que multiplican discursos, ideologías, y crean estados de opinión (Fowler, 1991). Son espacios más o menos democráticos en función de que permiten la diversidad de voces. "Cualquiera" puede hablar ahí, naturalmente siempre y cuando los marcos legales lo permitan y cuente con las

condiciones materiales mínimas para ello. Las participaciones que se hacen en ellas, quienes cuentan con la fortuna de hacerlo, son parte de las sociedades actuales de maneras políticas, económicas y sociales. Ello no sucede a partir de una única participación, sino por su acumulación; además dicha acumulación no es heterogénea, sino plural.

Así pues, esta investigación consistió en un intento de desentrañar la pluralidad de opiniones vertidas en dos redes sociales sobre las migraciones centroamericanas y caribeñas en su paso o estadía por México. De esta pluralidad, fue de particular interés el discurso xenofóbico. La pluralidad a la que se hace alusión es a la diversidad de voces que han participado en la discusión pública, viendo a las redes sociales como tal especie de *ágora* (Damiris & Wild, 1997), cuyo cúmulo de opiniones forman un discurso xenofóbico mediante una tendencia hacia el rechazo, la reproducción de estereotipos, o la exaltación de la violencia hacia tales migrantes. Esta es una concepción de discurso que descansa en la concepción de Fairclough (2010), es decir, a formas semióticas usadas para construir *aspectos* del mundo, como pueden ser las percepciones hacia grupos de personas; esa construcción sucede, además, mediante la acumulación de dichas formas semióticas.

En otras palabras, se presenta un análisis de corte multidisciplinar cuyo fin último fue la descripción lingüística de lo que acabamos de llamar discurso xenofóbico hacia migrantes centroamericanos y caribeños. Es una propuesta que se desarrolló en 4 fases correspondientes a las disciplinas y sus respectivos métodos que conforman esta propuesta (Ilustración 1).

Ilustración 1. Planteamiento y fases de la investigación.



Tales fases obedecieron al hecho de que se buscó explicar los discursos de migración por el efecto acumulativo de la activación léxica: en estricto sentido las redes sociales hacen esa acumulación sobre cualquier tema que pasa por ellas. Así pues, estas fases descansan en los principios teóricos, principalmente, del análisis crítico del discurso y de la teoría de la Activación Léxica. Si bien las particularidades de la primera de las disciplinas anteriores no figura en el esquema de las fases, es porque esta fue tratada como un eje transversal, de modo que permitiera dar cuenta de cómo se forman y refuerzan las ideas xenofóbicas en la red. De esta manera, esta diversidad de enfoques teóricos y métodos fue en concordancia con los lineamientos de este tipo de estudio, que no se limita a un solo método sino que aboga por la multidisciplinariedad para explicar las diferentes dimensiones de las que busca dar cuenta, a saber, discurso, cognición y sociedad (van Dijk, 2016).

Así pues, lo que el lector podrá encontrar en el capítulo **¡Error! No se encuentra el origen de la referencia.** será el Marco Teórico que sustenta la propuesta de análisis. Dentro de este capítulo, en **¡Error! No se encuentra el origen de la referencia.** se explica esta técnica, que es una práctica de Procesamiento de Lenguaje Natural (PLN) utilizada tanto para estudios de mercado como para detectar lenguaje llamado tóxico o de odio hacia personas o grupos en redes sociales, tal como se trató en este trabajo. Lo correspondiente a la concepción crítica se puede consultar en la sección **¡Error! No se encuentra el origen de la referencia.** Análisis crítico del discurso, donde se puede ver cómo este enfoque examina la forma en que el discurso promueve o justifica relaciones de poder, y cómo se opone a ellas. Entre tanto, en la sección **¡Error! No se encuentra el origen de la referencia.** Lingüística de corpus y Activación Léxica y en la sección **¡Error! No se encuentra el origen de la referencia.** Activación Léxica y Discursos Ideológicos se puede consultar la propuesta teórica central que se sostuvo en este trabajo, según la cual un elemento restringe la ocurrencia de los que lo acompañan. Esta condición, como se verá, no se limita a elementos léxicos, sino que se anida hasta, incluso, lograr la coherencia textual y, según se siguió en este trabajo, también tiene repercusiones en la intertexualidad. Para cerrar un marco teórico desde un enfoque funcionalista y experiencial, se echa mano de la Lingüística cognitiva (sección **¡Error! No se encuentra el origen de la referencia.**) para explicar cómo las propiedades léxicas participan para la formación de los patrones encontrados.

El capítulo **¡Error! No se encuentra el origen de la referencia.** consiste en una descripción del problema abordado. Ahí se planteó tanto la interrogante de cómo es que los discursos ideológicos se pueden sustentar en factores cognitivos y sociales, como la necesidad de identificar los patrones y elementos comunes en discursos xenofóbicos con el fin de comprender cómo tales discursos se construyen en un entorno como el de las redes sociales.

Entre tanto, en el capítulo **¡Error! No se encuentra el origen de la referencia.** se encuentra una revisión de literatura. Esta revisión solo se centró en los análisis de sentimientos en torno a discursos xenofóbicos o racistas, y en el análisis del discurso crítico del mismo tema abordados desde la Lingüística de corpus. Ello obedeció a que la propuesta teórica y

metodológica presentada en este trabajo valora las conclusiones obtenidas a partir del comportamiento de los datos en muestras representativas del lenguaje.

Entre tanto, los objetivos de esta investigación, así como las preguntas y las hipótesis a las mismas se pueden consultar en el capítulo **¡Error! No se encuentra el origen de la referencia.**.

En concordancia con los objetivos, y con las cuatro disciplinas mencionadas en el Marco Teórico, en el capítulo **¡Error! No se encuentra el origen de la referencia.** se puede consultar la propuesta metodológica. Particularmente, el diseño de cada uno de los corpus utilizados se puede revisar en el subcapítulo **¡Error! No se encuentra el origen de la referencia.**; el tratamiento de los mismo para el Análisis de sentimientos en **¡Error! No se encuentra el origen de la referencia.**; y los lineamientos para los análisis más cualitativos de esta investigación en **¡Error! No se encuentra el origen de la referencia.**.

Los resultados se encuentran en el capítulo **¡Error! No se encuentra el origen de la referencia.**. En primer lugar se describirán los resultados del análisis de sentimientos en **¡Error! No se encuentra el origen de la referencia.** para luego contextualizarlos en comparación con otros trabajos de clasificación de sentimientos xenofóbicos hechos sobre datos en español en **¡Error! No se encuentra el origen de la referencia.**. Posteriormente se encuentra una descripción lingüística de los corpus que comienza con la explicación de la generalidad de los patrones encontrados (**¡Error! No se encuentra el origen de la referencia.**), y continúa con explicaciones cada vez más particulares y cualitativas en tanto aborda las asociaciones semánticas encontradas (**¡Error! No se encuentra el origen de la referencia.**) y cómo las particularidades léxicas de los elementos componentes de dichas asociaciones permiten, paradójicamente, la coherencia intertextual del discurso xenofóbico en redes sociales (**¡Error! No se encuentra el origen de la referencia.**).

Entre tanto, en el capítulo **¡Error! No se encuentra el origen de la referencia.** se discute la pertinencia de este análisis. En primer lugar, se encontrará una revisión de lo que un análisis como éste, que mezcla técnicas automáticas y estadísticas abona a indagaciones cualitativas de la lengua, permita explicar. En segundo lugar, se discuten tres evidencias encontradas a favor de la hipótesis general.

Finalmente, en conclusiones, el lector podrá encontra una síntesis de la propuesta teórica y metodológica, de los hallazgos, y de cómo estos se sitúan en relación con el estado de la cuestión, así como de los posibles caminos a seguir a partir de esta investigación.

2 Marco Teórico

El presente proyecto busca realizar un aporte a los estudios del discurso sobre migración y, particularmente, al discurso mexicano sucedido en redes sociales sobre migrantes centroamericanos y caribeños. Si bien de corte multidisciplinar, el enfoque que se propone es principalmente lingüístico porque persigue la explicación de patrones lingüísticos que expliquen la constitución de un discurso xenofóbico que no es emitido por las élites, que es lo que tiende a denunciarse desde los Estudios Críticos del Discurso. Se les denomina *estudios* porque refieren a varios análisis hechos desde diversas disciplinas para explicar cómo, desde el orden de lo discursivo, se construyen jerarquías sociales, se justifican las existentes o, más recientemente, se oponen a ella (Londoño Zapata, 2013). El interés de estos análisis ha sido mayoritariamente sobre el discurso de las élites, lo que van Dijk llama los actores de las tres P*: Políticos, Periodistas y Profesores* (Londoño Zapata, 2013; van Dijk, 2016).

Por lo anterior, en este capítulo se encuentra la descripción del panorama multidisciplinar desde el que se aborda el así llamado discurso xenofóbico y en el cual se inscribe esta propuesta de análisis lingüístico.

En tanto se presenta un análisis orientado por datos, el marco teórico está circunscrito a las técnicas de clasificación automática de texto utilizadas en el procesamiento de lenguaje

natural, con el propósito de detectar los mensajes que, en uno de los corpus utilizados (Dalal & Zaveri, 2011), contengan proposiciones antimigrantes. Así, se verá en esta sección cómo desde esta disciplina las características de un mensaje pueden ser utilizadas para la detección de otros parecidos.

A pesar de que este último proceso suele ser usado para probar varios experimentos en los trabajos de clasificación automática, la exploración del discurso comienza, en términos lingüísticos y particularmente desde la Lingüística de corpus, desde aquellas características que el proceso empleado señale como "propias" del discurso que interesa describir. Es importante establecer, sin embargo, que las técnicas del PLN usadas no distinguen, en términos humanos, un discurso de otro, en cambio lo hacen a partir de la diferencia de presencia y frecuencia de dichas características. Así pues, a semejanza de algunas técnicas de la Lingüística de corpus, un texto, un discurso circunscrito a sus características extralingüísticas puede ser explicado por la ocurrencia (presencia, ausencia, frecuencia) de sus elementos lingüísticos, así como de las constantes interacciones en que estos sucedan.

La segunda parte del análisis es la descripción del discurso delimitado como xenofóbico. En ese sentido, se utilizan dos propuestas teóricas: nuevamente la Lingüística de corpus, esta vez con la propuesta teórica de Activación Léxica, y también desde el Análisis Crítico del Discurso. A grandes rasgos, explicados detalladamente más adelante, interesa explicar cómo desde la propuesta de Hoey (2005) es posible resolver preguntas propias del enfoque del ACD, a saber: ¿qué dice la coocurrencia de tokens de un discurso posicionado ideológicamente?, ¿qué dice del hablante que lo emite, qué de la cultura de la que forma parte?

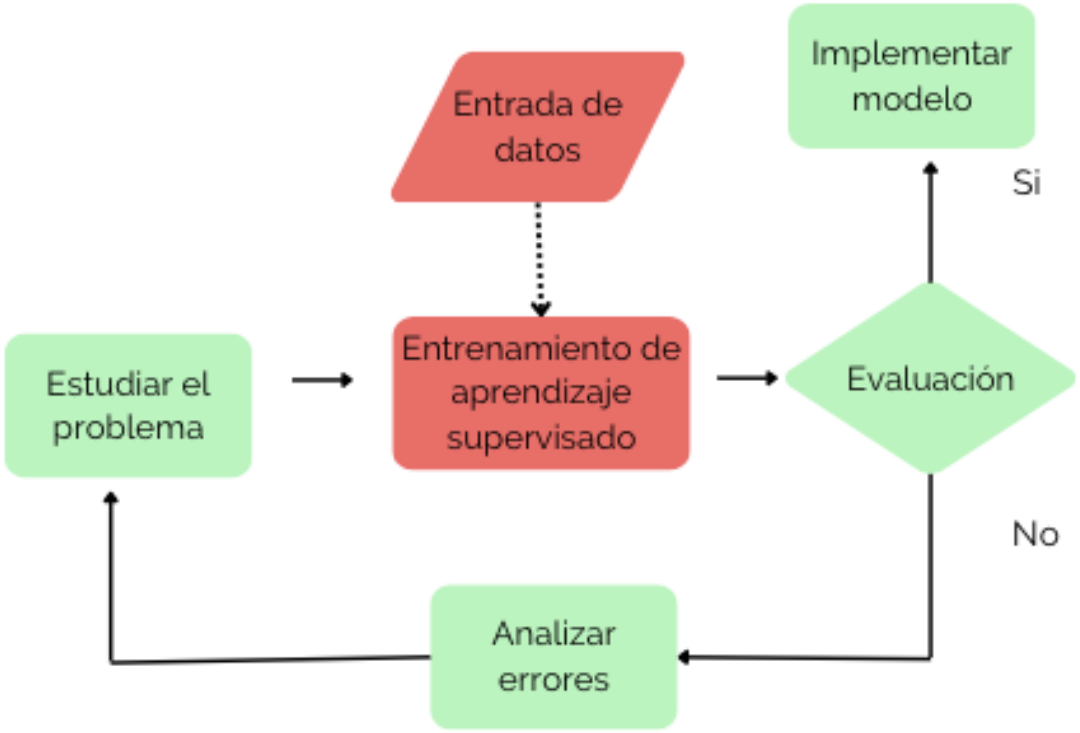
El último sitio del panorama multidisciplinar en el que sustentaremos las explicaciones en este trabajo es la Lingüística Cognitiva. En la última sección de este Marco Teórico se verá cómo este enfoque es utilizado con el propósito de describir algunos patrones léxicos encontrados en nuestros resultados, para ver cómo tales patrones pueden estar sustentados en operaciones mentales de los hablantes que las emiten.

2.1 Clasificación automática de textos

La lingüística computacional es una disciplina que investiga cómo funciona la generación y comprensión del lenguaje humano con el propósito de diseñar sistemas de inteligencia lingüística. Dentro de tal área, el Procesamiento de Lenguaje Natural (PLN) realiza una serie de tareas que responden a necesidades propias de la sociedad de la información. En este contexto, se entiende por Lenguaje Natural la lengua en uso. En la medida en que esta es cambiante en diferentes contextos y tiempos, las tareas que se realicen desde el PLN deben asumir dicha posibilidad de cambio (Bengfort et al., 2018). Para ello aplica herramientas propias del Machine Learning; es decir, echa mano de modelos que reconocen y aprenden patrones -para nuestro interés, lingüísticos- a partir de una base de datos. Este aprendizaje es la base de soluciones que se usan diariamente, como es el caso de motores de búsqueda, chatbots, reconocimiento de voz, clasificación de textos como reconocimiento de spam, reconocimiento de autoría, detección de mensajes con ciertas características en foros o redes sociales.

En ese sentido, es de utilidad para mejorar las herramientas que mencionábamos, propias de la sociedad de la información. Por ejemplo, en la implementación de soluciones referentes a la detección de spam en las bandejas de correos electrónico, Géron (2022) esquematiza la solución del problema desde el aprendizaje automático como en la Ilustración 2. Dicho proceso de mejora es un enfoque guiado por datos, de hecho, este proceso aprende a partir de un conjunto de datos entrenados. En este caso, desde la automatización se buscar *aprender* qué elementos de un correo electrónico se relacionan con dos *clases*, a saber, un correo spam de uno no spam, para, posteriormente, llevar a cabo la segregación de una y otra clase.

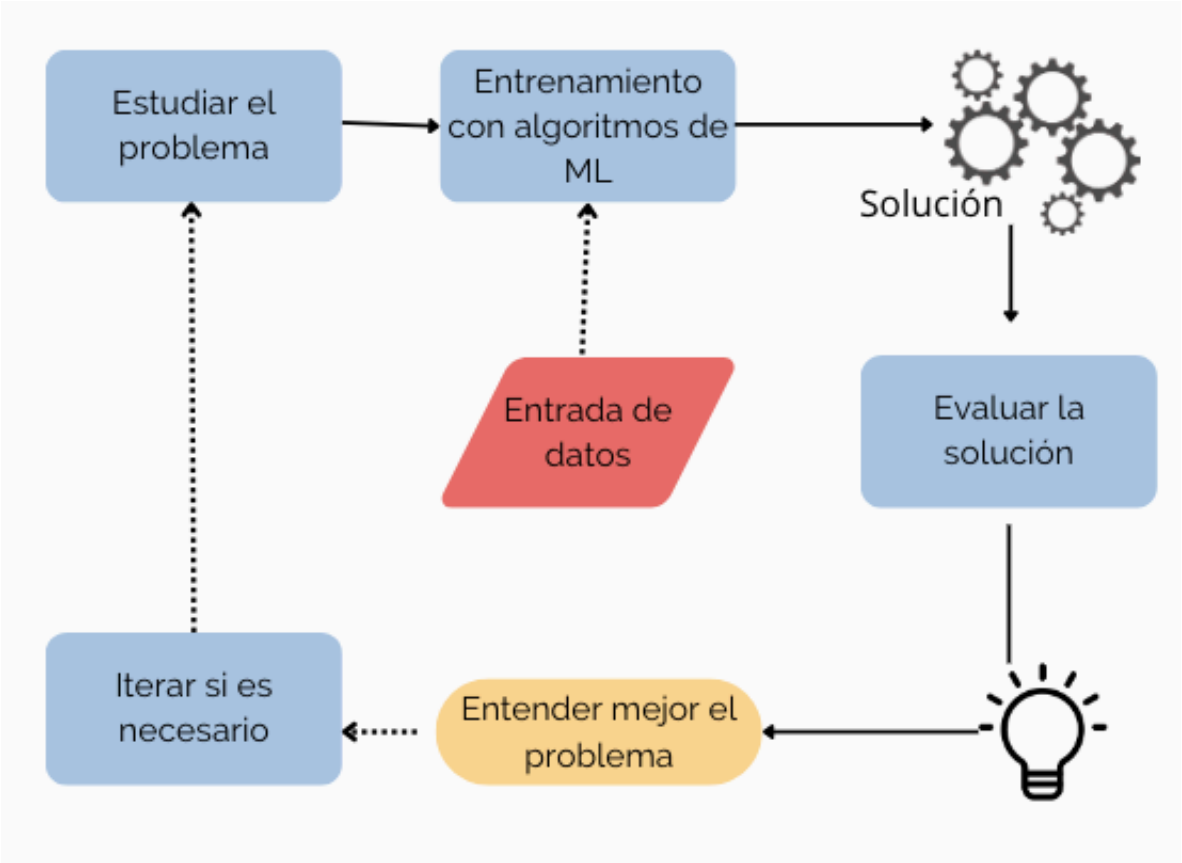
Ilustración 2. Enfoque del aprendizaje automático



Nota: Géron, A. (2022). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media, Inc.

Otra posibilidad es que estas técnicas ayuden al investigador a comprender el fenómeno para el que está entrenando a su sistema; precisamente en función de que estas son técnicas de aprendizaje. Retomando el ejemplo del spam, lo que el entrenamiento dirá es qué palabras o conjuntos de ellas son mejores predictores. Este proceso es llamado *data mining* (Géron, 2022) o minería de datos, en español (Ilustración 3). Gracias a la posibilidad de obtener estos datos "aprendidos" es que este análisis de discurso tiene como punto de partida una clasificación automática y supervisada de datos de redes sociales para comprender el discurso xenofóbico hacia migrantes centroamericanos y caribeños.

Ilustración 3. Usar aprendizaje automático para comprender un fenómeno



Nota: Géron, A. (2022). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media, Inc.

Se le llama "supervisada" pues en un proceso como este, el investigador interviene en el entrenamiento de los datos desde dos posibles enfoques

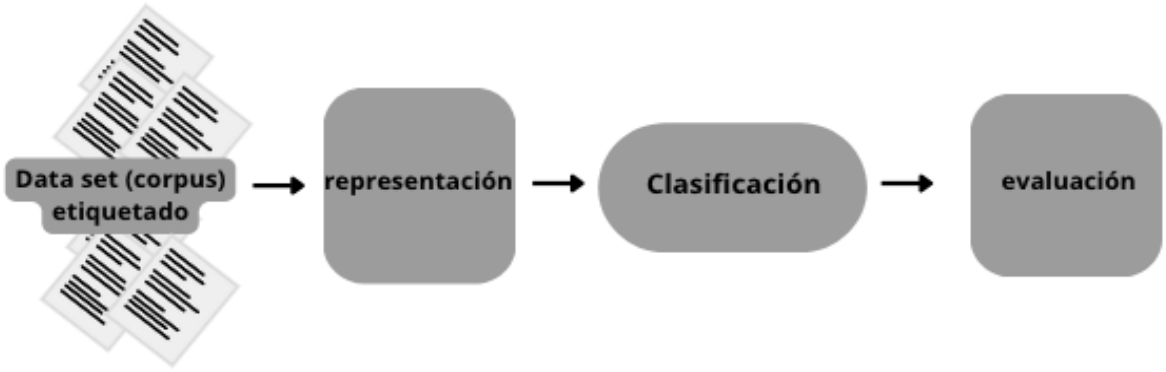
- de aprendizaje supervisado, esto es, se provee al algoritmo con un set de datos previamente etiquetados con el aprendizaje deseado,
- aprendizaje semisupervisado en la medida en que se provee un set de datos con etiquetas parciales,

Un tercer enfoque es posible ahí donde se provee el set de datos sin etiquetas. Se hablaría entonces de un enfoque no supervisado.

Entre los aprendizajes supervisados, la clasificación es una tarea típica. Particularmente, la Clasificación Automática de Textos es un procedimiento mediante el cual se asigna de manera automatizada un documento a una categoría entre dos o más previamente establecida(Dalal & Zaveri, 2011). Un documento es la unidad de análisis que se delimita

por aquello que se quiera analizar. Un documento puede ser un artículo, un libro, un estado de Facebook, un comentario de YouTube, un tweet. Definir con qué tipo de documentos se trabajará, así como su obtención para la formación de un conjunto de datos (al que en la Lingüística le llamamos corpus), será el primero de una serie pasos cuyo último objetivo es que cada documento sea clasificado en algún subconjunto (Ilustración 4).

Ilustración 4. Diagrama de flujo del proceso de Clasificación Automática de Textos



La lectura de los patrones por aprender dependerá de la vectorización. Esto es, la transformación de los datos textuales de cada documento a representaciones numéricas. En este proceso, un documento se convierte en una *instancia* (a su vez, el algoritmo de aprendizaje supervisado opera sobre cada instancia), esto es, un vector numérico de propiedades distinguibles, también llamadas *características* (Bengfort et al., 2018).

Una de las formas de representación de características es la *Bolsa de palabras* (Bag of Words, BoW). En este tipo de representación, el vector que se genere será a partir de la frecuencia de cada término presente en el conjunto de datos. La unidad es cada token o unigrama. Pero también ocurre que se selecciona a un conjunto de palabras como unidad básica; si tal conjunto está compuesto por dos palabras, se llamará bigrama; por tres, trigrama y así sucesivamente. De la oración (1), ejemplificamos en (2) el primer unigrama, en (3) el primer bigrama y en (4) el primer trigrama.

(1) Tendremos un seguimiento de los abusos policíacos a la población migrante

(2) Tendremos

(3) Tendremos un

(4) Tendremos un seguimiento

Para la etapa de clasificación se utilizarán diferentes algoritmos (o clasificadores). Un experimento es, de hecho, la selección en cada instancia de un conjunto de características que, una vez vectorizadas, son clasificadas por algoritmos, y así a cada instancia se le asigna una clase. En un proyecto de clasificación, como el que aquí presentamos, se ponen a prueba varios conjuntos de características vectorizadas con varios clasificadores en un análisis de sentimientos. Un análisis de sentimientos es un tipo de clasificación que determinará si un documento tiene un sentimiento, al que le corresponde una clase *x*, u otra, naturalmente correspondiente a otra clase *y*. Tomando ejemplos del tema que atañe en este trabajo, el sentimiento que interesa es el odio hacia migrantes, representado bajo la clase uno esto es 'xenofobia' (X), y la clase dos es 'no xenofobia' (Y).

Los algoritmos clasificadores pueden ser más o menos complejos al tomar en cuenta variables en la relación entre las características del data set. Con una relación relativamente simple están los modelos bayesianos ingenuos (o Naïve Bayes). Están basados en la técnica estadística del teorema de Bayes (Merino García, 2012; Roman, 2019), y tratan a las variables predictoras como independientes entre sí. En términos generales quiere decir que asumen que ninguno de los rasgos presentes en el corpus tiene una relación entre sí. Más específicamente quiere decir que cada característica cuenta para asignar una etiqueta a cada documento o instancia. Para ello, en primer lugar se calcula la probabilidad a priori de cada etiqueta, esto se logra mediante la frecuencia que tiene cada etiqueta en el set de entrenamiento. En segundo lugar, esta probabilidad a priori se combina con la contribución de cada característica para llegar a una probabilidad estimada. La contribución, a su vez, se entiende como un "voto en contra" de la característica que no ocurre muy frecuentemente (Bird et al., 2009).

En un caso práctico, por ejemplo, ante un conjunto de datos de noticias que deba ser clasificado en función de su tema, dicho clasificador actuaría de la siguiente manera. En primer lugar, se calcularía la probabilidad a priori de cada tema mediante el conteo del

número de noticias de cada tema en el conjunto de entrenamiento. En segundo lugar, se calcularía la contribución de cada característica para cada tema, contando el número de veces que aparece cada característica en las noticias de cada tema.

Los algoritmos basados en la Máquina de Soporte Vectorial (SVM), entre los que se encuentra el clasificador Sequential Minimal Optimization (SMO), y que se utiliza en los experimentos de este trabajo como puede verse en **¡Error! No se encuentra el origen de la referencia.**, pueden utilizarse tanto para regresión como para clasificación. Dentro de esta última son especialmente útiles en problemas de clasificación binaria (Vasquez et al., 2013). Su funcionamiento se sustenta en la construcción de un hiperplano en el que puedan separarse los datos de acuerdo con su clase en un contexto no paramétrico. Por ejemplo, en el contexto de la clasificación de textos, un algoritmo SVM podría utilizarse para distinguir entre textos de noticias y textos de opinión. En este caso, el hiperplano de separación podría representar la frontera entre las dos clases de textos. Los textos que se encuentren más cerca del hiperplano serían los que sean más difíciles de clasificar, ya que sus características son más ambiguas.

2.1.1 Evaluadores

Así, cada modelo está compuesto por la selección de las características, el método de extracción de estas y el algoritmo que le asigna una clase al documento. Cada modelo es evaluado por métricas con el propósito de mejorar los resultados y, finalmente, encontrar qué combinación de extracción de características y algoritmo clasificador es el mejor.

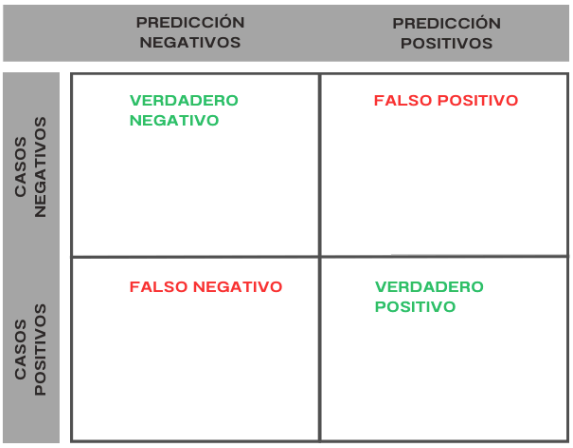
Las métricas evaluadoras más comunes en la literatura son precisión, exactitud, exhaustividad y valor F. El valor de cada una de ellas depende del peso que den a los documentos clasificados; para comprender cómo se pondera tal peso, se debe conocer en primer lugar las cuatro categorías posibles dependiendo de si cada documento es clasificado correcta o incorrectamente:

- Habrá un **verdadero positivo** cuando el modelo identifique un documento con la clase de interés (a la que se le llama clase positiva) y verdaderamente el documento pertenezca a tal clase,

- Será un **falso positivo** cuando el modelo identifique un documento con la clase de interés, pero en realidad no corresponda a esta,
- se llamará **verdadero negativo** cuando el modelo identifique un documento con la clase de oposición (a la que se llama clase negativa) y verdaderamente lo sea,
- finalmente, habrá un **falso negativo** cuando el modelo identifique como negativo a un documento positivo.

La distribución que obtengan estas categorías será representada gráficamente en la matriz de confusión (Ilustración 5). Si bien el mejor modelo será el que obtenga los valores más altos en la diagonal que expresa los valores verdaderos tanto negativos como positivos, es importante tomar en cuenta los objetivos que se busquen así como de la cantidad de datos con que se cuente para decidir cuál de los modelos es el conveniente de acuerdo con los valores de las diferentes métricas evaluadoras.

Ilustración 5. Modelo de matriz de confusión



Entre los evaluadores más utilizados se encuentra la exactitud (accuracy), que se refiere a la cantidad de instancias clasificadas correctamente sobre el número total de instancias a clasificar (*Figura 1*).

Figura 1. Fórmula Exactitud (Accuracy)

$$exactitud = \frac{VP + VN}{VP + VN + FP + FN}$$

La precisión (*precision)* es una medida ideal para cuando la cantidad de falsos positivos tiene una repercusión importante en la clasificación. Esta medida se refiere a la ponderación de los verdaderos positivos sobre la cantidad total de lo que se dijo que era positivo (Figura 2). Continuando con el ejemplo de lenguaje de odio en redes sociales, si el modelo clasifica 80 tweets como positivos sobre la migración, pero sólo 50 de ellos realmente lo son, la precisión sería de 50/80 = 0.625.

Figura 2. Fórmula Precisión (Precision)

$$Precisión = \frac{VP}{VP + FP}$$

Contrario a la métrica anterior, la exhaustividad (*recall*) compara la cantidad de casos positivos clasificados como tales, sobre todo lo que realmente era positivo (Figura 3). Es decir, que es una medida que se usa cuando hay un costo alto asociado con un falso negativo. Así pues, si de 100 tweets positivos sobre la migración, nuestro modelo solo clasifica correctamente 60, la exhaustividad es 60/100 = 0.6.

Figura 3. Fórmula exhaustividad (recall)

$$Exhaustividad = \frac{VP}{VP + FN}$$

Entre tanto, el valor F1 es una métrica de balance entre la precisión y la exhaustividad, esto es (Figura 4). Por ejemplo, si tenemos una precisión de 0.625 y una exhaustividad de 0.6 para un modelo, el puntaje F1 sería de 2*(0.625*0.6)/(0.625+0.6) = 0.612.

Figura 4. Fórmla del valor F1 (F1-score)

$$F1 = \frac{precisión * exhaustividad}{precisión + exhaustividad}$$

Así pues, en esta sección se presentó cómo funciona la clasificación automática de textos para asignar etiquetas a documentos. Las herramientas hasta aquí expuestas se utilizaron para detectar automáticamente mensajes con contenido xenófobo en redes sociales. En las próximas secciones de este capítulo describiremos los enfoques teóricos que nos ayudarán para estudiar la estructura y contenido lingüístico de tales mensajes.

2.2 Análisis crítico del discurso

El Análisis Crítico del Discurso (ACD) es la operacionalización del análisis de las creencias que comparte una sociedad para sostener una organización jerárquica, en la medida en que estas creencias se configuren en un plano semiótico. Desde este enfoque, el acercamiento es crítico porque trata de explicar cómo a través del discurso se promueven o se justifican relaciones de poder, o bien, cómo se construye la oposición a ellas (Fairclough, 2010). Del desarrollo teórico que ofrece esta gama de estudios aquí interesa responder ¿qué es un discurso? ¿cómo un discurso es capaz de sostener u oponerse a dichas relaciones de poder?, ¿cuáles son las prácticas de su producción, distribución y consumo?, ¿quiénes participan en dichas prácticas y qué lugar ocupan en tal jerarquía social?, ¿desde qué plataformas se emiten dichas prácticas y qué repercusiones tienen las diferentes plataformas en esas prácticas discursivas? Y ¿cómo este se ha analizado cuando este es un producto lingüístico? Para Fairclough (2010) el discurso es, en primer lugar, una categoría analítica que denomina a las formas semióticas de construir aspectos del mundo. Esta definición obliga a detenerse en varios puntos. En primer lugar, Fairclough elige formas semióticas para indicar que el discurso (el que construye aspectos del mundo) va más allá de la lengua, existe también, por ejemplo, en 'modalidad visual'. En segundo lugar, el teórico prefiere la noción de construir que la de representar, porque este verbo da cuenta de las posibilidades de ir cambiando aquello que es representado. En tercer lugar, un aspecto del mundo es una condición que atañe a un grupo de personas; por ejemplo, la vida de las personas pobres que, en principio, no solo se debe a las condiciones económicas y sociales, sino a los discursos que, junto con aquellas, la constriñen. Así, como parte de la visión dialéctica de

las relaciones que pretende explicar el CDA, desde esta perspectiva, dichas construcciones se relacionan con diferentes actores o grupos sociales.

Ahora bien, ¿cómo es que un discurso es capaz de sostener u oponerse a relaciones de poder? Al respecto Fairclough advierte que el análisis de estas formas semióticas será crítico ahí donde estas sean relacionadas con otros elementos sociales de manera que sea posible ver cómo dichas figuras semióticas sean utilizadas para establecer, reproducir o cambiar relaciones desiguales (Fairclough, 2010). Existe, diría Fairclough, una relación dialéctica entre estas formas semióticas y algunas circunstancias (o errores) sociales.

Desde la psicología social y la sociología, pero apoyando este punto, es posible explicar una relación igualmente dialéctica entre estratificación social y nuestras capacidades cognositivas (Massey, 2008). La estratificación social, explica el sociólogo Douglas S. Massey, se da porque toda sociedad humana se caracteriza por una estructura social basada en categorías que son una combinación de rasgos. Pero advierte "Antes de que la inequidad categorial pueda ser puesta en marcha socialmente, las categorías deben ser creadas a nivel cognoscitivo para clasificar mentalmente a las personas con base en alguna combinación de características logradas y adscritas" (Massey, 2008, p. 66).

El discurso, en ese sentido, estaría estrechamente relacionado también con esas capacidades cognoscitivas en las que se asienta el orden social. Cuando desde la lingüística de corpus se ha hecho CDA, se ha encontrado, por ejemplo, que para el caso sobre los discursos de migración la constante ocurrencia de las palabras *migrante + ilegal*, puede provocar la asociación de las personas migrantes con actos de ilegalidad (P. Baker & McEnery, 2005; Dobrić Basaneže & Ostojić, 2021; Montali et al., 2013; Taylor, 2014). Y aun fuera de la lingüística de corpus, los estudios del discurso han encontrado en la acumulación de formas semióticas con el mismo sentido, la construcción de discursos (Fairclough, 2003). Ahora bien, no es solo la acumulación, sino la oportunidad de expresar dichos discursos lo que, naturalmente, podrá lograr ir a favor o en contra de la estructura social. Por ello, son prácticas discursivas de las élites simbólicas (van Dijk, 2016) –aquellas que tienen el control principal sobre el discurso– las que primeramente fueron analizadas bajo este enfoque. Las élites simbólicas que se distinguen son la prensa, la política institucional y la academia.

Dentro de este enfoque, el racismo ha sido de interés prioritario. Además, el discurso sobre migración tiende a colocarse como un tipo de discurso racista, por ello, en esta sección nos centraremos en los esfuerzos de esta postura para entenderlo. Así, van Dijk describe un cuadro ideológico en el que inscribe al discurso racista. Se trata de una oposición entre el endogrupo y el exogrupo, de modo que los principios generales que organizan el discurso (racista) se basan en enfatizar lo positivo del Nosotros, enfatizar lo negativo de Ellos, desenfatizar lo positivo de Ellos, y desenfatizar lo negativo del Nosotros. Es decir, que son discursos que se darán en dos sentidos, por ejemplo, lo civilizado vs lo no civilizado.

Como ya decíamos, las formas semióticas en las que se construye el discurso son varias, sin embargo, aquí las que nos interesan son las formas lingüísticas. Al respecto, el CDA ha operacionalizado el análisis al investigar, por ejemplo, el énfasis en los temas negativos sobre «Ellos» en titulares y primeras planas; la repetición de temas negativos en historias cotidianas; la expresión de estereotipos en la descripción de miembros de grupos étnicos; la selección de términos (los miembros de nuestro grupo siempre son «luchadores por la libertad», mientras que los de los otros son «terroristas» traidores); el uso de pronombres y demostrativos que implican distancia («esas personas»); metáforas negativas («invasión», «olas» de inmigrantes); el énfasis hiperbólico en sus propiedades negativas («parásitos»); eufemismos de nuestro racismo («descontento popular»); y falacias en la argumentación para demostrar sus propiedades negativas.

2.3 Lingüística de corpus y Activación Léxica

En esta sección se desarrollará la propuesta teórica de Michael Hoey de Activación Léxica (2005), también se verá la utilidad de esta propuesta para explicar la expresión de posturas ideológicas en los corpus trabajados.

El concepto de activación léxica propuesto por Hoey apunta a un fenómeno psicolingüístico rastreable en corpus. Si bien el concepto propio de la psicolingüística evoca la relación entre dos elementos lingüísticos, a saber el *priming* que provoca una palabra *target*, la atención de Hoey se dirige al ítem priming en sí (2005, p. 8). En consecuencia, en rasgos generales se entiende por priming la activación –o restricción– de una palabra (o un grupo de palabras)

que también es restringida por determinados contextos como pueden ser grupos semánticos, funciones pragmáticas, y posiciones gramaticales.

La *colocación* es el concepto que subyace a esta noción de *priming* o *activación léxica*. Generalmente entendida como dos o más palabras que suelen ocurrir juntas (e.g. *inevitable* + *consequence*), esta definición de colocación pone de manifiesto dos cuestiones: la colocación es un fenómeno psicolingüístico; y la evidencia puede ser recogida estadísticamente por medio de corpus computarizados. Así, una colocación se entiende como una asociación psicológica entre palabras (más que entre lemas), con una ventana máxima de 4 palabras (a la izquierda y a la derecha), cuya evidencia de coocurrencia es observable en corpus más allá de una distribución aleatoria (Hoey, 2005, p. 5) . Con esta definición, Hoey persigue la explicación de cómo se consigue lo natural en una lengua para, entonces, explicar lo que es posible en ella.

La asociación psicológica entre palabras, que es la primera característica de la colocación y por lo tanto también lo es de la activación léxica en todos sus niveles lingüísticos, remite al léxico mental. La colocación, entendida bajo esta definición, es observable en un corpus por la co-ocurrencia de dos o más palabras. Atendamos los siguientes ejemplos propuestos por Hoey. En (5) se lee las primeras oraciones del libro de viajes de Bryson (Bryson, 2010); en (6) esas mismas oraciones parafraseadas por Hoey. Ambos ejemplos comparten, además del sentido, la ocurrencia de varias palabras.

(5) In winter Hammerfest is a thirty-hour ride by bus from Oslo, though why anyone would want to go there in winter is a question worth considering.

(6) Through winter, rides between Oslo and Hammerfest use thirty hours up in a bus, though why travelers would select to ride there then might be pondered.

La primera de las oraciones es, sin embargo, natural. La diferencia entre la naturalidad de la primera y la torpeza de la segunda subyace en la distinción entre colocación normal y colocación poco usual (Partington, 1998). En el corpus que Hoey usa (compuesto por un poco más de 95 millones de palabras, procedentes de textos periodísticos; complementado con un poco más de 3 millones de palabras del British National Corpus (texto escrito) y

230.000 palabras de datos hablados), *in* y *winter* ocurren juntas 507 veces (la colocación inicial de (5); en oposición a lo que sucede en (6), es decir la coocurrencia entre *through* y *winter* -perfectamente gramatical, es decir, posible mas no natural- sucede siete veces. Lo mismo ocurre con las colocaciones de otras palabras que comparten los ejemplos, hecho que puede corroborarse en (Hoey, 2005, pp. 6-7).

Lo anterior pasa porque cada palabra está restringida o activada para el uso colocacional (2005, p. 8); a esta característica Hoey la llama *ubicuidad* de la colocación. Sin embargo, dicho juego de activación y restricción entre palabras sucede también entre conjuntos de palabras, entre categorías gramaticales, o entre asociaciones semánticas; incluso a nivel textual, intertextual o contextual, es decir, a contextos sociales como los que el correr del tiempo puede marcar (H. Baker et al., 2017; Hoey, 2005).

Dicha ubicuidad, además, hace posible otra característica de la activación léxica, a saber la anidación. Es decir, la asociación colocacional activará o restringirá, a su vez, a la gramática o al sentido. Hoey ofrece el ejemplo de la asociación de sentido a partir de lo observado en (5). La asociación semántica hace referencia a dicha anidación de palabras, observable en los corpus, pero, a diferencia del concepto de colocación, estará situado a un nivel más abstracto en la medida en que permita no solo palabras frecuentemente asociadas, sino sentidos, como en los ejemplos de (7), que, como ya se dijo, fueron obtenidos a partir de (5) (Hoey, 2005).

(7) SMALL PLACE is a NUMBER-TIME-JOURNEY – (by VEHICLE) – from LARGER PLACE

(8) Ntobeye is a two-hour ride by four-wheel drive vehicle from the vast refugee camp at Ngara.

(9) The village is a four-hour drive from London.

(10) Pamuzindo is an hour’s drive from Harare.

Dichos ejemplos se deben a la consideración de la palabra *hour* dentro de la frase *thirty-hour ride* que, si bien es una frase que un hablante produce con relativa naturalidad, ocurre

apenas una vez en la evidencia de corpus que Hoey presenta. *Hour,* no obstante, sí co-ocurre con colocaciones como *half an, one, two, three, four* y *twenty*; además, en sus concordancias ocurren las frases NUMBER-TIME-JOURNEY dentro de anidaciones mayores que sugieren varias asociaciones semánticas que pueden o no corresponder con patrones gramaticales.

Con estas bases (a saber, la colocación como el origen de la activación léxica por su ubicuidad –casi todas las palabras tienen colocaciones– y su capacidad de anidación a contextos mayores), Hoey postula las hipótesis principales de su propuesta teórica (Hoey, 2005, p. 13).

1. Cada palabra está preparada para ocurrir con palabras específicas; tales ocurrencias forman colocaciones.
2. Cada palabra está preparada para ocurrir con determinados conjuntos o sets semánticos; tales conjuntos forman asociaciones semánticas.
3. Cada palabra está preparda para ocurrir en asociación con funciones gramaticales determinadas; formando así asociaciones pragmáticas.
4. Cada palabra está preparade para ocurrir, o para evitar ocurrir, en determinadas posiciones gramaticales, así como para desempeñar, o evitar desempeñar, determinadas funciones gramaticales; por estas restricciones gramaticales forman coligaciones.
5. Los cohipónimos y sinónimos difieren en cuanto a sus colocaciones, a sus colocaciones, a sus asociaciones semánticas y a sus coligaciones.
6. Cuando una palabra es polisémica, las colocaciones, asociaciones semánticas y coligaciones de cada uno de sus sentidos diferen entre sí.
7. Cada palabra está preparada para ser usada en uno o más roles gramaticales; estas son sus categorías gramaticales.
8. Cada palabra está preparada para participar o evitar ciertos tipos de relación cohesiva en un discurso; las relaciones en las que participa son sus colocaciones textuales.
9. Cada palabra está preparada para ocurrir dentro de un discurso en una relación semántica particupar; estas ocurrencias formarán asociaciones semánticas textuales.

10. Cada palabra está preparade para ocurrir o evitar ciestas posiciones dentro del discurso; estas ocurrencias formarán coligaciones textuales.

2.3.1 Activación Léxica y Discursos Ideológicos

Hasta aquí hemos dicho, a propósito de la activación léxica, que está circunscrita al contexto lingüístico, y al extralingüístico; que se refuerza o debilita por dichas anidaciones en la medida en que vuelve a ocurrir en los mismos contextos, o no sucede, o sucede otra cosa en su lugar; que se anida y se combina (como en (7)). En esta sección argumentamos que el enfoque teórico de Hoey es útil para estudiar la expresión de posturas ideológicas, la adhesión a ellas, así como la constitución de los discursos.

Dado que entendemos, por estos últimos, formas semióticas que construyen aspectos del mundo, definimos por *postura política* expresiones que reproducen, justifican o ponen en entredicho esos aspectos. Explicaremos que en la relación entre expresiones de posturas ideológicas y constitución del discurso xenofóbico, la acumulación de las primeras dará como resultado al segundo. De este modo, aquí revisaremos cómo dichas expresiones son posibles por los efectos de restricción y anidación –en asociaciones semánticas– propias de la activación léxica; y cómo a partir de la acumulación de tales asociaciones se forma un discurso cohesionado identificable en un corpus temático.

La explicación a la expresión de posturas ideológicas es posible, en parte, por la restricción que condiciona la elección léxica. Esta condición está sujeta a los encuentros anteriores en los que ocurrió una palabra, lo que implica una diferenciación entre tipos de colocación. Partington (1998), por su parte, ofrece una distinción entre *colocación normal* y *colocación poco usual*. Desde una perspectiva diacrítica, se ha hablado de *colocaciones estacionales* frente a *colocaciones consistentes* (H. Baker et al., 2017; P. Baker et al., 2008); las primeras se entienden como colocados muy frecuentes en poco número de años, mientras que las segundas hacen referencia a dos palabras mutuamente restringidas por, al menos, siete décadas de un siglo. En función de que lo que se busca es explicar el refuerzo de los elementos activados, esta última categoría resulta oportuna; si los continuos encuentros son los que refuerzan la restricción, un enfoque diacrónico puede dar cuenta de la *fuerza* de la misma cuando ambos elementos están restringidos.

En contraste, este enfoque también pretende explicar el debilitamiento entre elementos

lingüísticos. De hecho, en este debilitamiento descansa la propuesta de Hoey para dar razón del cambio lingüístico al introducir el concepto de *craqueo* de la activación. Si el refuerzo sucede con la acumulación de los mismos elementos, el debilitamiento ocurre ante la exposición de un elemento nuevo (en términos psicolingüísticos, ante input poco usual). Entre tanto, en lo que a esta tesis se refiere, este contraste entre el refuerzo y el debilitamiento es uno de los elementos que permiten inferir que un hablante expresa, o se adhiere a una u otra postura ideológica.

Lo anterior se apoya en los conceptos de *activación productiva* y *activación restringida* (Baker et al., 2017; Hoey, 2005). La primera, que Baker et. al. (2017) definen como activa, es aquella que se introduce como posible entre las opciones del hablante, y este de hecho la reproduce. La activación restringida, o pasiva, igualmente es input para el hablante, pero no hay probabilidad, o incluso posibilidad, de que este lo reproduzca. Más que hablante, se está en posición de un lector o un oyente, por ejemplo, si la asociación se introduce por un texto del s. XVIII no ocurrirá la adopción (Hoey, 2005). En cambio, la producción de un priming conocido estará sujeta a tres factores: el primero es personal, en la medida en que convenga al hablante; el segundo es contextual, en la medida en que la situación o contexto fomente la producción; y el tercero es social, por ejemplo, el empleo de ciertos primings que reflejen la membresía a un grupo que se desea pertenecer (Baker et al., 2017).

La propuesta que aquí hacemos de la cohesión de un discurso ideológico, entre tanto, sigue la propuesta de Hoey de aplicar el análisis de la cohesión de un texto, dentro de un corpus temático (Hoey, 2017). A su vez, su propuesta de la cohesión textual es una extensión de las nociones de colocación y de asociación semántica (Hoey 2005). Parte de dos diferentes definiciones de colocación para llegar a una tercera. Por un lado, la común dentro de la Lingüística de corpus, que la define como dos palabras cuya co-ocurrencia, en una ventana delimitada, sea observable por medidas estadísticas; por otro, aquella que postula que la relación entre el vocabulario de un texto ayuda a crear la cohesión del mismo. Tomando una y otra definición como criterios, Hoey habla entonces de *colocación textual*, aquello para lo que está preparado el léxico y por cuyo efecto se logrará la cohesión textual.

Nuevamente es necesario partir del elemento léxico. De acuerdo con Hoey, el uso de una palabra requiere que el hablante conozca que esta es capaz de formar relaciones cohesivas. Con ello, Hoey lleva su propuesta, que actuaba principalmente en los dominios de la

cláusula, al texto. Mientras el hablante conoce esta propiedad de una palabra creará expectativas en su interlocutor, que según puede suponerse para fines explicativos, tiene el mismo conocimiento. Esta expectativa es consecuencia de la naturalidad que se explicaba anteriormente. Estas expectativas, Hoey las expone en dos ejes.

El primero de estos ejes postula que una palabra, o un conjunto anidado de estas, está restringido, o no, para participar en *cadenas* o *links* cohesivos; y del mismo modo, cada uno de estos vínculos está restringido a ser de un tipo gramatical específico (coligación textual). Una cadena se define como tres o más elementos léxicos vinculados entre sí por una coligación textual. En un link, el vínculo sucede solo con dos. Además, la extensión está relacionada con el grado cohesivo que logre en el texto: a mayor extensión, mayor cercanía con el tópico y mayor cohesión. Tomemos como ejemplo lo sucedido (11).

(11) With a spare hour on my hands before lunch in Lebanon this week, I revisited the joys of my childhood, crunched my way across the old Beirut marshalling yards and climbed aboard ***a wonderful 19th-century rack-and-pinion railway locomotive***. Although scarred by bullets, the green paint on ***the wonderful old Swiss loco*** still reflects the glories of steam and the ***Ottoman empire***.

For it was ***the Ottomans*** who decided to adorn their jewel of Beirut with ***the latest state-of-the-art locomotive***, ***a train*** which one carried the German Kaiser up the mountains above the city where, at a small station called Sofar, the Christian community begged for his protection from the Muslims. 'We are a minority,' they cried, to which the Kaiser bellowed: 'Then become Muslims!'

Estos párrafos, tomados de 'The irresistible romance of a steam train scarred with the bullet holes of battle' de Robert Fisk, por Hoey, contienen una cadena y un link. La primera está compuesta por la frase nominal *a wonderful 19th-century rack-and-pinion railway locomotive* cuyo núcleo, la locomotora, es posteriormente referenciada por otras tres frases nominales, cada una menos extensa que la otra*: the wonderful oldo Swiss loco, the lastest state-of-the-art locomotive* y *a train*. Entre tanto, el link compuesto por dos frases nominales *Ottoman empire* y *the Ottomans*, nos advierte Hoey (2005, p. 117), no vuelve a ser

mencionado en el texto; con un lugar más periférico en la temática del texto, tiene, del mismo modo, una menor participación en la cohesión textual.

Ahora bien, el segundo de los ejes postula que cada ítem léxico (o bien la combinación de ellos), estará condicionado, o no, a ocurrir en un tipo específico de relación semántica o pragmática. Entre las primeras se encuentra, por ejemplo, contraste, comparación, secuencia temporal, causa y efecto, ejemplificación, problema y solución. Entre las segundas, se encuentran las pautas para cambio de turno usadas en el análisis conversacional; o dentro de un texto, la relación entre el escritor y el lector.

Estas relaciones semánticas y pragmáticas apuntan al hecho de que la acumulación de una palabra no solo persigue la repetición de la información, sino la agregación. Volviendo al referente de la ya mencionada cadena (11), se observa que es introducida como el medio que ayuda al narrador a *volver a visitar* sus alegrías de infancia. En otras palabras, cuando el escritor introdujo *I revisited the joys of my childhood*, creó la expectativa, entre otras, de conocer el medio que lo llevo a dichas alegrías. Así mismo, la primera frase nominal de la cadena indica la época y las características de la locomotora; la segunda agrega su procedencia. No se trata, pues, solo de recuperar la materia anterior, sino de entender un asunto anterior en un contexto nuevo (Hoey, 2017).

Por último, si bien ya se mencionó que la restricción ocurre entre un par de elementos léxicos a nivel de la cláusula, o también a nivel textual, se debe advertir que esto es por el conocimiento que el hablante tiene sobre un ítem dado, conocimiento que se justifica en la noción de un lexicón mental que prepara a los hablantes para todas estas restricciones. Cuando estas restricciones ocurren a nivel textual, los hablantes deben poner en función también su memoria.

Con estos mecanismos de cohesión en mente, Hoey se pregunta si, dado que la cohesión textual es resultado de los encuentros repetidos de elementos lingüísticos, la repetición del encuentro entre elementos léxicos en el terreno de la intertextualidad de un corpus temático no podría también suponer la cohesión del mismo (Hoey, 2017). Siguiendo esta misma lógica, esta propuesta es llevar la revisión de dichas repeticiones no solo al nivel intertextual de nuestros corpus temáticos, sino vincularlo con el tenor ideológico de lo que aquí se revisa.

2.4 Lingüística cognitiva

A las corrientes revisadas hasta ahora en este Marco Teórico, sumamos la lingüística cognitiva. El inicio de este capítulo explora técnicas propias del PLN para un análisis macro que permita encontrar patrones lingüísticos en un corpus de redes sociales. A medida que avanzamos en este texto, lo que se busca resolver es cómo operan esos patrones. En ese sentido, el enfoque del ACD trata de explicar dichos patrones como parte de un discurso culturalmente arraigado; entre tanto, la lingüística de corpus mediante la explicación de la activación léxica intenta explicar tanto estadística como psicológicamente cómo es que dichos patrones se afianzan en el habla. En esta investigación se echa mano, finalmente, de la lingüística cognitiva como complemento de una de tales explicaciones, a saber, cómo es que el léxico está preparado para coocurrir con otros elementos. De este modo, la lingüística cognitiva supone la conclusión del análisis, conclusión que se realizará mediante el microanálisis de las representaciones simbólicas de las oraciones que se extrajeron en las redes sociales en la sección de resultados. Así bien, con las diferentes corrientes teóricas que se utilizan, la propuesta es hacer un análisis bajo un enfoque funcionalista y experiencial, por lo tanto, sustentado en datos empíricos.

En contraste con el enfoque racionalista que predominó en el siglo XX, la lingüística cognitiva es un enfoque sustentado en la relación entre el lenguaje, la percepción y la cognición. Esta corriente lingüística se fundamenta en la idea de que el lenguaje se basa en una base experiencial dada a partir de la relación entre el mundo, la percepción y la cognición. También en contraste con un enfoque racionalista y formalista, este enfoque dirigió su atención al significado, este lo explica a la luz de tales procesos cognitivos en tanto el primero es considerado “un fenómeno mental, y los significados de las expresiones lingüísticas se corresponden con representaciones conceptuales de los sujetos” (Ibarretxe-Antuñano & Valenzuela, 2012). Por lo tanto, se retoman algunos postulados de esta corriente teórica con el propósito de dar cuenta de elementos léxicos que tienden a discursos xenofóbicos gracias a conceptualizaciones y operaciones mentales de los hablantes (Ibarretxe-Antuñano & Valenzuela, 2012). Particularmente, interesa un marco que dé cuenta los conceptos de subjetividad y deixis.

El punto de partida se encuentra en la comprensión del significado a partir de la experiencia

en el mundo de los hablantes. A semejanza de la teoría de la activación léxica en que dos o más elementos lingüísticos se afianzan a fuerza de la coocurrencia, para la Lingüística Cognitiva «[…] el significado y su valor se entroncan en la naturaleza de nuestros cuerpos y nuestros cerebros, a medida que se desarrollan a través de las continuas interacciones con diferentes entornos que a su vez tienen unas dimensiones sociales y culturales. La naturaleza de nuestra experiencia corporeizada motiva y restringe la manera en la que las cosas nos resultan significativas» (Johnson, 1997 en Ibarretxe-Antuñano & Valenzuela, 2012). Así, este enfoque no se comprende sin considerar la percepción. Esta se refiere a la capacidad de los seres humanos para interpretar y comprender el mundo que les rodea a través de la interacción con su entorno. La lingüística cognitiva considera que la percepción es un aspecto fundamental en la estructuración de las categorías de significado de la realidad, ya que los conceptos e ideas de los hablantes están influenciados y conformados por la estructura de sus cuerpos y la experiencia del mundo que los rodea. En este sentido, la percepción es un componente crucial en la formación de la cognición y el lenguaje, ya que influye en la organización y almacenamiento de conocimientos sobre el mundo y la realidad, así como en la producción y comprensión lingüística.

Al respecto, es relevante el concepto de esquema de imagen de cognición propuesto por Johnson (Johnson, 1990). Los esquemas de imagen constituyen estructuras mentales que se abstraen a partir de interacciones recurrentes del hablante con el entorno. Por ejemplo, se deriva una estructura común a partir de la multitud de experiencias físicas en las que se percibe el desplazamiento de objetos en el espacio. Estas experiencias heterogéneas comparten un núcleo abstracto o "esquema" conformado por tres elementos: un punto de origen, una trayectoria y un punto de destino. Este esquema ORIGEN-TRAYECTORIA-DESTINO puede facilitar la identificación del alcance referencial de diversos eventos. Otro esquema es el del CONTENEDOR, que sucede mediante la enunciación de eventos como entrar/salir de habitaciones, meter/sacar objetos de cajas, etc. Implica una zona interna, una externa y un límite que las separa.

En la explicación del significado por medio de estos esquemas mentales subyacen las nociones de subjetividad y deixis. Ambas involucran al hablante como el punto de partida –

o mejor dicho, de referencia– de estas trayectorias; la primera de ellas mediante su abstracción, la segunda mediante la codificación de elementos lingüísticos con la capacidad de describir dichas trayectorias. Como se verá más adelante, estos conceptos son relevantes en la descripción de los discursos revisados en esta investigación. Son necesarios, además, en la medida en que ya se han revisado a la luz del comportamiento de partículas con valor deíctico que sobresalieron en nuestro corpus. Hablamos específicamente del adverbio “aquí” (Maldonado, 2013).

El concepto de deixis se entiende, de acuerdo con Maldonado (2013), como la capacidad del lenguaje para señalar o referirse a elementos del discurso en relación con el contexto en el que se produce la comunicación. Entendido así, algunas palabras tienen la capacidad de señalar elementos específicos en el espacio, el tiempo, el discurso y en relación con los participantes en la comunicación. En este caso, se analiza el contraste entre los adverbios deícticos "aquí" y "acá" en el español, centrándose en su relación con la proximidad, la subjetividad y la experiencia del hablante. Entre tanto, la subjetividad se entiende como la capacidad de los adverbios deícticos "aquí" y "acá" para reflejar la experiencia y la perspectiva del hablante en relación con la proximidad de los elementos señalados. Se establece que "aquí" se especializa en demarcar una región próxima al emisor, pero con la suficiente distancia para ver las cosas con una mayor objetividad, lo que se denomina como subjetividad media o cuasiobjetiva. Por otro lado, "acá" implica una proximidad mayor que da paso a la emergencia de una amplia gama de significados subjetivos asociados a la experiencia del hablante, lo que se denomina como subjetividad profunda. La subjetividad, pues, se relaciona con la experiencia y la perspectiva del hablante en la comunicación.

En síntesis, este marco teórico articula postulados teóricos, metodológicos y herramientas del PLN, el ACD, la lingüística de corpus y la lingüística cognitiva. Esta convergencia permite abordar integralmente y de manera innovadora la construcción discursiva de posicionamientos xenófobos en el contexto digital, particularmente dentro de un corpus en el que, como se verá, la subjetividad emerge en el uso de adjetivos posesivos (“nuestro/a”) y otras marcas que evidencian una valoración personal de los hechos, construyendo nocionalmente categorías como "nosotros" y "ellos" en relación a un territorio

determinado.

3 Planteamiento del problema

Los países del norte de Centroamérica, México y los Estados Unidos de América conforman el corredor migratorio más grande del mundo (Caribe, 2019). Ello implica, para México, dinámicas de movilidad como país de origen y de destino. En tanto país de destino, en los últimos años ha recibido personas, tanto de Centroamérica como del Caribe, que han optado por ingresar al país desde su frontera sur por medio de las llamadas caravanas, un fenómeno de migración masiva (Caribe, 2019; Vázquez Meneley, 2019). Este fenómeno ha contado con cobertura mediática y, ya sea por esto o por las mismas características del ingreso, ha contado también con interés público que es posible observar en redes sociales y que se expresa, a menudo, como rechazo no solo al hecho, sino a las personas que lo realizan.

Estas posturas de rechazo, ya sean dadas en discursos, ya con políticas que regulen el flujo migratorio, tienen su contraparte en concepciones para las cuales las migraciones humanas deben ser comprendidas como un fenómeno inherente al humano, comprensión que sea la base para tratarlas como un derecho resguardado (Rodríguez Albor, 2019).

¿De qué dependen entonces las concepciones que adscriben una postura de rechazo y de qué aquellas que se suman a una postura de aceptación? Seguramente son varios los elementos detrás de ello, y por lo tanto, una perspectiva multidisciplinar ayude a responder esa pregunta.

Al respecto, Fiske (2009) sugiere que existe un proceso de cognición social que nos lleva a conceptualizar a las personas. Dicha conceptualización no es meramente una inferencia racional, sino un proceso en el que las emociones también están involucradas: “Las emociones almacenadas en el sistema límbico pueden ser positivas o negativas pero cuando se asocian con clases particulares de personas u objetos contribuyen al *prejuicio*, el que constituye una orientación emotiva predeterminada a los individuos o los objetos” (Fiske et al., 2009).

Se hablaría entonces de un elemento cognitivo que podría encontrar su confirmación en la organización social. En este sentido, Massey (2008) sugiere que la evaluación que se mencionó en el párrafo anterior tiene repercusiones en la segregación social; esto es, no es

solo que incumba a lo que se dice, se piensa o se siente ante un grupo de personas, sino que esa evaluación psicosocial puede justificar las condiciones (favorables o desfavorables) en las que vive dicho grupo de personas.

De hecho, el ACD (Fairclough, 2010; van Dijk, 2016) está basado en una premisa parecida a la de Massey: los discursos construyen un aspecto del mundo. La segregación social es un aspecto del mundo. Así, un ejemplo tanto de segregación como de aspecto es que existan personas que se trasladan desde su lugar de origen a otro distinto, y que haya quien las crea sujetos de derechos o quien crea que se les debe negar tales derechos. El discurso, a su vez, es todo recurso semiótico que lleve a conceptualizar a las personas migrantes como merecedoras de derechos, pero también todo recurso semiótico que lleve a conceptualizar a las personas migrantes lejos del estatus de merecedoras de recursos.

Desde esta perspectiva, habría una disputa entre discursos por la construcción del aspecto en cuestión. Así lo sugiere, al menos, tanto el mismo uso del verbo *construir* entre los conceptos del enfoque (Fairclough, 2010), como la distinción entre los discursos de las élites frente el discurso de las resistencias (van Dijk, 2016). El par de concepciones a las que se hace referencia en los primeros párrafos de este apartado pueden ser un ejemplo de ello. Además, dicha disputa se llevaría a cabo en términos de producción, reproducción y consumo del discurso. Mientras más expuesta esté una persona ante un discurso, más probable es que lo asuma (Fairclough, 2010).

Por un lado, cuando se menciona al componente cognitivo detrás de un prejuicio pareciera que el primero condiciona al segundo. Por otro, se sugiere que hay en disputa diferentes conceptualizaciones y alguna puede imponerse, formando el prejuicio. Sin embargo, el propósito de enumerar diferentes perspectivas disciplinares nunca fue decidir entre una y otra explicación sino presentarlas como coordenadas cuya intención común es situar el problema de por qué conceptualizamos diferentemente al mismo grupo de personas. Intención que es común a este proyecto en el que proponemos la mirada al mismo fenómeno desde la propuesta de la Activación Léxica (Hoey, 2005).

Como el nombre de tal enfoque sugiere, descansa en el concepto de la activación léxica de la psicolingüística, esto es, proponemos una mirada que considera un componente

cognitivo. Así, este concepto se ha usado para explorar la relación entre elementos léxicos, semánticos y gramaticales, mediante el registro de la actividad cerebral que se genera en el hablante ante un estímulo lingüístico conocido, esto es, un priming que activaría, o no, otro llamado target. Por ejemplo, un par de lexemas con una relación semántica reconocida por el hablante genera que este los asocie en un tiempo de reacción menor, a un par de lexemas sin relación conocida. Este tipo de técnicas han sido, de hecho, utilizadas para medir la activación del prejuicio mediante la actividad cerebral (Wang et al., 2011; White et al., 2009). No obstante, la propuesta no es mirar la actividad cerebral, sino el discurso dicho recopilado en corpus.

Con esta visión de corpus, las respuestas que se buscan están condicionadas a las evidencias de la lengua en uso; particularmente, se busca el rastro del discurso xenofóbico dado entre los usuarios de las redes sociales. Aquí, es necesario detenerse a pensar la migración como una cuestión pública en tanto se pretende regular desde la institucionalidad; explicar desde la academia; y es fuente de debates en, y entre, diferentes estratos de la vida social. Con todo ello, genera discursos que además pueden vincular tales estratos: la prensa, por ejemplo, hablándonos de la nueva caravana; el candidato proponiéndonos tal o cual medida regulatoria de la migración o con miras a cuidar los derechos de las personas migrantes. Luego estamos nosotros, sin ostentar puesto o títulos, que comentamos todo eso, porque sí, en sitios de confianza, a saber, en redes sociales.

Pues bien, dentro de ese entorno, ¿qué es el discurso xenofóbico contra migrantes? Este estudio parte de dos suposiciones. La primera es que tal discurso se deriva de la exposición real o imaginada (dando por hecho que la relación entre cada quien con la migración es diferente) al hecho del ingreso de migrantes centroamericanos y caribeños a México, es decir, es una reacción. La segunda es que las “coordenadas” teóricas de las que se habló anteriormente, tiene componentes cognitivos, por lo que queda preguntarse ¿qué evidencias hay de ellos en un conjunto de comentarios hechos en redes sociales?

En tanto al enfoque social de la construcción del discurso, se considera que las características propias de las redes sociales las hacen una especie de ágora. Un lugar para expresar, entre otras cosas, animadversiones. ¿Qué consecuencias, entonces, tiene ello

sobre las expresiones xenofóbicas? En otras palabras, ¿es posible rastrear, a partir de la acumulación de dichas expresiones, ideas comunes que configuren o construyan un discurso xenofóbico en esta masa, caracterizada por la falta de indentificación individual que se manifiesta cuando todos hacemos uso de las redes sociales?

4 Antecedentes

Los discursos en torno a la migración han sido objeto de numerosos estudios desde diversas disciplinas y enfoques teórico-metodológicos. En las últimas décadas, la lingüística computacional y la lingüística de corpus han aportado herramientas valiosas para analizar estos discursos de manera empírica y sistemática. Precisamente, en esta sección interesa revisar una serie de antecedentes y estudios previos que han aplicado estas disciplinas al análisis de discursos a propósito de la migración.

Si bien ya se ha visto cómo la lingüística computacional ofrece métodos y recursos para el procesamiento del lenguaje natural (sección **¡Error! No se encuentra el origen de la referencia. ¡Error! No se encuentra el origen de la referencia.**), se verá en el primer apartado de este capítulo, sección 4.1 Detección de Discursos Xenofóbicos, cómo el análisis de sentimientos ha sido una herramienta destacada que permite detectar automáticas actitudes, emociones y evaluaciones sobre personas migrantes en diferentes textos.

Entre tanto, en la sección **¡Error! No se encuentra el origen de la referencia. ¡Error! No se encuentra el origen de la referencia.**, veremos cómo la Lingüística de corpus permite construir y explorar grandes conjuntos de datos textuales (corpus) con el fin de identificar patrones lingüísticos recurrentes. Mediante técnicas como el análisis de frecuencias, listas de palabras clave y colocaciones, esta disciplina ha contribuido a desvelar las representaciones e ideologías subyacentes en los discursos sobre migración.

4.1 Detección de Discursos Xenofóbicos

Como se vio en el marco teórico, el análisis de sentimientos es una técnica de clasificación de documentos a partir de su polaridad. Llamado así porque la identificación de cada documento se realiza a partir de la oposición entre sentimientos, este tipo de análisis es útil cuando se trata de detectar el uso de la lengua contra un grupo de personas; a este uso también se le conoce como lenguaje de odio. Cuando se trata de expresiones maliciosas hacia migrantes se le denomina como *lenguaje xenofóbico, antimigrante*, o bien se le agrupa bajo el concepto de *lenguaje racista*. En ese sentido, en esta sección, dedicada a la revisión de literatura, se revisarán algunos análisis de sentimientos que se han hecho sobre lenguaje

de odio hacia personas migrantes o hacia refugiados, con especial atención a aquellos proyectos hechos con conjuntos de datos en español.

Se revisarán dos aspectos de los trabajos presentados pues son estos los elementos necesarios en las tareas de detección de clasificación de texto (Rico-Sulayes, 2018). El primero de ellos se refiere al proceso de preparación de los datos para los experimentos. El segundo de estos aspectos son los experimentos que se hicieron en cada uno de los trabajos que se presentan, esto es, tanto los lineamientos que siguieron los experimentos (particularmente, la extracción de características, y los clasificadores utilizados) como los resultados de los mismos.

4.1.1 La preparación del conjunto de datos

Cuando desde el PLN se plantea una tarea de aprendizaje supervisado, como la que aquí se propone, la participación humana es una etapa vital del proceso, pues de ella dependerán los insumos de los que los algoritmos de clasificación deberán aprender. Si bien dicha participación está presente a lo largo de todo el proceso, la preparación de data set es un momento donde la intervención humana es total en tres pasos. El primero de ellos es la decisión del esquema de anotación a elegir; el segundo se refiere a los criterios de anotación; y el tercero es la aplicación de dichos criterios a cada documento por anotadores humanos así como la posterior comparación entre las anotaciones hechas.

Los primeros de estos pasos están vinculados a la delimitación de lo que se entiende por lenguaje de odio. En la medida en que una tarea de clasificación es una tarea de polarización, supone oposiciones que deben quedar suficientemente marcadas en los esquemas de etiquetación. Entre tanto, debe considerarse que tanto estos esquemas como los criterios de clasificación son factores de variación en los resultados obtenidos en los experimentos pues el lenguaje de odio cambia dependiendo del contexto (Nobata et. al, 2016). Por estas razones, de los antecedentes aquí presentados se consideran ambos pasos (Tabla 1).

De acuerdo con Poletto et al. (2021), es posible encontrar en la literatura tres esquemas de clasificación del lenguaje de odio. El primero de dichos esquemas, de etiquetación binaria, cuenta con dos clases mutuamente excluyentes para marcar la presencia (ej. lenguaje

antimigrante), o ausencia (ej. lenguaje a favor de la migración o neutral) de un fenómeno. Dentro de los esquemas de etiquetado no binario se observarán más de dos valores que marcan los matices de un fenómeno (ej. nivel alto de agresivad, nivel medio, nivel bajo, y sin agresividad). Los esquemas multinivel pueden implicar una o varias series de rasgos diferentes (ej. agresivo vs no agresivo), o una o varias escalas de variación (como diferentes niveles de agresividad), o bien la combinación de series polarizadas con escalas de variación. A semejanza de la revisión de literatura hecha por estos autores, aquí se encontró que, cuando se trata de delimitar lo que es el lenguaje de odio, los esquemas de anotación son claros, pero no se tiende a reportar los criterios con los que se hacen los esquemas y, cuando se hace, no se tiende a unificar criterios.

Tabla 1. Criterios para anotación de los data set en los antecedentes

Autores	**Esquema de anotación**	**Criterios de anotación para documentos xenofóbicos o con lenguaje de odio**	**Ejemplos**
Arcila Calderon et. Al (2020)	**Anotación multinivel** 1) Rechazo (binario: rechazo - no rechazo); 2) justificación de rechazo (no binario) -Carga económica, -Amenaza de la seguridad, -Amenaza a la identidad, -Amenaza de invasión, -Rechazo manifiesto, -Prejuicio social; 3) tipo textual -informativo -opinión.	1) asociación de migrantes con eventos negativos, 2) asociación de migrantes con carga económica, 3) representación de migrantes y refugiados como invasión, avalanche, y empobrecimiento.	
Pitropakis et. Al (2020)	**Anotación no binaria:** 1) antimigrante, 2) no negativo, 3) no decidido, 4) no relacionado.	1) presencia de calumnias xenófobas, ataques, críticas, o descalificaciones a un grupo o individuo que es parte de él, 2) presencia de hashtags xenófobos,	

		3) defensa de xenofóbia.	
(*DETOXIS-IberLEF*, 2021)	**Anotación multinivel:** 1) Binaria (tóxico, no tóxico), 2) detección del grupo objetivo (lenguaje misógino, y lenguaje antimigrantes).	NA	levemente tóxico '*Vienen a pagarnos las pensiones'*, tóxico *'asi me gusta, que se maten entre ellos y en alta mar. Mas inmigrantes asi porfavor* '(Sic), muy tóxico *'A esos moros hay que echarlos pero ya. O los políticos hacen algo o la gente tendrá que actuar*'.
HatEval SemEval-2019	**Anotación multinivel:** A partir de 2 niveles binarios 1) Lenguaje de odio (presencia o ausencia), 2) alcance (grupal o individual), 3) agresividad (presencia o asuencia).	NA	NA
Sanguinetti et. al. (2018)	**Anotació multinivel:** 1) Detección lenguaje de odio, 2) detección de agresividad, 3) detección de ofensa, 4) detección de ironía, 5) detección de estereotipo.	Criterio para lenguaje de odio: 1) contenía grupo objetivo 2) contenía acción (fuerza ilocucionaria del enunciado (Searle, 1969): difunde, incita, promueve o justifica el odio o la violencia hacia el destinatario dado, o un mensaje que pretende deshumanizar , deslegitimar, herir o intimidar al objetivo.	Ejemplo de lenguaje de odio: *'la prossima resistenza la dovremo fare subito contro gli invasori islamici!'* (¡la próxima resistencia la tendremos que hacer inmediatamente contra los invasores islámicos!)
Waseem & Hovy (2016)	**Anotación no binaria:** 1) Lenguaje agresivo, 2) lenguaje sexista, 3) lenguaje racista.	1) usa insultos sexistas o racistas; 2) ataca a una minoría;	

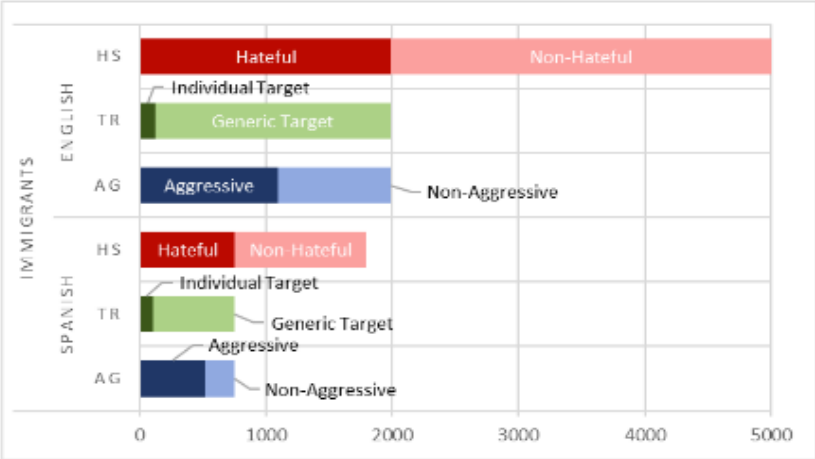
		3) busca silenciar a una minoría; 4) critica a una minoría, 5) promueve, pero no usa directamente, discurso de odio o crimen violento; 6) critica a una minoría y usa un argumento de hombre de paja; 7) tergiversa la verdad o busca distorsionar las opiniones de las minorías con afrmaciones infundadas; 8) muestra compatibilidad hashtags problemáticos; 9) estereotipan negativamente a una minoría.	
(*HUHU*, 2023)	**Anotación multinivel:** 1) Tweets que expresan prejuicios que usan humor vs tweets que expresan prejuicio y no usan humor; 2) detectar grupo minoritario objetivo; 3) detección del grado del prejuicio (escala 1-5).	NA	NA

Dentro del campo que aquí interesa (**¡Error! No se encuentra el origen de la referencia.**), lo que es y no lenguaje antimigrantes es etiquetado bajo esquemas no binarios (Pitropakis et al., 2020; Waseem & Hovy, 2016) y, principalmente bajo esquemas multinivel (Arcila Calderón et al., 2020; Basile et al., 2019; *DETOXIS- IberLEF*, 2021; *HUHU*, 2023; Sanguinetti et al., 2018). Estos últimos a su vez tienen al menos una tarea de clasificación binaria; no obstante, sobresale que el esquema binario en solitario no suele considerarse. Lo anterior se debe a que las tareas de clasificación y los esquemas de etiquetado no son

necesariamente equivalentes, pero también a la pretensión de explicar el fenómeno a mayor detalle: un esquema binario corresponde con explicaciones más bien generales. Así, puede observarse en la **¡Error! No se encuentra el origen de la referencia.**, que cuando se hace una tarea multinivel, el primero de los esquemas es para tareas de detección del lenguaje de odio, y una vez detectado este, se detalla el objetivo de dicho odio (*DETOXIS-IberLEF*, 2021; *HUHU*, 2023), o el grado de odio expresado (Basile et al., 2019), o bien si el lenguaje de odio empleado usa recursos lingüísticos como humor (*HUHU*, 2023), o ironía (Sanguinetti et al., 2018).

El ejemplo que se observa a través de las tareas de HatEval, explicadas por Basile y colegas (2019) corresponden a la combinación de tres niveles (cada uno de ellos binario). El primero de los niveles distinguió si un tweet era o no lenguaje de odio, y cada tweet que lo era fue clasificado en otros dos niveles secundarios, uno que explicitaba si el odio iba dirigido al miembro de un colectivo (mujer o persona migrante) o al colectivo; otro que lo clasificaba como tweet agresivo o no. La orientación del enfoque de lo general a lo particular se aprecia en la Ilustración 6, que muestra la composición del conjunto de datos para tweets sobre inmigrantes (para consultar también aquellos sobre lenguaje misógino ver Basile et al., 2019). Ahora bien, observamos que la visión de los niveles secundarios no es menos compleja. El primero de dichos subniveles abona a la visión sociológica del problema, principalmente en la medida en que puede vincular el odio del que es objeto un individuo debido a su pertenencia a un colectivo (en tanto mujer o migrante, para los ejemplos de HatEval). Mientras el segundo de estos subniveles abona a la explicación lingüística dado que la diferenciación de un lenguaje agresivo cuya ocurrencia no es igual a la de lenguaje de odio obliga a la inferencia de que la agresión es solo una de las formas de expresar el odio.

Ilustración 6. . Distribución de documentos clasificados en Basile et. al. 2019.



Resalta, por otro lado, que aquí estamos ante un intento de unificar criterios de etiquetación de un tópico (lenguaje antimigrante) que involucra diferentes lenguas: Basile et al. (2019) tomaron como base a Sanguineti et al. (2018), quienes diseñaron el esquema con sus respectivos criterios con tweets en italiano. Para estos autores, la presencia del lenguaje de odio existía en aquellos tweets en que hubiera un grupo objetivo del odio, y una acción definida como la fuerza ilocutiva del enunciado apreciada en mensajes que difundan el odio hacia el grupo objetivo. Dicho odio, además, era clasificado en cuatro subniveles (agresividad, ofensividad, ironía y estereotipo), además clasificados en subniveles de intensidad.

Además de la diferencia de la profundidad que ofrecen los subniveles del trabajo de Sanguinetti et al. (2018), destaca que estos autores solo trabajaron con lenguaje de odio contra personas migrantes. Entre tanto, Basile et al. (2019) trataron con lenguaje de odio tanto contra este grupo como contra mujeres. A semejanza de estos autores, Wasem & Hovy (2016) desarrollaron un detector de mensajes sexistas y racistas, sin embargo, este proyecto opuso como clases a contraponer el sexismo vs el racismo, esto es, distinguían el lenguaje de odio con base a su grupo objetivo. Los autores utilizaron un conjunto de datos compuesto por 16, 914 tweets en inglés. Igualmente, a semejanza de Sanghuineti et al. (2019), pero con

mayor interés para nuestro trabajo, por tratarse de un proyecto elaborado con un conjunto de datos en español, encontramos el esquema de anotación de Iberian Languages Evaluation Forum (IberLEF 2021), DETOXIS, que además de oponer los mensajes llamados "tóxicos" vs los "no tóxicos" para mensajes misóginos y antimigrantes, contó con una tarea que dependía de esta primera oposición en la que la toxicidad era subdivida en tres niveles de la misma.

Dentro de los esquemas con criterios lingüísticos se encuentra la nueva edición de IberLEF 2023, que pretende detectar además el uso del humor en tweets perjudiciales contra grupos objetivos. La primera tarea es binaria y consiste en la detección del humor en mensajes odio hacia cuatro grupos de minorías. La segunda tarea es la detección del grupo objetivo hacia el que se dirige el mensaje de odio, a saber, 1) mujeres y feministas; 2) comunidad LGBTIQ; 3) inmigrantes y personas racializadas; 4) personas con sobrepeso. La tercera clasifica en cinco niveles el nivel del perjuicio de cada mensaje.

Arcila Calderon et al. (2020) proporcionan también una anotación multinivel con criterios de codificación detallados. Trabajaron con un data set en español que, al igual que lo que se presentará más adelante (sección **¡Error! No se encuentra el origen de la referencia.**. **¡Error! No se encuentra el origen de la referencia.**, y **¡Error! No se encuentra el origen de la referencia.**. **¡Error! No se encuentra el origen de la referencia.**) trata el lenguaje de odio hacia migrantes latinoamericanos (particularmente venezolanos) emitido por otro país latinoamericano (Ecuador). Sobresale que los criterios de su tarea binaria corresponden con lo encontrado en estudios de discurso antimigrante desde la lingüística de corpus, como se verá en este mismo capítulo más adelante. Así pues, se consideró un documento como lenguaje antimigrante y anti refugiados aquel que contenía una asociación de tales personas con delincuencia, con una carga económica, con invasiones, con avalancha, o con empobrecimiento. Sin embargo, no especificaba si hubo criterios lingüísticos para distinguir cada una de esas asociaciones.

Los esquemas no binarios de anotación pueden igualmente contribuir a visiones generales del fenómeno. Pitropakis et al. (2020) clasificaron un conjunto de datos de 47,976 tweets en inglés en cuatro categorías (o clases), tres de ellas opuestas a su clase de interés de una

polaridad emocional negativa, o bien, de lenguaje antimigrante. En la Tabla 1 (p.47) puede identificarse cada una de ellas, la mayor oposición es aquella que distingue al lenguaje antimigrante, que, a grandes rasgos, distingue al grupo objeto del odio (migrantes) y a las formas de expresar tal odio.

Ahora bien, además de lo necesario para dejar preparado un corpus experimental que nos permita entrenar un sistema de clasificación automática de sentimientos, es necesario procesar tal conjunto de datos de modo que se seleccione un conjunto de rasgos o características que le permitan distinguir los sentimientos de interés. Posteriormente, tales características serán vectorizadas, o lo que es lo mismo llevadas a una representación numérica para que los algoritmos de clasificación las procesen de modo tal que puedan asignarle prioridad a aquellas que consideren representativas de la clase de interés. Finalmente, una vez que los algoritmos trabajen con las características vectorizadas, la clasificación que hagan será evaluada. Dado que son las medidas de evaluación las que nos permiten conocer el desempeño de un experimento –cada experimento es la combinación de las características escogidas procesadas por un clasificador– serán estos valores los que nos permita reconocer a continuación qué tan eficientes han resultado este tipo de sistemas para clasificar lenguaje contra migrantes.

Hemos visto en esta sección que la preparación de las clases dentro del data set obedece a la necesidad de encontrar el llamado lenguaje de odio contra personas migrantes. A semejanza de lo encontrado por Poletto et al. (2021), se observa que ni los esquemas de anotación ni los criterios con que los sustentan son los mismos, no obstante, la presencia de dos elementos son constantes: el grupo objetivo y las expresiones maliciosas contra él. Fueron precisamente las expresiones aquellas que presentan mayor divergencia en los esquemas de anotación.

En el trabajo de Arcila y Calderón (2020), la aproximación al conjunto de datos fue mediante la selección de tres categorías gramaticales para su vectorización, a saber, verbos, sustantivos y adjetivos; de todos ellos solo se utilizaron los cinco mil más frecuentes. El clasificador Naïve Bayes para modelos multimodales fue el mejor clasificador en términos de precisión, logrando 79,88% en este aspecto. Además, presentó una exactitud del 74,64%,

una exhaustividad del 72,23% y una puntuación F1 de 75,86%. La precisión se refiere al porcentaje de casos correctamente identificados como positivos en relación con todos aquellos que se identificaron como positivos. La exactitud es una métrica que indica el porcentaje de casos correctamente predichos por el modelo, ya sean de la clase de interés o no. La exhaustividad representa el porcentaje de casos verdaderamente positivos en relación con todos los casos que realmente son positivos. Por último, la F1-score es una medida que equilibra los casos positivos detectados al calcular la armonía entre la precisión y la exhaustividad.

En el contexto de su participación en DETOXIS 2021, Paula y Schlicht (2021) llevaron a cabo una comparación entre dos tipos de clasificadores estadísticos, específicamente los bayesianos y los basados en el concepto de Máxima Entropía (MaxEnt), para abordar una tarea de clasificación binaria. Para lograr esto, modificaron la cantidad de atributos empleados mediante la evaluación de diferentes variantes de n-gramas, incluyendo (1-1)-gramas, (1,2) gramas y (1-3) gramas. Los n-gramas se refieren a secuencias de palabras adyacentes en el texto. Representaron estas características de dos formas distintas: una mediante la técnica de bolsa de palabras (BOW), en la que se considera la presencia o ausencia de un término en un documento, y otra mediante la abstracción de TF-IDF, en la que se expresa una característica en función de la frecuencia con la que aparece un término en comparación con otros documentos. En relación a los dos esquemas de clasificación, se observó que el enfoque BOW arrojó los puntajes más elevados en términos de F1-score (0.4679 para los modelos MaxEnt y 0.5355 para los modelos bayesianos). Además, esta representación también proporcionó los mejores resultados al aplicar los modelos basados en MaxEnt, con una precisión de 0.7126 y una exhaustividad de 0.419. Frente a esto, los clasificadores bayesianos alcanzaron la exhaustividad más alta (0.8004). No obstante, la técnica de vectorización mediante TF-IDF obtuvo la precisión más alta (0.8928) para los modelos basados en MaxEnt, así como la mejor calificación tanto para la exactitud (0.6933) como para la precisión (0.7282) en los modelos bayesianos. A pesar de las variaciones en la selección de atributos en los diversos modelos, no se lograron deducciones concluyentes ni generalizables debido a que las métricas que arrojaron los mejores resultados estuvieron

vinculadas a distintos números de n-gramas para cada métrica en particular. En otras palabras, un modelo que exhibió un alto desempeño en una métrica no necesariamente garantizó resultados satisfactorios en otras métricas, como se evidenció en el caso del modelo con la mejor precisión (0.7126) para la tarea binaria, que generó una precisión de 0.6188, una exhaustividad de 0.3128 y un F1-score de 0.4101.

Pitropakis y sus colegas (2020) desarrollaron modelos utilizando secuencias de palabras y de caracteres; para los primeros, los n-gramas fueron de distancia de 1 a 3, para los segundos de 1 a 4. Esta elección estuvo basada en la suposición de que serían los n-gramas de palabras los que arrojarían los resultados más favorables. Las características fueron vectorizadas mediante TF-IDF. Para evaluar sus modelos, se recurrió a métricas de exhaustividad, precisión y F1-score. Los modelos que implementaron una Regresión Logística demostraron igual eficacia para identificar contenido xenofóbico tanto con n-gramas de palabras como con n-gramas de caracteres, logrando una exhaustividad del 87% en ambos casos. Sin embargo, la extracción de n-gramas de palabras logró un mejor desempeño en cuanto a precisión (85%) y F1-score (84%). Tomando como base la noción de que la detección de aspectos textuales de naturaleza temática, como el análisis de sentimientos que concierne aquí, tiende a beneficiarse más del uso de n-gramas, y tomando en consideración los resultados en este experimento de Pitropakis y su equipo (2020), este trabajo consideró la evaluación y prueba de varios modelos basados en n-gramas de palabras.

Como se ha visto, tanto para de Paula y Schlicht (2021) como para Pitropakis et al. (2020) utilizar n-gramas de diferente longitd es un recurso con la intención de explorar la pertinencia de la variación del tamaño de vocabulario para obtener mejores resultados en la clasificación. Por otro lado, son también útiles cuando se trata de ver el lenguaje en su contexto. Al respecto, a partir del data set de tweets en español que se proporcionó en el SemEval19, Plaza del Arco et al. (2020) compara un modelo de clasificación basado en un lexicón construido a través de términos ofensivos contra migrantes, contra modelos de aprendizaje supervisados basados en una vectorización en Tf para unigramas y bigramas. Los mejores resultados se obtuvieron cuando combinaron ambas longitudes –a lo que

agregaron un clasificador de votación–, con una precisión de 0.707, una exhaustividad de 0.713, una F1-score de 0.707 y una exactitud de 0.711; no obstante, los resultados de F1-score obtenidos para el modelo basado en lexicón son casi los mismos que para uno de los modelos de aprendizaje supervisado, a saber, Decision Tree que obtuvo una F1-score de 0.686. De acuerdo con los autores, ello obedeció precisamente a la importancia de considerar una palabra ofensiva en su contexto; que, como bien ejemplifican con el término misógino *puta*, este puede expresar una polaridad positiva en contextos lingüísticos y coloquiales específicos (*puta madre*, con un significado de fantástico), o bien puede reforzar su polaridad negativa en otros (*hijos de puta*).

En esta sección, se revisó una serie de trabajos anteriores de clasificación automática, relevantes en tanto que abordan el tema del lenguaje de odio hacia personas migrantes, principalmente en español. Esto obedeció a la necesidad de conocer cómo esta técnica se ha aplicado al tema de este estudio, así como para contextualizar los hallazgos encontrados; qué significan los resultados de esta tesis en relación con estos antecedentes es un tema que será abordado en el capítulo **¡Error! No se encuentra el origen de la referencia.**.

Ahora bien, como ya se ha mencionado, esta investigación pretende dar cuenta también a un nivel descriptivo del discurso xenofóbico, de modo que en la siguiente sección se verá cómo la Lingüística de corpus ha incorporado la perspectiva crítica para dar cuenta de discursos discriminatorios contra las personas migrantes.

4.2 Estudios del Discurso sobre Migración desde la Lingüística de corpus

Esta es la segunda y última de las secciones donde se exponen los hallazgos de trabajos que utilizan herramientas computacionales para aproximarse a los discursos ideológicos sobre fenómenos migratorios. Si desde la perspectiva del análisis de sentimientos lo que interesa es detectar mensajes que representen alguna forma de violencia contra las personas migrantes, desde la lingüística de corpus, en conjunción con el ACD, lo que atañe responde al hecho de cómo una postura ideológica puede empatar con ciertas formas lingüísticas. No se profundiza aquí en el debate sobre cómo un enfoque cuantitativo contribuye a la mirada crítica del estudio del discurso –que puede consultarse en (Baker et al., 2008; McEnery &

Hardie, 2012; Stubbs, 1997)–, sino que se examina cómo las técnicas propias de los estudios de corpus se han utilizado para ello en lo que se refiere al estudio de los discursos contra migrantes.

La lingüística de corpus es una ciencia de datos en la medida en que estos tienen un papel protagónico en la investigación. Por un lado, es conocido que a partir de las propuestas metodológicas que se han desarrollado desde este enfoque pueden llevarse a cabo estudios tanto basado en datos, como guiados por ellos (McEnery & Hardie, 2012). Por otro lado, ya entrado en un análisis específico, se espera que, de los datos agrupados por sus comportamientos cuantificados en categorías basadas en medidas estadísticas (por ejemplo, la colocación), se desprendan conclusiones teóricas (por ejemplo, cómo a partir de la colocación es posible hablar de la preferencia semántica). Específicamente aplicada al estudio de discursos ideológicos, lo que esta disciplina debería lograr es la explicación tanto lingüística como semiótica de la lengua empleada para tales fines. Por lo tanto, lo que en esta revisión de literatura se verá es cómo han sido obtenidos y tratados los datos, qué explicaciones sobre la lengua se han obtenido, y cómo explican el funcionamiento del discurso ideológico sobre migración.

4.2.1 Punto de partida: construir un corpus

Dentro de los estudios críticos del discurso, conocer tanto a los actores como las condiciones de producción y enunciación del discurso es necesario precisamente para lograr una perspectiva crítica. A grandes rasgos se trata de resolver la cuestión sobre quién dice qué sobre quiénes, cuestión capital cuando se explica cómo una sociedad jerarquizada es representada en los discursos. Recordemos que, a grandes rasgos, se ha resuelto esta cuestión como el estudio de las élites o de las resistencias. Las primeras, al contar con mayores recursos para la producción y distribución de *sus ideas*, lograrían la construcción de lo que Fairclough denominó *aspectos del mundo*. Dichos aspectos tienden a "organizar" las relaciones sociales en que se involucran jerárquicamente tanto tales élites como otras personas en situaciones de menores privilegios o incluso de desventaja. Un discurso de las resistencias se opondría a tales construcciones. Un discurso de personas que no

corresponden ni a las élites ni a las resistencias, como el que es problema de esta investigación puede colocarse entre esos dos.

En aras de resolver tal cuestión, la metodología de la lingüística de corpus aporta soluciones desde la misma construcción de los corpus por medio de los cuales se estudiarán los discursos. Esta disciplina buscar dar cuenta de la lengua en uso, para lo cual suscribe el análisis de los rasgos de la misma a las fronteras del corpus. Dichas fronteras deben ser reconocidas desde el momento mismo de la construcción de los corpus, pues estos están sujetos a criterios situacionales que permitan extraer muestras *representativas* de diferentes realidades del habla (Biber, 1993). Los alcances explicativos del proyecto de investigación dependerán precisamente de cómo se fijen los criterios de tales muestras.

Se presentan aquí tres tipos de criterios de delimitación. El primero es el actor social que emite el discurso, es decir, determina si se habla de un discurso de élite, de resistencia o de un tercer actor; obedece a la necesidad de dar cuenta de la ideología vinculada a tal actor social. El segundo criterio es la tipología textual en la medida en que esta puede estar vinculada con la variación de las expresiones lingüísticas utilizadas. Los primeros dos criterios, como se verá a continuación, tienden a superponerse, debido a que desde el planteamiento teórico de los ACD la prensa, por ejemplo, es identificada como un actor social de las élites, pero es, desde estudios lingüísticos, un tipo textual. El tercer criterio es contextual y da las particularidades sociohistóricas del hecho sobre el que se trate en las producciones discursivas. Al respecto, aquí se hablará, por un lado, de países en tanto son las fronteras de estas demarcaciones las que sitúan políticamente el tipo de migración del que se trata en estos trabajos; por otro lado, se hablará de temporalidad que demarca al corpus pues esta tiende a ser escogida en virtud de sucesos particulares que pueden ser factor para una u otra postura ideológica.

Así pues, para esta revisión de literatura se da cuenta de 21 trabajos que utilizaron algún tipo de cómputo de datos textuales para dar explicaciones lingüísticas sobre discursos ideológicos del tema migratorio, la mayoría de estos utilizaron herramientas propias de la lingüística de corpus, salvo dos trabajos. La primera de estas excepciones (Hartnett, 2019) resulta útil en la medida en que usa técnica del PLN, particularmente un modelado de

tópicos, que permitió perfilar el análisis en un sentido crítico. La segunda de las excepciones (2019), al contrario, opta por un conteo "manual" que considera únicamente las frecuencias brutas de algunas palabras, no obstante, llama la atención que trata el discurso mexicano sobre migración, aunque la autora trata el tema de la migración mexicana.

Tabla 2. Tipología textual de los discursos analizados

Tipología textual	**Prensa**	**Documentos jurídicos**	**Organismos internacionales**	**Blogs académicos**	**Redes sociales**
# artículos	18	2	1	1	1

Cabe aclarar que, para esta revisión de literatura solo se encontraron trabajos que analizaran el llamado discurso de las élites. De ello se da cuenta en la Tabla 2, donde también se distribuye las tipologías textuales que se revisaron y entre las cuales *prensa* es mayoritaria. Dentro de los ACD en general es reconocida esta tendencia a investigar sobre todo el discurso de las élites (van Dijk, 2016). Entre tanto, la particular preferencia por estudiar la prensa puede deberse a que esta es considerara como una cámara de eco que refleja prácticas lingüísticas populares al tiempo que sus productos están modelados por propósitos específicos e ideologías (Hartnett, 2019).

Ahora bien, *élite* o *resistencia* no deben entenderse sino como categorías analíticas útiles para explicar sociedades organizadas jerárquicamente. Para propósitos de un estudio crítico del discurso, ello quiere decir que si bien un grupo social debe cumplir con ciertas características para considerarse en una u otra categoría (a saber, mayor oportunidad para la difusión de sus ideas), no necesariamente se relaciona la pertenencia a una de estas para profesar una única posición ideológica. Una práctica dentro de la lingüística de corpus para dar cuenta de esta diversidad es la de la comparar conjuntos de textos diferenciados por su tipología textual o bien por el actor social de quien se tomó el discurso.

De hecho, es esta práctica la que provoca que en la Tabla 2 la distribución de las tipologías rebase en número a los artículos reportados, pues algunos de estos compararon dos tipologías textuales. Entre estos trabajos se encuentra el de Baker y McEnery (2005) que

comparan el discurso de la prensa británica y el de la Oficina del Alto Comisionado de las Naciones Unidas; o el de Turnbull (2018) que consiste en una comparación entre lo que sucede tanto en la prensa británica como en blogs de corte académico que versan sobre temas migratorios. Los resultados de estos trabajos muestran que es acertada la decisión metodológica de contrastar dos diferentes tipos de discursos –aunque los dos correspondan a actores que son parte de las élites– pues tanto los discursos académicos como los de un organismo internacional manifiestan una postura más favorable hacia poblaciones migrantes en oposición a los discursos periodísticos.

En tanto el principal interés de los trabajos reportados aquí es el de dar cuenta de la relación de las formas lingüísticas con la que se expresan las posturas ideológicas, la diferenciación de las tipologías tiende a supeditarse frente a la necesidad de dar cuenta del discurso de un actor. Así, en los trabajos de Baker y McEnery (2005) y Turnbull (2018) el contraste entre las tipologías textuales es importante en la medida en que dan cuenta de actores sociales diferentes. En un sentido semejante, también sucede que es posible contrastar únicamente una de tales tipologías textuales, pero dando cuenta también de diferentes actores sociales. Ello es lo que sucede en el emblemático trabajo de Gabrielatos et al. (2008) en el que se construyó un corpus de 140 millones de palabras compuesto por artículos de prensa publicados desde 1996 hasta el 2005 en Reino Unido. Los datos provinieron de periódicos nacionales y regionales, así como de publicaciones de mayor seriedad (*broadsheet*) y de corte sensacionalista (*tabloids*).

Con unos criterios semejantes para el diseño de corpus, pero para el caso de la sociedad alemana, Hartnett (2019) escogió cuatro periódicos. De una cobertura regional se seleccionó *Berlinier Zeitung* y *Der Tagesspiegel*; con cobertura nacional, *Der Spiegel* y *Die Zeit*. A grandes rasgos, se destaca que tanto *Berlnier Zeitung* como *Der Tagesspiegel* presentan una perspectiva más cercana a las experiencias migratorias. Si bien ambas publicaciones destacan experiencias personales y familiares de los migrantes, la primera de estas las presenta con una orientación hacia el futuro de Alemania, mientras que la segunda enfatiza la situación legal, así como los derechos de los migrantes en dicho país. Entre tanto,

Der Spiegel y *Die Zeit* se enmarcan en contextos más amplios, tanto nacionales como de perspectiva internacional.

Por su parte, Islentyeva (2021) construyó un par de corpus de prensa basada en criterios ideológicos, de modo que se preocupó por obtener muestras de periódicos británicos que pudieran dar cuenta de las perspectivas políticas en medios impresos tanto de izquierda, como de derecha. Identificadas con un enfoque político de izquierda se agruparon las publicaciones de *The Guardian, The Observer* y *The Daily Mirror*; entre tanto, la perspectiva política de la derecha fue representada por las publicaciones de *The Daily Telegraph, The Daily Mail* y *The Sun*.

Trabajos como el que presentan Pérez Paredes et al. (2016) y Bevitori (2018) se aproximan al discurso de las élites a través de textos legislativos. Los resultados de ambas investigaciones indican una representación de las personas que migran mediada por el control del número de personas que ingresan al Reino Unido (Bevitori, 2018) o bien del uso de mecanismos administrativos para tales fines (Pérez Paredes et al., 2016). Sobresale que este tipo de actor, el político –y en oposición a la prensa– la representación negativa de los migrantes es menor, aunque no por ello deja de percibirse una prosodia negativa sutil.

El trabajo de Camargo Fernández (2021) es el único entre la literatura revisada que, como este proyecto, dirige el análisis hacia el discurso sucedido en las redes sociales. No obstante, también se centra en el discurso de las élites y, a semejanza de las investigaciones de Pérez Paredes et al. (2016) y Bevitori (2018), se interesa en el discurso producido por un actor político; particularmente el producido en Twitter por VOX, partido político español de derecha o extrema derecha. Este partido político fue fundado en 2013 por exmiembros del Partido Popular (PP) y se caracteriza por su ideología nacionalista española, su oposición a la inmigración, su crítica a las autonomías y su defensa de la unidad de España. En concordancias con estas características, los investigadores encontraron seis ejes temáticos recurrentes que, a grandes rasgos, presentan a las personas que migran como una amenaza; concluyen así, que existe una estrategia discursiva centrada en la fabricación y diseminación de bulos sobre inmigrantes para generar rechazo y miedo hacia ellos.

Sobre el tercer criterio para la delimitación de estas *muestras de lenguaje*, el contextual, debe decirse que la mayoría de los antecedentes aquí presentados tienen como objeto de estudio el discurso sobre migración que se emite desde el Reino Unido (Baker et al., 2008; Baker & McEnery, 2005; Bevitori, 2018; Fotopoulos & Kaimaklioti, 2016; Gabrielatos & Baker, 2008; Islentyeva, 2021; Lawson, 2015; O'Regan & Riordan, 2018; Pérez Paredes et al., 2016; Schrötei et al., 2019; Taylor, 2014; Turnbull, 2018). Entre estos se distingue un conjunto de trabajos "fundadores", que son relevantes aquí porque se trata de los primeros análisis hechos desde la lingüística de corpus, con una perspectiva crítica, sobre el tema de migración. Ya se mencionaron algunos de ellos, a saber, Baker & McEnery (2005), Gabrielatos & Baker (2008) y Baker et al. (2008). El primero de estos trabajos, sin una referencia mayor al periodo histórico, señala que el corpus está compuesto por tabloides y periódicos publicados en el Reino Unido durante el año 2003 (Baker & McEnery 2005). Los trabajos posteriores, no obstante, son parte de un proyecto mayor donde, a partir de los lineamientos planteados por Baker y McEnery (2005) se construye un corpus más amplio y, más importante, más apegado a la necesidad de construir un corpus representativo, que ha servido para varias investigaciones. Dicho corpus está compuesto por las mismas tipologías textuales y abarca 6 años, desde 1999 al 2005.

Cuando se debate el aporte que la lingüística de corpus puede hacer a la dimensión crítica de los estudios del discurso, se tiende a señalar que mediante sus técnicas cuantitativas pueden construirse corpus representativos de un tipo de discurso. En otras palabras y como ya se mencionaba unos párrafos atrás, desde la lingüística de corpus se plantea como crítica al ACD que una forma lingüística no puede ser asociada a una postura ideológica sin una selección representativa de textos para evitar conclusiones basadas en impresiones. Gabrielatos (2007) plantea esta cuestión en la selección –dentro de una base de datos– de los documentos relevantes de un tema específico, a saber, el de la migración, de modo que puede construirse un corpus especializado. Así, propone la medida RQTR (Relative Query Term Relevance, o Relevancia Relativa del Término en Consulta) que es útil en la construcción de un corpus especializado porque ayuda a seleccionar términos de consulta que contribuyan a recuperar documentos relevantes de un tema, basándose en indicadores

objetivos de la relevancia de los términos candidatos. KhosraviNik (2009) retoma esta selección de documentos para un análisis meramente cualitativo, pero ya habiendo encontrado cinco momentos en los que aumentó la frecuencia de artículos referentes a RASIM (refugee(s), asylum seeker(s), immigrant(s), and migrant(s)) Baker (2008) y relacionándolos con eventos migratorios, a saber:

1) Marzo de 1999, invasión de la OTAN en Kosovo y refugiados kosovares,
2) septiembre de 2001, los atentados del 11-S, los problemas de los solicitantes de asilo en Gran Bretaña, el caso de los "boat people" australianos,
3) mayo de 2002, la segunda vuelta de las elecciones presidenciales francesas LePen contra Chirac, la escolarización de los hijos de los solicitantes de asilo, el asesinato de Pim Fortyun,
4) marzo de 2004 - el atentado de Madrid, la ley de asilo, los controles de inmigración en Europa del Este, la ampliación de la UE, y
5) mayo de 2005 - las campañas previas a las elecciones generales británicas.

Se observa así, al tiempo que, gracias a la compilación de textos bajo criterios representativos, pueden obtenerse periodos históricos relevantes para analizar, también puede obtenerse evidencia de temas mediáticamente importantes. Particularmente, hay temas que también se verán en los planteamientos de otros trabajos que no revisan el discurso británico, a saber, las maneras polémicas en que viajan las personas para ingresar a un país (como en tema 2), campañas electorales o las elecciones mismas (tema 3 y 5), menores o niños migrantes (tema 3), y leyes o controles migratorios (tema 4). Es decir, independiente de las fronteras políticas, estas son cuestiones que sobresalen en la discusión sobre migración; qué pasa con tales cuestiones en los corpus de esta investigación se revisará en el capítulo **¡Error! No se encuentra el origen de la referencia. ¡Error! No se encuentra el origen de la referencia.**.

Por otro lado, dentro de esa base de datos se ve que estos temas rebasan, en su mayoría, las fronteras de Reino Unido; a menudo, también rebasan las fronteras de Europa, y, de hecho, la relación entre la migración y la Unión Europea (UE) es una cuestión capital en los planteamientos de problema de este y otros trabajos aquí revisados (Islentyeva, 2021;

Turnbull, 2018) cuando se trata de analizar el discurso producido desde esta demarcación sobre migración.

El trabajo de Pérez Paredes (2017) estudia un corpus que comprende textos administrativos de dominio público producidos después del proyecto RASIM, revisa la relación entre la concepción que se tiene por parte de actores gubernamentales de las personas que migran a Reino Unido desde 2007 hasta el 2011. Bevitori (2018), entre tanto, utiliza un corpus que abarca cinco años (desde el 2010 hasta el 2015), el análisis se enfoca en el año 2015, que fue clave para la crisis migratoria global por lo que se mostró un aumento de los debates sobre asilo y refugiados. Durante el 2015, además, hubo un cambio histórico en el Parlamento Británico con la llegada de un gobierno de coalición entre Conservadores y Demócratas Liberales; las elecciones de ese año marcaron un punto de inflexión con la pérdida de escaños de los Demócratas Liberales y la victoria por mayoría estrecha de los Conservadores; en ese contexto, se lanzó el programa *Vulnerable Persons Resettlement Programe* (Programa de Reasentamiento de Personas Vulnerables), lo que generó una discusión en torno a la nueva Ley de Inmigración 2015/16 que buscaba crear un ambiente hostil para migrantes irregulares con nuevas sanciones y medidas. O'Regan & Riordan (2018) trabajan también con un corpus del 2015 (desde el mes de septiembre a noviembre) con el propósito de entender cómo la prensa estaba tratando la crisis de refugiados de ese año. Los trabajos de Turnbull (2018) e Islentyeva (2020) trabajaron con corpus de este contexto histórico pero centrados en el debate pre referéndum de Brexit, durante el referéndum y posterior al mismo.

Por otro lado, encontramos trabajos que comparan discursos sucedidos entre varios países europeos. Entre esos se han identificado tres estudios que contrastan los discursos entre el Reino Unido y otros países. Fotopoulos & Kaimaklioti (2016), por ejemplo, toman el discurso de prensa de Grecia, Alemania e Inglaterra en torno al Acuerdo Unión Europea-Turquía del 20 de marzo de 2016 a 31 de mayo del 2016; el objetivo de este estudio fue revisar las maneras en que se enmarcan los migrantes y refugiados en la prensa de dichos países en medio de la llamada crisis de refugiados. Entre tanto, Schöretei et al. (2019) construyeron un corpus de prensa que abarca 24 años, a saber, desde 1998 hasta el 2012, para cuatro

países Reino Unido, Francia, Alemania e Italia. Y finalmente, Taylor (2014) compara el discurso sucedido en prensa tanto en Reino Unido e Italia con el objetivo de abonar a la mirada sociológica sobre la representación de los migrantes en ambos países.

En este análisis de la literatura, también se encontraron investigaciones que utilizaron un conjunto de datos de una sola nación, aunque con el mismo énfasis en un país europeo. En este contexto, encontramos investigaciones como la realizada por Montali et. al (2013), en la que los autores centraron su investigación en Italia durante los años comprendidos entre 1992 y 2009. Este marco temporal les permitió examinar el desarrollo del discurso en torno a los migrantes en ese país en particular. En consecuencia, descubrieron que, entre 1992 y 2002, el discurso giraba principalmente en torno a tópicos relacionados como 'inmigrantes', 'inmigrantes ilegales' y 'extranjeros'. Por el contrario, el período comprendido entre 2003 y 2009 reveló temas adicionales, como las motivaciones detrás de los viajes de los migrantes, los riesgos asociados a la migración, los desafíos a los que se enfrenta al llegar al país de acogida y las contribuciones económicas de los migrantes. Al emplear un conjunto de datos que se centra en la evolución del discurso, así como en las disparidades entre estos dos períodos, los autores también observan puntos en común, específicamente, la exclusión de los propios migrantes del discurso.

El estudio de Taylor (2009) tiene como objetivo investigar la llamada otredad a través de un estudio de corpus de la prensa italiana. El artículo es parte de un proyecto mayor llamado IntUne (*Integrated and United: A quest for citizenship in an ever closer Europe)*, fundado por la Unión Europea. La autora sitúa su estudio en el contexto nacional italiano en el que la movilización de los italianos tanto fuera como dentro de su país han sido importantes.

Entre tanto, Salahshour (2016) centró su análisis en Nueva Zelanda, que es uno de los países que más reciben migrantes en el mundo, y se centró particularmente en Auckland, que es la ciudad con mayor población migrante en el país. El periodo de interés para la recolección de los textos fueron los años 2007 y 2008, pues durante estos años los efectos de la recesión económica mundial se hicieron sentir en dicho territorio. Considerando este contexto, el estudio buscar dar cuenta del cambio en que los migrantes fueron representados antes y después de la recesión.

Hartnett (2019) examinó los patrones de comunicación de la prensa en relación con el movimiento de personas en Alemania tras la crisis europea de los refugiados de 2015. Esta investigación se sitúa en el contexto de las políticas que permiten a los refugiados cruzar libremente las fronteras (Müller, 2015) y ser acogidos por la población alemana. Este fenómeno se conoce como *Willkommenskultur*, que significa un compromiso con la humanidad, la aceptación y la fidelidad a los principios democráticos (Hartnett, 2019), y si bien este fenómeno se incrementó del 49% entre la ciudadanía alemana al 59% en 2015, la crisis europea de refugiados ha sido un tópico de división. En medio de este contexto, este estudio tiene como objetivo analizar los detalles lingüísticos de los artículos de noticias para comprender el impacto de las palabras escritas y habladas en el discurso que rodea a las poblaciones migrantes en Alemania. Al examinar el lenguaje utilizado en la cobertura mediática, el estudio busca descubrir las consecuencias de estas palabras y su papel en la formación de la opinión pública y actitudes hacia los migrantes.

Por último, entre los trabajos que analizaron el discurso de algún país europeo, encontramos el trabajo de Camargo Fernández (2021), quien exploró el discurso de la extrema derecha española. Particularmente, se enfocó en el auge de Vox como partido de extrema derecha en España durante el período de 2020 a 2021, destacando la inmigración como un eje central en su ideario.

Ahora bien, esta revisión de literatura también contó con estudios centrados en la región latinoamericana, en concordancia con el objeto de estudio de esta tesis. Entre estos trabajos tenemos dos que revisaron discursos producidos en lengua española (Galindo Gómez, 2019; Guerra Salas & Gómez Sánchez, 2017) y uno en portugués (Ferreira et al., 2017). Este último abordó la situación en Brasil, específicamente en junio de 2015, durante lo que se conoció como la "crisis de los refugiados" en Europa. Entre tanto, Guerra & Gómez (2017) contrastan el discurso español con el de otros países latinoamericanos hispanohablantes, mientras Galindo (2019) compara el discurso mexicano contra el estadounidense.

Guerra & Gómez (2017) se centra en la cobertura de los movimientos migratorios en periódicos de países de habla hispana en 2016, lo que indica un contexto social de interés en comprender cómo se retratan estos movimientos en los medios de comunicación.

Particularmente contrastan discursos de España con países latinoamericanos como Chile, Argentina, Perú, Colombia, México y Puerto Rico. Por medio de esta comparación, se analizan las representaciones de realidades migratorias, como crisis migratorias, políticas migratorias y debates relacionados, tanto en América del Norte como en Europa, con una predominancia de información sobre América del Norte.

Como en este trabajo, la investigación de Galindo (2019) estudió el discurso de México sobre la migración; la autora además contrasta este con el discurso sucedido en Estados Unidos. En concordancia con la mayoría de los estudios de esta revisión, estudió el discurso producido por la prensa. Su trabajo, en contraste este, investiga el discurso sobre los migrantes mexicanos en Estados Unidos; es precisamente en este hecho donde reside el valor de este estudio, esto es, en la comparación del discurso del mismo país de procedencia de los migrantes con el de el que los recibe.

Hasta aquí hemos revisado cómo el primer aporte de la lingüística de corpus a los estudios críticos del discurso es la construcción de los corpus mismos. Idealmente, estos se construyen con criterios tales que sea posible la representatividad del habla, como en el ejemplar trabajo de Gabrielatos (2007). Como sea, el tamaño del corpus debe ser tal que sea posible encontrar patrones del habla, en la siguiente sección veremos, precisamente, cómo son encontrados y tratados dichos patrones.

4.2.2 Explorar un corpus y encontrar patrones

Una vez que el corpus está constituido, es momento de explorarlo. En término propios de la disciplina de la lingüística de corpus, lo que interesa es la búsqueda de patrones que den cuenta de discursos ideológicamente delimitados. Para ello, como se vio en la sección anterior, se busca la construcción de bases de datos –corpus– suficientemente grandes, de modo que los patrones que se encuentren tengan fundamento estadístico. De estos hallazgos estadísticos se desprenderán las generalizaciones que permitirán vincular ciertas formas lingüísticas con un determinado discurso. Hay diferentes técnicas para tal propósito, que obedecen a diferentes niveles de análisis. Desde la lingüística de corpus todas tienen como base la frecuencia, de ahí la necesidad de la construcción de corpus representativos,

como ya se vio en la sección anterior. Entre las herramientas de esta disciplina, la colocación es la predilecta en los estudios sobre discursos ideológicos; además, también juega un rol importante en esta investigación en la medida en que está detrás del concepto central de esta tesis, a saber, la activación léxica. Se hablará también de la lista de palabras clave pues es una herramienta que apunta a elementos lingüísticos como candidatos ideales para comenzar un análisis más profundo como el ACD requiere; por ello, será sobre el uso de esta herramienta en los trabajos revisados de lo que hablaremos en primer lugar.

Se decía que el primer paso para aportar explicaciones sobre la pertenencia de rasgos lingüísticos al discurso de un actor social era la construcción de varios corpus a contrastar, o de incluir subcorpus si es necesario. Dicho contraste es el que llevaría a los rasgos lingüísticos *clave* del discurso estudiado. Es decir, una palabra es *clave*, o bien, representativa de un discurso (o un corpus), en la medida en que su frecuencia es mayor en tal discurso *en comparación* con otro. Como parte de este contraste, la frecuencia de cada token se normaliza, y sobre una base de normalización es que se puede decir que dicho token ocurre más o menos veces en un corpus que en otro. Así, finalmente se obtiene la llamada lista de palabras clave, y sobre esta el investigador debe aplicar su criterio mediante filtros y evaluaciones –claramente subjetivas (Baker, 2004, p. 353)– que le permitan llegar a la generalización de lo que es propio de un tipo de discurso (Taylor & Marchi, 2018). Por tanto, dentro de los estudios críticos del discurso hechos desde la lingüística de corpus, la lista de palabras clave es una herramienta exploratoria con la que el investigador comienza a asignar rasgos lingüísticos a un tipo de discurso ideológico.

En el trabajo de Gabrielatos (2007), la lista de palabras clave es usada, en un primer momento, para comparar la frecuencia de uso de ciertas palabras (y colocaciones) entre diferentes tipologías textuales. Una vez detectadas, estas palabras se convierten en punta de lanza para un análisis descriptivo de los discursos encontrados. En el caso del estudio de Gabrielatos, en primer lugar, estas palabras fueron agrupadas en tópicos, actitudes o topos argumentativos y, posteriormente, analizadas en el contexto más amplio de sus concordancias. Otro ejemplo se encuentra en el trabajo de Lawson (2015), un estudio guiado por palabras clave de acuerdo a su frecuencia relativa en un corpus de prensa

británica. Las palabras clave fueron agrupadas bajo cuatro categorías conceptuales: dos relacionadas con el nombramiento y la categorización de las asociaciones en términos de ser británico, y el concepto de ser un migrante, un tercer grupo se relacionó con el concepto de una vida nueva o mejor, y el cuarto con aspectos de la integración.

Ahora bien, decíamos que en esta revisión de literatura se daba cuenta de trabajos que utilizaron algún tipo de cómputo de datos textuales, la mayoría de ellos hacen tal cómputo desde las herramientas de la lingüística de corpus; no obstante, se reportó uno que utiliza una herramienta propia del PLN. El trabajo de Sabina Hartnett (2019) se acercó a las palabras representativas de un corpus de poco más de mil artículos noticiosos mediante un modelado de tópicos. Se trata de una técnica propia del PLN –como lo es también el análisis de sentimientos que se aplica en este proyecto– que agrupa en un tema o tópico un conjunto de términos de acuerdo a su prominencia y relevancia, esto es, a grandes rasgos, de acuerdo con la frecuencia de uso de las palabras que componen un conjunto de datos o corpus así como de la relación entre las mismas palabras. De este modo, encontró que el término *Willkommenskultur* se utilizó más frecuentemente hacia el 2015, uso que decayó durante el 2017.

A menudo, en lugar de la lista de palabras clave, los estudios comienzan con una serie de palabras representativas del discurso a estudiar previamente identificadas. Este es, de hecho, el caso mayoritario en este estudio. Estas palabras son lo que se conoce como *palabras clave del discurso* (DKW, Discourse Key Words) (Stubbs, 2001).En el primero de los trabajos que aquí revisamos, en que esta técnica de aproximación fue utilizada, es el de Baker & McEnery (2005), con los términos *refugees* y *asylum seekers* para iniciar una búsqueda de frecuencia y posteriormente de concordancia para su análisis. Estas palabras o lemas fueron posteriormente extendidas a actores sociales clave en el discurso de migración, de donde se desprende el acrónimo RASIM.

Entre los trabajos revisados cuya metodología inició con estas DKW se encuentra el de Taylor (2009), quien en un primer lugar buscó equivalente léxicos para estas palabras en italiano y, en segundo lugar, se dispuso a revisar la representación de tales actores sociales dentro de un encuadre de pánico moral. Esta autora (Taylor, 2014) también comparó la frecuencia de

los lemas RASIM, igualmente en italiano, pero en relación con los adjetivos posesivos para la primera persona del plural (*nostra, nostre, nostri, nostro*). Por su parte, Fotopoulos & Kamiklioti (2016) partieron de tales palabras clave para hacer un análisis comparativo en prensa de tres países (Grecia, Reino Unido, y Alemania), en el marco del acuerdo entre Turquía y la Unión Europea en marzo del 2016.

Tomando parcialmente los lemas de RASIM encontramos los trabajos de Pérez Paredes (2016) y de Turnbull (2018). El primero de tales se trata de un análisis colocacional del lema *migrant*, mismo que fue confirmado como término clave mediante una lista de palabras clave (keyword list). El segundo inicia con un conteo de las frecuencias de las palabras *migrant, migrants* y *migration*. Entre tanto, Schöretei y colaboradores (2019) usaron como términos clave sustantivos que ya no designan actores clave del fenómenos de migración, pero que sí han sido clave en el discurso europeo sobre migración, a saber *multicultural* y *multiculturalism*; así, estos términos fueron buscados en prensa británica, y sus equivalentes en artículos periodísticos franceses, alemanes e italianos.

Ahora bien, se decía que la colocación ha sido la herramienta predilecta en los estudios de corpus que buscan revisar críticamente los discursos sobre migración. Esto se debe a que la constante ocurrencia de dos términos contribuye a las propiedades semánticas de ambos; en palabras de Firth, “you shall know a word by the company it keeps” (Honeybone, 2005). De esta propiedad se desprenden dos conceptos que son fundamentales en los estudios sobre discursos ideológicos, a saber, *preferencia semántica* y *prosodia semántica*.

El primero de estos conceptos se refiere a la tendencia de ciertas palabras a co-ocurrir con categorías semánticas específicas o conjuntos léxicos, lo que indica una preferencia por ciertas combinaciones de palabras. La prosodia semántica, por otro lado, se refiere a las actitudes subyacentes o connotaciones del discurso asociadas con ciertas palabras o relaciones colocacionales. Va más allá de la mera coocurrencia de palabras y examina el significado afectivo o evaluativo general transmitido por el lenguaje utilizado.

Estos conceptos han sido ampliamente utilizados para estudiar las representaciones de los actores sociales involucrados en la migración comúnmente referidos mediante el acrónimo RASIM. El análisis de colocaciones de estos referentes ha permitido identificar patrones

subyacentes en diversos discursos sobre migración. El estudio pionero realizado por Baker & McEnery (2005), del que ya hemos hablado, reveló una tendencia a anteceder el lexema "refugee" con modificadores cuantificadores, como se ilustra en los ejemplos:

- "Our coverage of the deaths **eight** *refugees* found in a container near Wexford"
- "**Four million** *refugees* have fled in a quarter of a century of..."
- "**more** *refugees* fleeing in terror and dying in agony"

Al identificar este patrón, se le asigna una categoría semántica que agrupa los colocados que reflejan dicha "relación". En este caso, la categoría NÚMERO abarca los colocados que cuantifican a los refugiados. Siguiendo esta metodología, Baker, McEnery y otros autores han encontrado en sus datos no solo la relación NÚMERO, sino también otras categorías semánticas relevantes, como se muestra de manera no exhaustiva, en Tabla 3.

Tabla 3. Categorías de relación semántica encontradas en la revisión de literatura

Categoría de relación semántica	*Autores*
NÚMERO	(Baker et al., 2008; Baker & McEnery, 2005; Fotopoulos & Kaimaklioti, 2016; Gabrielatos & Baker, 2008; Salahshour, 2016; Taylor, 2009; Turnbull, 2018)
MOVIMIENTO	(Baker et al., 2008; Baker & McEnery, 2005; Camargo Fernández, 2021; Gabrielatos & Baker, 2008; Taylor, 2009, 2021)
AGUA	(Ferreira et al., 2017; Salahshour, 2016; Turnbull, 2018)
INVASIÓN	(Baker & McEnery, 2005; Camargo Fernández, 2021; Taylor, 2009, 2021)
IDENTIDAD NACIONALIDAD CIUDADANÍA	(Bevitori, 2018; Galindo Gómez, 2019; Guerra Salas & Gómez Sánchez, 2017; Hartnett, 2019; Islentyeva, 2021; Lawson, 2015; Montali et al., 2013; O'Regan & Riordan, 2018; Taylor, 2009)

No se ha profundizado en los resultados ni en la utilización del concepto de preferencia semántica en los antecedentes, ya que esto se explica con mayor detalle en la sección de resultados utilizando los datos propios. Además, se discuten las coincidencias en esta investigación y las que aquí se han presentado en la sección **¡Error! No se encuentra el origen de la referencia. ¡Error! No se encuentra el origen de la referencia.**. Por lo pronto, lo que sí se reslata es que los conceptos de preferencia semántica y prosodia semántica a menudo se solapan: la categoría INVASIÓN en sí misma habla de una representación negativa de los actores RASIM. Además, dado que el estudio de colocaciones sugiere un significado compartido no solo entre dos elementos léxicos continuos, sino entre varios elementos lingüísticos (Sinclair, 2004, p. 20), es posible que varias de estas categorías semánticas den pie a las demás. Continuando con el ejemplo del trabajo de Baker y McEnery (2005), se aprecia que a través de las categorías semánticas de NÚMERO y MOVIMIENTO (entre otras que pueden ser consultadas en tal estudio) dan cuenta de un discurso que enmarca a los refugiados como invasores.

Como se ha visto hasta aquí, la lingüística de corpus ha demostrado ser una valiosa herramienta para el estudio crítico de los discursos sobre migración. A través del análisis de frecuencias, listas de palabras clave y colocaciones, esta disciplina permite identificar patrones lingüísticos recurrentes que pueden revelar ideologías y representaciones subyacentes. Los conceptos de preferencia semántica y prosodia semántica son particularmente útiles para examinar las connotaciones y actitudes transmitidas por el lenguaje utilizado en torno a los actores sociales involucrados en la migración.

Si bien el punto de partida suele ser términos clave o lemas previamente identificados, como los del acrónimo RASIM, el análisis de corpus brinda un enfoque sistemático y empírico para profundizar en el estudio del discurso. Al vincular ciertos rasgos lingüísticos con discursos ideológicos específicos, la lingüística de corpus contribuye a desvelar las construcciones sociales y las narrativas que se tejen en torno a fenómenos complejos como la migración. Estos hallazgos pueden informar debates públicos más ecuánimes y sensibilizar sobre el impacto del lenguaje en la percepción de grupos vulnerables.

En la primera sección de este capítulo, se examinó cómo la técnica de análisis de sentimientos, propia del campo de la lingüística computacional, ha sido empleada para identificar discursos xenófobos. Posteriormente, se abordó la manera en que la lingüística de corpus ha sido utilizada para estudiar los discursos sobre migración desde una perspectiva crítica. En el siguiente capítulo, dedicado a la metodología, se detallará cómo se han integrado ambas disciplinas en la presente investigación con el objetivo de realizar un estudio sobre el discurso xenófobo en las redes sociales, enfocándose particularmente en los discursos relacionados con la migración de personas centroamericanas y caribeñas hacia México.

5 Objetivos e Hipótesis

Se ha dicho que esta investigación representa un acercamiento al discurso que sucede en redes sociales sobre migración; también se ha descrito las disciplinas y los enfoques teóricos entre los que se circunscribe el proyecto; así como expuesto una revisión de la literatura de los abordajes que, desde la lingüística computacional y desde la lingüística de corpus, se han hecho a este tipo de discursos. En este capítulo expondran los lineamientos de la investigación. Es decir, cuáles son los objetivos que se persiguieron, a qué preguntas se responde, así como las hipótesis que planteadas para dichas preguntas.

5.1 Objetivos

El objetivo principal es delimitar el discurso xenofóbico en torno a las migraciones centroamericanas y caribeñas en México por los temas que lo componen mediante un estudio guiado por datos.

Entre tanto, los objetivos específicos son:

- Construir dos corpus de redes sociales.
- Identificar de manera automática sentimientos y temas sobre la migración en México.
- Identificar los rasgos lingüísticos prominentes del discurso clasificado como xenofóbico por el análisis de sentimientos.
- Identificar campos semánticos del discurso xenofóbico en redes sociales (Twitter y YouTube) a través de los rasgos lingüísticos prominentes.
- Identificar patrones gramaticales en los que ocurren los rasgos lingüísticos ubicados por el modelo de clasificación automática.
- Identificar las asociaciones semánticas, así como las anidaciones en las que participen, dichos rasgos en su contexto lingüístico inmediato.
- Comparar el comportamiento de los rasgos estudiados en las diferentes redes sociales.
- Describir el discurso, o discursos, xenofóbico que sucede en redes sociales.

- Investigar las relaciones entre técnicas propias de la lingüística computacional y la lingüística de corpus, a saber, la lista de elementos con ganancia de información y la lista de palabras clave respectivamente.

5.2 Preguntas de investigación

Con el propósito de definir el discurso xenofóbico contra migrantes que opera en redes sociales, este proyecto buscar responder la siguiente pregunta general:

1. ¿Cómo se constituye el discurso xenofóbico en Twitter y YouTube?

Se busca realizar un análisis del discurso desde los principios de la lingüística de corpus (que busca, a su vez, representatividad). Esto implica que la constitución del discursos se hará con elementos representativos obtenidos de la lista de elementos con ganancia de información del análisis de sentimientos xenofóbicos. Con esto en mente:

   a. ¿Con qué campos semánticos está asociado el discurso xenofóbico mexicano que se utiliza en redes sociales?, así como
   b. ¿En qué anidaciones se encuentran tales asociaciones semánticas?

A su vez, revisados en sus contextos, estos campos formarán temas propios de un discurso xenofóbico, que estarán expresados en asociaciones semánticas; por tanto:

   c. ¿Cuáles son las temáticas que dan forma al discurso xenofóbico?

Por último, siguiendo la propuesta de Hoey, que toma los contextos extralingüísticos como un factor de activación o inhibición de los elementos lingüísticos:

   d. ¿Es un mismo discurso xenofóbico el de YouTube y el deTwitter?

5.3 Hipótesis

Se asume que la expresión del discurso xenofóbico sucede como una producción intertextual. Esto es, el discurso xenofóbico en ambas redes sociales es una unidad, sin embargo, está compuesta por diferentes emisiones protagonizadas por distintos autores anónimos. El conjunto de estas emisiones forma un discurso cohesionado por medio de palabras y asociaciones semánticas repetidas.

a. Los campos semánticos encontrados serán NÚMERO, MOVIMIENTO, INVASIÓN e IDENTIDAD NACIONAL.
b. Las asociaciones semánticas darán cuenta de temas como el territorio, invasión del territorio y grupos enfrentados en esos territorios.
c. Las asociaciones semánticas de las que da cuenta la lista de ganancia de información estará presente tanto en YouTube como en Twitter, la diferencia se deberá al registro. La primera de estas redes tenderá más a un registro informal.

6 Metodología

Este estudio, orientado a la identificación y descripción de discursos xenófobos contra migrantes en México, adoptó un enfoque multidisciplinario guiado por datos. Se centró en explicar las formas en que los hablantes expresan xenofobia. El diseño de investigación transitó de lo general a lo particular, utilizando una herramienta de la lingüística computacional para filtrar patrones característicos de expresiones de odio hacia migrantes. Posteriormente, estos patrones se analizaron desde una metodología propia de la lingüística de corpus para observar cómo dan cuenta del discurso xenófobo en redes sociales. En concordancia con el planteamiento, la investigación siguió 4 fases relativas al tratamiento de datos, cuyas metodologías respectivas ya se anunciaron en la Ilustración 1. *Planteamiento y fases de la investigación*., página 15, y se detallan en esta sección.

A la primera de estas fases le correspondió tanto la extracción de los datos como la composición de los corpus (**¡Error! No se encuentra el origen de la referencia.**). Se describirán, por lo tanto, las dos redes sociales seleccionadas: Twitter y YouTube, y cómo las particularidades de cada una obligaron a una metodología distinta de extracción datos. Advertimos que las redes sociales no fueron utilizadas de manera equitativa durante la investigación, por lo que también se revisarán los usos que se dieron a cada red, lo que obligó procesamientos de corpus diferenciados. En ese sentido, dado que los datos de Twitter fueron usados para la etapa de clasificación automática de texto, se explica su clasificación manual de polaridad (xenofóbica y no xenofóbica); así como su composición, también manual, en temáticas propias de los fenómenos migratorios de acuerdo a su sitio de ocurrencia (frontera norte y frontera sur), y un tercer tema (migración general). Mientras que para YouTube se describe la selección de datos y la composición del corpus.

La segunda fase corresponde a la etapa de clasificación automática por lo que la metodología que se siguió y se describe está en función de los principios de la lingüística computacional. En este segundo apartado (**¡Error! No se encuentra el origen de la referencia.**) se encontrará, por tanto, lo referente a la selección de atributos para vectorización, la selección de clasificadores, la implementación de experimentos y la extracción de la lista de ganancia de información.

Este último proceso marca la pauta para la detección de patrones a analizar desde la lingüística de corpus, dando con ello pie a las fases 3 y 4 que se detallan en el último apartado de esta sección (**¡Error! No se encuentra el origen de la referencia.**). Se revisará cómo se aplica la propuesta del Marco Teórico, referente a la activación léxica, propuesto por Hoey (2005) (sección **¡Error! No se encuentra el origen de la referencia.**). Este enfoque se utilizará para describir los patrones léxicos, gramaticales y semánticos que caracterizan el discurso xenofóbico en las redes sociales. Dicho análisis permitirá revelar las estructuras lingüísticas recurrentes y los campos semánticos prominentes en este tipo de discurso, lo cual contribuirá a una comprensión más profunda de las estrategias discursivas empleadas en la expresión de actitudes y creencias xenófobas en el contexto de las plataformas digitales.

6.1 Diseño de Corpus

Este apartado tiene como objetivo describir los corpus utilizados en esta investigación multidisciplinar basada en datos. Inicialmente, es pertinente detallar la aplicación de estos recursos para las distintas etapas del estudio; en ese sentido, en la Tabla 4 se muestra que los corpus se utilizaron en dos etapas. La primera de estas etapas fue el análisis de sentimientos, en el que se uso el corpus llamado Twitter no balanceado; sus características, tales como su composición por número de tweets y número de palabras puede revisarse también. En tanto, la segunda etapa correspondió al análisis del discurso que fue hecho desde la lingüística de corpus, y en concordancia con los principios de tal disciplina tenemos un corpus de referencia; para el caso de esta investigación se utilizaron además dos corpus objeto de análisis (Twitter no balanceado y el corpus de Youtube).

Tabla 4. Corpus usados por momentos de la investigación

Análisis de sentimeintos	**Análisis del discurso**		
Twitter balanceado	**Twitter no balanceado**	**YouTube**	**Corpus de referencia (**SketchEngine)
712 tweets Xenofóbicos: 293 No xenofóbicos: 418	3,832 tweets Xenofóbicos: 293 No xenofóbicos: 3,539	Comentarios a 12 videos Palabras: 351, 637	Palabras: 1,028, 321

Palabras: 20, 848	Palabras: 103,874		

Para los corpus de Twitter se extrajeron tweets escritos desde el 21 de abril del 2020 hasta el 30 de noviembre del mismo año; entre tanto, para YouTube se extrajeron comentarios a 12 videos publicados del 20 de octubre del 2018 al 7 de agosto del 2020. Cabe destacar, además, que el corpus de Twitter balanceado, se obtuvo del corpus de Twitter no balanceado –como se explicará más adelante. Una descripción de los corpus puede consultarse en las secciones **¡Error! No se encuentra el origen de la referencia. ¡Error! No se encuentra el origen de la referencia.**.

En lo que respecta al corpus de referencia extraído de Sketch Engine (SKE), se trata de textos académicos disponibles en páginas web de diversas disciplinas de ciencias sociales y humanidades como ciencias políticas, sociología, demografía, pedagogía, antropología, psicología, e incluso teología. Las temáticas se centran en familias, género e infancias, pues se trató de no coincidir con el tema de la migración para lograr una oposición; sin embargo, sí hubo algunos textos que trataron este tema de manera interseccional. Se escogio un corpus académico como corpus de referencia con la expectativa de encontrar una menor presencia de lenguaje emotivo y subjetivo, de modo que lo que suceda en los corpus objetivos refleje las preocupaciones y perspectivas de los usuarios.

6.1.1 Twitter

Hacia el año 2020, cuando se hizo el levantamiento de los datos utilizados en esta tesis, se reportaban 89 millones de usuarios mexicanos de redes sociales de entre 16 a 64 años, entre estos el 61% usaba Twitter, lo que la hacía la sexta red más usada en México (*Digital*, 2020). Además, fue una red social que contó con un reconocido sistema para marcar la pauta a la opinión pública en diferentes momentos de la vida política y pública alrededor del mundo (Campos-Domínguez, 2017).

En octubre de 2022, Elon Musk compró Twitter. A partir de entonces, la red social ha atravesado varios cambios, si bien el más notable es el cambio de nombre a X, hay otros que ya harían diferente esta investigación, probablemente en sus conclusiones, pero sin duda en su metodología. En cuanto a la obtención de datos se refiere, se advierte que anteriormente, mediante la API de Twitter se podía acceder a una cantidad suficiente de

datos, al menos para los fines de un trabajo como este. Hubo también cambios importantes en las formas en que se comunican los usuarios, entre estas destaca la alteración al sistema de verificación de cuentas; anteriormente, este sistema distinguía como cuentas verificadas a figuras públicas, instituciones y medios de comunicación; actualmente, cualquier persona puede comprar la verificación de su cuenta. Probablemente estos cambios pueden dificultar replicar el enfoque original y obligaría a replantear diversas aristas del estudio en un contexto tecnológico cambiante.

En lo que a Twitter se refiere, era una red social que permitía una conversación abierta en la que el usuario se integraba de varias maneras: ya sea que respondiera a otro tweet; o bien, sin vincular su mensaje a ningún otro, se emitía una opinión acerca de los temas de moda (*trending topic*); o incluso, se tuiteaba sin responder a ningún tema, ni a ningún usuario, pero abriendo la posibilidad a nuevas interacciones. Esta dinámica condicionó la recopilación de datos para un corpus de análisis de discurso, puesto que el discurso que interesaba estaba disperso. De modo que para la construcción de dicho corpus se siguieron los pasos mencionados a continuación.

Se utilizó la base de datos COVID México, del Grupo de Ingeniería Lingüística (GIL, UNAM). Se trata de un *listener* construido desde el 21 de abril del 2020, esto es, semanas después de iniciado al encierro por pandemia en el país, hasta el 23 de enero del 2023 (*COVID-19 México*, 2020). Se construyó este sistema con la API de Twitter en su versión gratuita, esto es, solo es posible hacerse con el 1% de los tweets emitidos. A pesar de ello, como se verá a continuación, se obtienen suficientes datos que tienen la virtud también de ser proporcionados de forma aleatoria de modo que es posible construir corpus robustos y diversos. Se optó por esta base de datos puesto que sus criterios de recolección es que los tweets guardados sean publicados a partir de la fecha ya mencionada (pues esto hace un *listener*, recopilar los datos en cuanto son emitidos) y dentro del territorio mexicano. Esto último es posible porque entre los criterios de búsqueda de esta red, había datos geográficos, particularmente las coordenadas del país.

Para la descarga de los datos, se identificaron "semillas" o términos clave para la descarga de tweets sobre el fenómeno migratorio en México (*migr, frontera, Saltillo, Juárez,*

Matamoros, Tapachula, Chiapas, Suchiate, salvadoreñ, hondureñ, caravana, hacinados). Cabe destacar que se consideraron para los términos claves aquellos que puedan hacer referencia a los fenómenos migratorios que han tenido cobertura mediática dentro del periodo en que se recogieron los datos; de hecho se trata de términos clave que sucedieron en encabezados de periódicos del 13 de noviembre del 2018 al 29 de septiembre del 2021 . Así pues, algunos referencían al modo de tránsito relacionado con ingresos sin documentos; gentilicios de los migrantes; y la raíz base "migr" con el propósito de recoger ejemplos donde ocurrieran palabras como migración, migrante(s), inmigrante(s). De los datos recogidos, se hizo una revisión manual que permitió filtrar aquellos tweets que hablan del fenómeno migratorio de los que no, el resultado de este proceso se puede observar en (Tabla 5).

Tabla 5. Tweets descargados por término clave

SEEDS	TWEETS DESCARGADOS	TWEETS FILTRADOS
MIGR	7500	3487
FRONTERA	2000	84
SALTILLO	1000	6
JUÁREZ	1000	2
MATAMOROS	1000	0
TAPACHULA	1500	12
CHIAPAS	2000	1
SUCHIATE	557	10
SALVADOREÑ	1000	94
HONDUREÑ	1000	83
CARAVANA	1000	29
HACINADOS	1000	24

Cabe mencionar que durante este momento se distinguieron tres temáticas:

a) *frontera norte*, para designar aquellos tweets en los que los usuarios identificaran la migración centroamericana o haitiana en los estados del norte de México;
b) *frontera sur*, que trata sobre los mismos orígenes de las personas migrantes del punto anterior pero en los estados de la frontera sur;
c) *migración general* que tratan sobre cualquier fenómeno migratorio humano con la salvedad de los dos tópicos anteriores.

Ahora bien, dado el parámetro por medio de coordenadas, lo que la base de datos COVID garantiza es que los tweets hayan sido emitidos desde México. Mediante el uso de semillas para extraer tweets de la base y, sobre todo, mediante la revisión a ojo humano, lo que se garantizó es la unidad temática del material. No obstante, el objetivo de esta investigación fue analizar el discurso mexicano sobre la migración centroamericana y caribeña ingresando a México, además dicho discurso no debía ser ni de las élites ni de las resistencias, pero, como se concluye, se obtuvo un conjunto de mensajes emitidos por actores diversos como medios de comunicación, y actores políticos y de usuarios que "firman" a título personal. Si bien son estos últimos donde recayó el interés principal, se mantuvieron los tweets de todos los actores. Esta decisión, que puede parecer contradictoria, se debió al hecho de que las posibilidades de comunicación de Twitter permiten un flujo de ida y vuelta entre todos los usuarios, lo que puede estar contribuyendo a las creencias populares sobre la migración. Un ejemplo de tweet polémico para tal objetivo se observa en (12).

(12) @M_OlgaSCordero @SSalud_mx @HLGatell Exigimos cerrar fronteras a inmigrantes y no a la contratación de médicos extranjeros así como la urgente habilitación de sistemas sanitarios en accesos principales a nuestro país. #AMLOseVA

Si bien este tweet parece haber sido emitido por un actor político, probablemente no se trate de un actor identificable en la vida institucional, es decir, no se puede identificar ni como un político, ni como un partido, ni como una institución; más bien se considera como un mensaje emitido por un grupo de personas opositoras a la administración federal (del periodo 2018-2024) que hicieron uso de bots para posicionar una postura contra presidencia, no necesariamente contra migrantes, aunque fuera este un "daño colateral". Esta suposición nace de la lectura de todos los tweets del corpus –para su conformación–, en tal lectura se corroboró que el mensaje aparecía varias veces entre los datos descargados, más o menos modificado.

Hasta aquí las características de uno de los corpus de Twitter utilizados en esta investigación, a saber, el que se usó para la etapa de estudio en corpus, estos es, el Corpus de Twitter no balanceado. Fue llamado así porque, como se verá a continuación, fue necesaria una segunda etapa de clasificación manual de tweets de acuerdo con su polaridad de

sentimientos que, si bien a todos los datos se les dio esta clasificación, no todos fueron usados para el corpus de tweets balanceado (que se utilizó para el análisis de sentimientos), por la disparidad entre tweets con polaridad negativa y positiva.

Una vez filtrados por tema, se continuó con la división de los tweets por su polaridad. El resultado de este proceso fue el corpus de Twitter balanceado (Tabla 4), que es el que se usó para el análisis de sentimientos. Al igual que el proceso anterior, este es un momento que requiere de la observación humana y debe designar la orientación valorativa del hablante hacia el fenómeno migratorio del que discurre. Se fijó para ello una "escala de valores" guiada por criterios. Dicha escala fue binaria, esto es, se distinguió entre tweets xenofóbicos (o de polaridad negativa en torno al fenómeno de migración) o no xenofóbicos (o de polaridad neutra o positiva). Los criterios para establecer ambos criterios se explican a continuación.

Etiqueta 1 (no xenofóbicos):

- **Actitud crítica ante los eventos negativos que viven los migrantes**. En (13), el usuario deja claro que lo que debe estar bajo la mira son los abusos policiacos contra migrantes y no estos últimos; en (14) se condena la presunta participación de policías en actos delictivos contra migrantes.
- **Reporta el fenómeno migratorio sin emitir un juicio**. Los tweets bajo este criterio parecen ser emitidos por medios de comunicación (15) a (17). En esto tweets se observa que se asocia al migrante a eventos negativos, se les considera no xenófobicos, no obstante, porque no se manifiestan a favor de tales adversidades.

(13) Tendremos un seguimiento de los abusos policíacos a la población migrante haitiana, centroamericana en Tijuana? Seguro hay mucha tela de donde cortar.

(14) Terrible que policías estatales en tamaulipas sean los sicarios de 19 migrantes, cuantas policías del país tienen nexos con el crimen organizado? El pronóstico puede ser desolador, tiene mucha chamba la GN antes de que también se contamine 🤦‍♂️🤦‍♂️

(15) Reitera la periodista sonorense @rynram que @PdPagina documentó las condiciones de migrantes en Matamoros.

"Vamos a informar", dice el presidente. https://t.co/Tst9FFq4GH

(16) Continúan agresiones y amenazas hacia migrantes por parte de elementos #Saltillo

(17) Abandonaron en la cajuela de un vehículo a una mujer migrante con hijos, quien había pagado para cruzar ilegalmente a EE.UU. https://t.co/BbGOYG0RDp

Etiqueta 2 (xenofóbicos)

- Descontento explícito contra el ingreso, estadía o paso por México de los migrantes, expresado como:
    - Estereotipo por evento negativo o "característica" del migrante. Se entrecomilla característica, porque es como en (18), donde se generaliza que todos los migrantes roban. En (19) se sentencia, sin atenuaciones, que son delincuentes.
    - Motiva o está a favor de eventos negativos para los migrantes, como el retorno obligado (18) y (21).
    - Creencia de complot político motivando migración. En (20) se dice que les dejan entrara para no molestar al presidente.

(18) Esperamos que si hayan mandado a su país a los migrantes, no los queremos robando en México; y disculpen, pero eso es lo que hacen no discuto que hay mexicanos ladrones, pero al dejar entrar tantas maras, hay muchos criminales, que trabajan para los narcos y por@au cuenta

(19) @adn40 son delincuentes y deberían de ser tratados como tales sin ningún privilegio, convirtieron tapachula en el chiquero más grande de la república mexicana, los detienen robando, derechos humanos los libera, claro no son todos pero la gran mayoria

(20) Avanza caravana de haitianos, hondureños, cubanos, guatemaltecos, salvadoreños y diferentes nacionalidades, rumbo a los Estados Unidos. Salieron hoy por la mañana de Tapachula. Les dieron vía libre para que no molesten al peje.

(21) Estubo bien una limpia mandar a los migrantes a su ciudad de origen todos internamente nos alegramos por quitar malos inmigrantes claro las reglas tomen sus excepciones

Como se puede observar en los ejemplos tanto para los criterios de la etiqueta no xenofóbico, como para los xenofóbicos, los datos no pasaron por una etapa de homologación ortográfica.

El etiquetado manual se llevó a cabo, primero, por la autora de esta tesis y en un segundo momento por una revisora que fue escogida por su formación lingüística y su interés en temáticas sociales. Ambas observaciones fueron comparadas mediante el estadístico Kappa score, que mide la concordancia inter-observador .

Figura 5. Fórmula Kappa score

$$K = \frac{P_0 - P_e}{1 - P_e}$$

P_0 es la proporción de acuerdos observados;

P_e la proporción de acuerdos esperados por azar.

De este modo, el coeficiente de Kappa puede tomar medidas entre -1 y +1. Si K=0 la concordancia observada es la se espera precisamente por causa del azar; entre tanto, mientras más cercano a +1 mayor es el grado de concordancia entre los observadores (Vindell, 2021). La escala de valores completa que puede tomar K se muestra en

Tabla 6. Valores para el coeficiente de Kappa.

Tabla 6. Valores para el coeficiente de Kappa

KAPPA	INTERPRETACIÓN
0-0.2	Ínfima concordancia
0.2-0.4	Escasa concordancia
0.4-0.6	Moderada concordancia
0.6-0.8	Buena concordancia
0.8-1.0	Muy buena concordancia

Nota: Msc Juan José Vindell. (2021). *Kappa de Cohen en R*. https://rpubs.com/VINDELL2981/kappa

La concordancia entre ambas observadoras obtuvo un valor de 0.64. En otras palabras, una buena concordancia. Con esto en mente,se continuó con la constitución del corpus de esta red social.

Finalmente, el balanceo es un proceso por el cual se asegura proveer al modelo de clasificación automática una muestra representativa de cada etiqueta. Después de anotar y revisar los tweets, se obtuvo un conjunto confiable de 712 mensajes en Twitter, donde 294 eran xenofóbicos y 418 no lo eran, es decir, el corpus de Twitter balanceado. Este set fue el que alimentó el proceso de análisis de sentimientos (**¡Error! No se encuentra el origen de la referencia.**). El no balanceado fue utilizado para las etapas de descripción de patrones desde la Lingüística de corpus y el Análisis Crítico del Discurso (**¡Error! No se encuentra el origen de la referencia.**).

6.1.2 YouTube

A inicios del año 2020, el 96% de los 80.60 millones usuarios de Internet de entre 16 a 64 años usaban YouTube, lo que la hacía la red social más utilizada en México (*Digital*, 2020).

Es una plataforma que no requiere tener una cuenta para usarla, en esta modalidad el usuario podrá ver la mayor parte del contenido que existe en ella. Sin embargo, para interactuar en ella es necesario ser usuario registrado y crear un canal. Este último es el que permite a los usuarios subir sus propias producciones; sin un canal es posible, además, interactuar desde los comentarios. En otras palabras, el usuario (que puede ser una persona, o cualquier actor de la iniciativa privada o pública) sube un vídeo, decide si este está abierto o no a comentarios; en el primer caso, cualquier usuario comenta al video, o bien responde a otro usuario.

Nacida entre estas condiciones, la comunicación que Youtube hace posible a) es electrónicamente mediada; b) no inmediata, puesto que los videos estarán ahí hasta que los propietarios decidan y, mientras esto ocurra y los videos queden abiertos, cualquier usuario puede "llegar" a comentar; c) más importante, la comunicación que esta dinámica hace posible es cerrada temáticamente al asunto que el video haya puesto sobre la mesa. Más importante porque esta cuestión facilita la recopilación de los datos además de en sentido técnico, en un sentido práctico: ya se está hablando del discurso de la migración.

No obstante, una vez que el video está en la plataforma, y los comentarios están activos, cualquier persona que cumpla con los requisitos del párrafo anterior, puede comentar; ello hace que los comentarios que forman parte del corpus no sean solo de usuarios mexicanos. Sin embargo, los comentarios tienden a contar con contexto para saber esta información, ya sea porque el usuario se nombre desde su nacionalidad (22) y (23); por elementos deícticos, como (24) y (25), donde la elección léxica del verbo *regresen* y de la perífrasis verbal *quieren entrar,* así como sus flexiones por persona, dan cuenta de que el usuario se está conceptualizando en México; referencia a símbolos de la mexicanidad, patrios (26), o no (27). Además, contamos con la oportunidad de revisar el comentario en contexto más amplio en AntConc ((Anthony, 2022) como en la Ilustración 7.

Ilustración 7. Concordancias en AntConc

.

(22) Yo soy hondureño y todo esto pasa por el presidente que tenemos en nuestro país

(23) Yo soy mexicana, y la verdad yo quisiera ayudarlos…

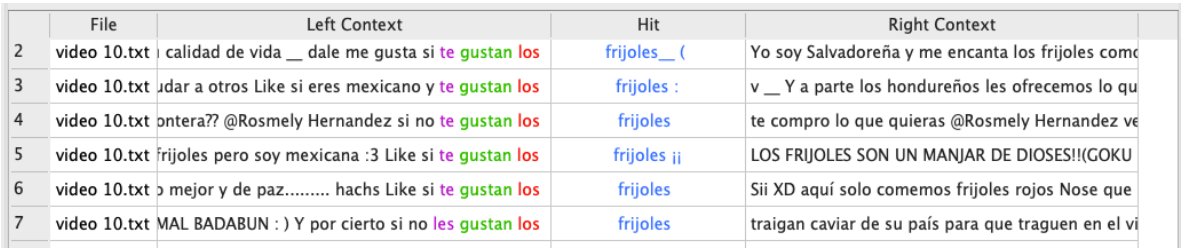
(24) Nel regresen a su país. Aquí muy apenas hay trabajo para la gente

(25) Pero por que no quieren entrar de 50 and 50 como Mexico se los pide?

(26) hora si como dice nuestro himno. Nacional Mexicanos al grito de guerra. Nos están invadiendo y el presidente donde está?

(27) Y por cierto si no les gustan los frijoles traigan caviar de su país para que traguen en el viaje.

Ilustración 7. Concordancias en AntConc



	File	Left Context	Hit	Right Context
2	video 10.txt	calidad de vida __ dale me gusta si te gustan los	frijoles__ (	Yo soy Salvadoreña y me encanta los frijoles comc
3	video 10.txt	udar a otros Like si eres mexicano y te gustan los	frijoles :	v __ Y a parte los hondureños les ofrecemos lo qu
4	video 10.txt	ontera?? @Rosmely Hernandez si no te gustan los	frijoles	te compro lo que quieras @Rosmely Hernandez ve
5	video 10.txt	frijoles pero soy mexicana :3 Like si te gustan los	frijoles ¡¡	LOS FRIJOLES SON UN MANJAR DE DIOSES!!(GOKU
6	video 10.txt	o mejor y de paz..... hachs Like si te gustan los	frijoles	Sii XD aquí solo comemos frijoles rojos Nose que
7	video 10.txt	MAL BADABUN :) Y por cierto si no les gustan los	frijoles	traigan caviar de su país para que traguen en el vi

Los videos de los que se extrajeron los comentarios son 12 reportajes de diferentes medios de comunicación sobre caravanas centroamericanas sucedidas desde el 10 de octubre del 2018 hasta el 7 de agosto del 2020. Sus comentarios fueron extraídos desde YouTube Comment Scrapper (*YouTube Data Tools*, s. f.). Si bien es posible contar, además, con los videos y subtítulos, no se revisan aquí estos contenidos porque no estaba entre nuestros objetivos analizar el discurso de la prensa; no obstante, la Tabla 7 cuenta con el título del video y su medio de procedencia para conocer, si bien someramente, el contenido al que responden los comentarios.

Tabla 7. Composición del corpus de YouTube

#	TÍTULO	MEDIO	# DE PALABRAS
1	Caravana migrante, primera movilización en masa centroamericana durante del 2020	France 24 español	12,049
2	La Nueva Caravana Migrante llega a México	El País	7,178
3	Caravana migrante entra a la fuerza a México	Milenio	37,091
4	Frenen caravana migrante al sur de México	Imagen Noticias	7,070
5	Hondureños presionan para cruzar a México	Noticias Telemundo	110,010
6	México, país destino de migrantes	Imagen Noticias	10,113
7	Migrantes centroamericanos varados en el Estado de México	AJ+Español	671

8	Migrantes en Chiapas: cómo tratan a los centroamericanos en Tapachula	Plumas Atómicas	992
9	Fuerzas mexicanas detienen 800 migrantes de la caravana 2020	El Tiempo	15,642
10	La verdadera cara de los migrantes que nadie muestra	Badabun	142,937
11	Empleo en México “el sueño hecho realidad” de algunos migrantes	AFP Español	6,018
12	Las duras historias de los migrantes centroamericanos	RT	1,886

YouTube Comment Scrapper descarga varios tipos de datos que pueden servir a múltiples análisis que den cuenta del comportamiento de los usuarios en esta red social, por ejemplo el id del usuario, información sobre si se trata de una respuesta a otro comentario, un conteo de cuántos “likes” recibe, el nombre del autor, el texto (es decir, el comentario), el autor del canal, el conteo del número de respuesta que recibe el comentario. Aquí interesaron únicamente los comentarios y, gracias a ellos, el corpus de Youtube está compuesto por 351, 637 palabras.

6.2 Experimentos para Análisis de sentimientos

Una vez conformados los corpus, el análisis de sentimientos fue el primer acercamiento a ellos.

Para el entrenamiento de los datos, se utilizó el corpus balanceado de Twitter como conjunto de datos, que cuenta con 712 tweets de los cuales 418 fueron previamente etiquetados como no xenofóbicos y 294 como xenofóbicos (Tabla 4), los criterios para ambas etiquetas se explicaron en la sección **¡Error! No se encuentra el origen de la referencia. ¡Error! No se encuentra el origen de la referencia.**. Entre tanto, para la extracción de características, la clasificación y la obtención de una lista de ítems con mayor ganancia de información se utilizó el software WEKA (Frank et al., 2016).

Durante el proceso de extracción de características, se consideró para todos los experimentados aquí realizados:

- Mil características vectorizadas,

- dentro de estas fueron consideradas todas las posibles flexiones de la palabra (no se usó un stemmer, o identificador de raíces morfológicas),

Por otro lado, también se consideraron los siguientes aspectos, para los que se adelanta que no hubo cambios en los resultados obtenidos:

- mayúsculas y minúsculas como en el texto original; y texto homologado a minúsculas,
- con y sin stopwords, o palabras función o vacías

Además, se experimentó con mayor éxito:

- con unigramas;
- con unigramas y bigramas;
- con unigramas, bigramas, trigramas y tetragramas.

Entre tanto, los algoritmos que se utilizaron para los modelos de clasificación fueron Naïve Bayes, Naïve Bayes Multinominal, SMO y Logistic (Tabla 7).

Tabla 7. Características extraídas y clasificadores utilizados para el análisis de sentimientos

	NAÏVE BAYES	***NAÏVE BAYES MULTINOMINAL***	***LOGISTIC***	***SMO***
UNIGRAMAS	✓	✓	✓	✓
1-2 GRAMAS	✓	✓	✓	✓
1-4 GRAMAS	✓	✓	✓	✓

El objetivo en este punto fue comprobar cuál de estos algoritmos, acompañado de cuáles características, distinguía mejor los mensajes xenofóbicos de los no xenofóbicos. En la Tabla 8 puede verse cuáles fueron las características vectorizadas en los experimentos de cada clasificador.

Tabla 8. Experimentos de análisis de sentimientos realizados.

		Naïve Bayes	*Naïve Bayes Multinominal*	*Logistic*	*SMO*
Unigramas	*Mayúsculas y minúsculas, con stopwords*	✓	✓	✓	✓
	Mayúsculas y minúsculas, sin stopwords	✓	✓	✓	✓
	Sin mayúsculas y minúsculas, con stopwords	✓	✓	✓	✓
	Sin mayúsculas y minúsculas, sin stopwords	✓	✓	✓	✓
1-2 gramas	*Mayúsculas y minúsculas, con stopwords*	✓	✓	✓	✓
	Mayúsculas y minúsculas, sin stopwords	✓	✓	✓	✓
	Sin mayúsculas y minúsculas, con stopwords	✓	✓	✓	✓
	Sin mayúsculas y minúsculas, sin stopwords	✓	✓	✓	✓
1-4 gramas	*Mayúsculas y minúsculas, con stopwords*	✓	✓	✓	✓
	Mayúsculas y minúsculas, sin stopwords	✓	✓	✓	✓
	Sin mayúsculas y minúsculas, con stopwords	✓	✓	✓	✓
	Sin mayúsculas y minúsculas, sin stopwords	✓	✓	✓	✓

El conjunto de datos, además, se dividió en 10 pliegues para validación cruzada. Se trata de proceso por el cual se divide el conjunto de entrenamiento por partes iguales. En este caso fueron 10 partes iguales. Ello significó que se entrenó el modelo con las primeras 9 partes y luego se validó el entrenamiento con la décima parte; posteriormente se entrenó secuencialmente con 9 partes para poder evaluar una décima distinta, esto se repitió 10 veces, siempre dejando libre un campo para la validación del modelo. El desempeño de cada

una de estas clasificaciones es promediado y con ello se obtiene el valor del desempeño global obtenido (Romero-Vega et al., 2021).

Además de los resultados de los experimentos, de este proceso interesó obtener la lista de ganancia de información de las características extraídas del modelo con el mejor resultado. La ganancia de información es una medida utilizada cuando es necesario conocer la relevancia que tiene un atributo dentro de un conjunto de datos. En concordancia con las cuatro fases de la investigación aquí planteadas, se aplicó esta medida con la intención de identificar qué términos lingüísticos fueron detectados como relevantes para el discurso xenofóbico en Twitter por los clasificadores; el objetivo fue tomar tales elementos y revisar sus frecuencias normalizadas en el corpus de Twitter no balanceado, el corpus de YouTube y el corpus de referencia. Los lineamientos de este proceso pueden revisarse en la próxima sección de este capítulo.

6.3 Análisis de Patrones Léxicos en Corpus

Uno de los objetivos específicos de esta tesis fue incorporar los resultados de un análisis de sentimientos a un estudio de corpus guiado por datos. Esto se hizo mediante el uso de una lista de elementos con ganancia de información del mejor modelo de clasificación obtenido. Precisamente, en esta sección describimos cómo fue utilizada para guiar el análisis en corpus. Este análisis continuó con una comparación de la ocurrencia –en dos corpus de redes sociales y uno de referencia– de algunos elementos con ganancia de información, proceso que también se describe a continuación.

La lingüística de corpus cuenta con la lista de palabras clave de un corpus entre sus herramientas tradicionales. Se obtiene a partir de la comparación de la ocurrencia y frecuencia de las palabras que componen al corpus objeto de estudio con las de un corpus de referencia. Por ello, se espera que las palabras resultantes de esa comparación sean representativas del corpus estudiado y como representativas pueden guiar también análisis posteriores; por ejemplo, sobre las colocaciones de esos términos clave, o su revisión en contextos más amplios, como las concordancias.

En otras palabras, la lista de palabras clave lo que hace es dirigir la atención del investigador a un conjunto de términos representativos de un discurso. Ese papel lo tomó aquí la ya mencionada lista de elementos con ganancia de información (ver apéndice **¡Error! No se encuentra el origen de la referencia.**). De ella se tomaron los primeros 100 elementos que consisten en n-gramas, o cadenas de palabras, con longitud de hasta 4 palabras. Entre estos, se descartaron aquellos con uso funcional, de modo que el análisis se centró en los que tenían alguna carga semántica sin necesidad de revisarlo en contexto más amplios. Los que restaron se dividieron en campos semánticos de POLÍTICA, LUGAR, COLECTIVIDAD y MOVIMIENTO, la descripción de cada uno de estos campos puede revisarse en la sección **¡Error! No se encuentra el origen de la referencia.** Descripción semántica de los elementos con ganancia de información obtenidos del análisis de sentimientos.

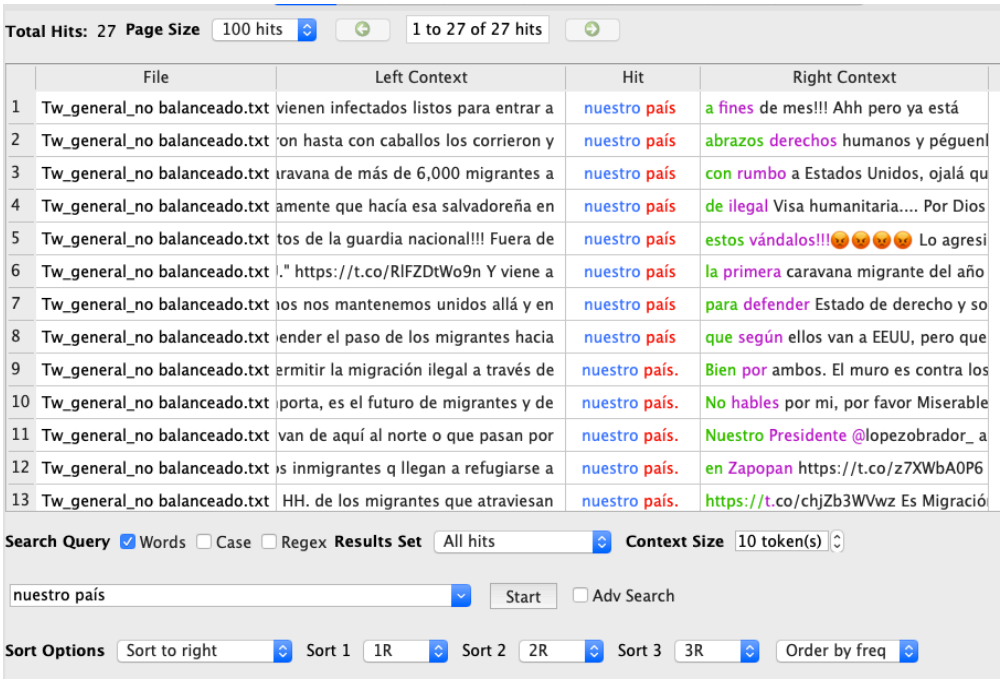
Posteriormente, para conocer si los n-gramas obtenidos eran representativos del discurso xenofóbico en Twitter, o representativos del discurso xenofóbico en redes sociales, se revisó la frecuencia normalizada sobre un millón de palabras de los elementos de cada campo semántico en los dos corpus, esto es, Corpus de Twitter no balanceado y Corpus de YouTube, así como en uno de referencia, es decir, el Corpus de Referencia de SketchEngine.

Una vez agrupados los campos semánticos, se seleccionó el campo semántico LUGAR para un análisis de concordancias, para el que se utilizó AntConc (Anthony, 2022). De este campo se seleccionaron tres n-gramas que compartieran dos características, a saber, que no tuvieran algún lexema que coincidiera con alguna de las semillas de búsqueda de Tweets en la base de datos COVID para evitar hacer un análisis circular; además, fue necesario que dichos elementos ocurrieran –de acuerdo con su frecuencia relativa– más veces en ambas redes sociales en comparación con el corpus de referencia.

Este análisis se hizo con dos objetivos en mente. El primero de ellos fue el de detectar los patrones gramaticales y semánticos en los que ocurrían los n-gramas; el segundo fue el de encontrar asociaciones semánticas y anidaciones en las que dichos patrones sucedieran. Para ello, en primer lugar, por red social se revisaron los elementos en sus concordancias ordenadas con una ventana de tres tanto a la derecha (

Ilustración 8) como a la izquierda (Ilustración 9).

Ilustración 8. Concordancias para "nuestro país" ordenadas a la derecha



Total Hits: 27 Page Size 100 hits 1 to 27 of 27 hits

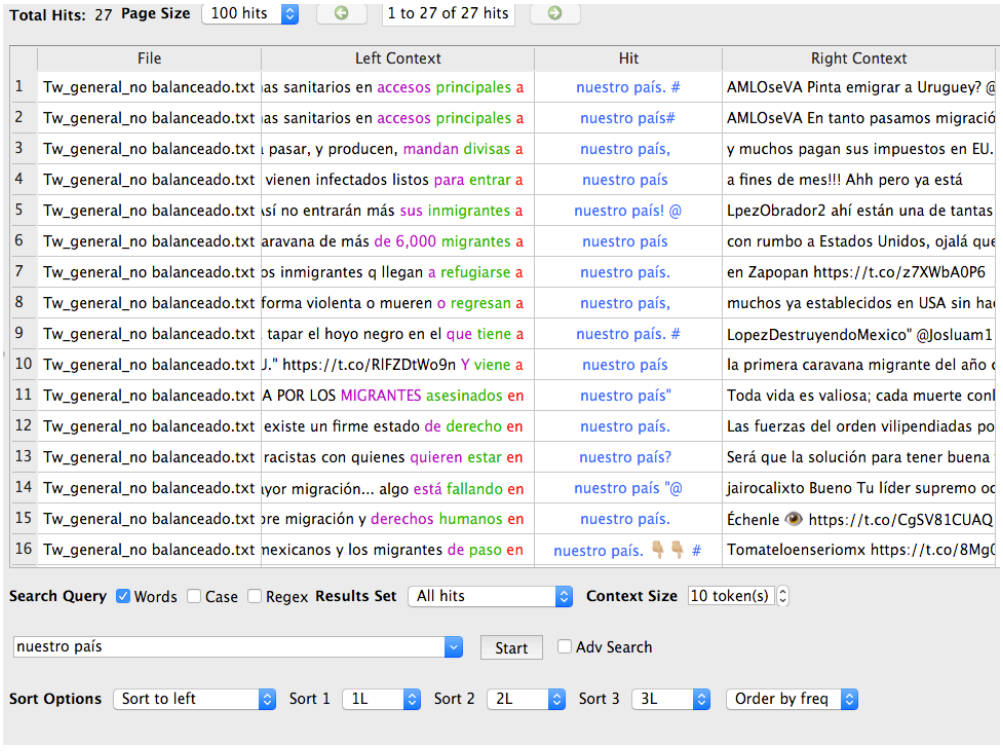
	File	Left Context	Hit	Right Context
1	Tw_general_no balanceado.txt	vienen infectados listos para entrar a	nuestro país	a fines de mes!!! Ahh pero ya está
2	Tw_general_no balanceado.txt	on hasta con caballos los corrieron y	nuestro país	abrazos derechos humanos y péguenl
3	Tw_general_no balanceado.txt	aravana de más de 6,000 migrantes a	nuestro país	con rumbo a Estados Unidos, ojalá qu
4	Tw_general_no balanceado.txt	amente que hacía esa salvadoreña en	nuestro país	de ilegal Visa humanitaria.... Por Dios
5	Tw_general_no balanceado.txt	tos de la guardia nacional!!! Fuera de	nuestro país	estos vándalos!!!😡😡😡😡 Lo agresi
6	Tw_general_no balanceado.txt	." https://t.co/RlFZDtWo9n Y viene a	nuestro país	la primera caravana migrante del año
7	Tw_general_no balanceado.txt	os nos mantenemos unidos allá y en	nuestro país	para defender Estado de derecho y so
8	Tw_general_no balanceado.txt	ender el paso de los migrantes hacia	nuestro país	que según ellos van a EEUU, pero que
9	Tw_general_no balanceado.txt	ermitir la migración ilegal a través de	nuestro país.	Bien por ambos. El muro es contra los
10	Tw_general_no balanceado.txt	porta, es el futuro de migrantes y de	nuestro país.	No hables por mi, por favor Miserable
11	Tw_general_no balanceado.txt	van de aquí al norte o que pasan por	nuestro país.	Nuestro Presidente @lopezobrador_ a
12	Tw_general_no balanceado.txt	s inmigrantes q llegan a refugiarse a	nuestro país.	en Zapopan https://t.co/z7XWbA0P6
13	Tw_general_no balanceado.txt	HH. de los migrantes que atraviesan	nuestro país.	https://t.co/chjZb3WVwz Es Migració

Search Query Words Case Regex Results Set All hits Context Size 10 token(s)

nuestro país Start Adv Search

Sort Options Sort to right Sort 1 1R Sort 2 2R Sort 3 3R Order by freq

Ilustración 9. Concordancias para "nuestro país" ordenadas a la izquierda



Total Hits: 27 Page Size 100 hits 1 to 27 of 27 hits

	File	Left Context	Hit	Right Context
1	Tw_general_no balanceado.txt	as sanitarios en accesos principales a	nuestro país. #	AMLOseVA Pinta emigrar a Uruguey? @
2	Tw_general_no balanceado.txt	as sanitarios en accesos principales a	nuestro país#	AMLOseVA En tanto pasamos migració
3	Tw_general_no balanceado.txt	pasar, y producen, mandan divisas a	nuestro país,	y muchos pagan sus impuestos en EU.
4	Tw_general_no balanceado.txt	vienen infectados listos para entrar a	nuestro país	a fines de mes!!! Ahh pero ya está
5	Tw_general_no balanceado.txt	sí no entrarán más sus inmigrantes a	nuestro país! @	LpezObrador2 ahí están una de tantas
6	Tw_general_no balanceado.txt	aravana de más de 6,000 migrantes a	nuestro país	con rumbo a Estados Unidos, ojalá que
7	Tw_general_no balanceado.txt	os inmigrantes q llegan a refugiarse a	nuestro país.	en Zapopan https://t.co/z7XWbA0P6
8	Tw_general_no balanceado.txt	forma violenta o mueren o regresan a	nuestro país,	muchos ya establecidos en USA sin ha
9	Tw_general_no balanceado.txt	tapar el hoyo negro en el que tiene a	nuestro país. #	LopezDestruyendoMexico" @Josluam1
10	Tw_general_no balanceado.txt	J." https://t.co/RlFZDtWo9n Y viene a	nuestro país	la primera caravana migrante del año
11	Tw_general_no balanceado.txt	A POR LOS MIGRANTES asesinados en	nuestro país"	Toda vida es valiosa; cada muerte conl
12	Tw_general_no balanceado.txt	existe un firme estado de derecho en	nuestro país.	Las fuerzas del orden vilipendiadas po
13	Tw_general_no balanceado.txt	racistas con quienes quieren estar en	nuestro país?	Será que la solución para tener buena
14	Tw_general_no balanceado.txt	yor migración... algo está fallando en	nuestro país "@	jairocalixto Bueno Tu líder supremo oc
15	Tw_general_no balanceado.txt	re migración y derechos humanos en	nuestro país.	Échenle 👁 https://t.co/CgSV81CUAQ
16	Tw_general_no balanceado.txt	nexicanos y los migrantes de paso en	nuestro país. 👇👇 #	Tomateloenseriomx https://t.co/8Mg0

Search Query Words Case Regex Results Set All hits Context Size 10 token(s)

nuestro país Start Adv Search

Sort Options Sort to left Sort 1 1L Sort 2 2L Sort 3 3L Order by freq

De estos contextos lingüísticos inmediatos el interés recayó en las frases verbales en las que ocurrieron los n-gramas seleccionados. Con esto, se obtuvo información gramatical. De estas frases verbales se extrajó información semántica. La primera de esta información permitió agrupar los tipos de verbos con los que sucedía cada n-grama, de modo que se encontró asociaciones semánticas. Como se mencionó anteriormente, una asociación semántica comparte con la colocación la constante coocurrencia de dos elementos, solo que, para la anidación, la relación entre los dos elementos estará situada a un nivel más abstracto, pues no solo se trata de dos palabras frecuentemente asociadas, sino de sentidos, como en (7) a (9) página 32, donde se observan cuatro sets semánticos que ocurren constantemente entre sí: LUGAR PEQUEÑO, TIEMPO DE VIAJE, VEHÍCULO y LUGAR GRANDE. En otras palabras, una asociación semántica está compuesta por sets semánticos, valga la redundancia, constantemente coocurrentes. Este análisis partió de tres n-gramas obtenidos del campo semántico LUGAR, esto es, se trataba de ver con qué otros sets ocurría. Se observó que en estos últimos había carga léxica de movimiento, de daño, de protección, de problemas y soluciones; esta información permitió concluir tres grandes sets semánticos, a saber, MOVIMIENTO, PELIGRO y ACTITUD. De las verbos que constituyeron estos sets, también se buscaron rasgos de agentividad, así como información sobre el sujeto gramatical de tales acciones, estos resultados se pueden consultar en la sección **¡Error! No se encuentra el origen de la referencia. ¡Error! No se encuentra el origen de la referencia.** (p. **¡Error! Marcador no definido.**).

Hasta aquí, el análisis de corpus planteado sugiere una combinación de técnicas cuantitativas y cualitativas con diferentes grados de automatización. Con ello se buscó la comprensión de las expresiones xenofóbicas como un discurso cohesionado, compartido por los usuarios. Para contribuir con dicha explicación es necesario explicar cómo es que suceden las ocurrencias que se encontraron gracias a la metodología recién descrita, por esa razón, el análisis culmina con una descripción de las propiedades léxicas de los n-gramas para comprender su contribución a la construcción del discurso xenofóbico (ver **¡Error! No se encuentra el origen de la referencia. ¡Error! No se encuentra el origen de la referencia.**).

7 Resultados

Esta investigación exploró el discurso de los mexicanos en redes sociales respecto al tema de la migración centroamericana en tránsito hacia Estados Unidos. Si se define cultura como el modo en que los individuos y grupos sociales se relacionan con la realidad, entonces, a partir de los datos lingüísticos examinados en este trabajo, se puede identificar la presencia muy notable de un discurso discriminatorio, ofensivo y xenofóbico.

Este capítulo da respuesta a las preguntas de investigación de este proyecto, empezando por la más general e importante: ¿cómo se constituye el discurso xenofóbico en Twitter y YouTube? Un discurso es una producción intertextual que construye aspectos del mundo (Fairclough, 2010). En este caso esa producción intertextual se conforma con las opiniones acumuladas de autores anónimos en redes sociales. Así, primero vamos a mostrar cómo tal acumulación ocurre.

Nuestro enfoque metodológico emplea técnicas de PLN así como técnicas propias de la lingüística de corpus. Así, la primera sección de este capítulo (**¡Error! No se encuentra el origen de la referencia. ¡Error! No se encuentra el origen de la referencia.**) reporta desde el PLN un análisis de sentimientos. En este apartado se muestran los resultados en razón de su polaridad de sentimientos xenofóbicos y no xenofóbicos (**¡Error! No se encuentra el origen de la referencia. ¡Error! No se encuentra el origen de la referencia.**), particularmente se exponen los algoritmos que generaron una mejor clasificación. Estos resultados se comparan con otros obtenidos en trabajos de clasificación de sentimientos xenofóbicos en conjuntos de datos en español en el apartado (**¡Error! No se encuentra el origen de la referencia. ¡Error! No se encuentra el origen de la referencia.**).

La segunda sección de este capítulo (7.2 Descripción semántica de los elementos con ganancia de información obtenidos del análisis de sentimientos) presenta, desde la lingüística de corpus, un análisis de las colocaciones, las asociaciones (o prosodia) semánticas y n-gramas con que aparecen los términos del listado de palabras con ganancia de información. Este listado lo conforman los elementos cuya probabilidad de ocurrir en una expresión xenofóbica es mayor que en las expresiones no xenofóbicas y fue utilizado aquí a manera de la tradicional lista de palabras clave de la lingüística de corpus. En otras palabras, fue el punto de partida de un análisis descriptivo detallado y crítico de las particularidades

del discurso xenofóbico mexicano que comenzó con la división en campos semánticos de tal lista, continuó con la revisión de uno de tales campos (a saber, LUGAR) en sus anidaciones gramaticales y semánticas, y concluyó con una explicación sobre cómo es que los elementos léxicos de la lista están preparados para coocurrir y cómo, a raíz de dicha preparación *favorecen* un discurso xenofóbico.

7.1 Análisis de sentimientos: detección de discurso xenofóbico en Twitter

En esta sección se revisarán los resultados de los experimentos del análisis de sentimientos. Con ello se muestra la primera fase de esta investigación, al tiempo que se responde al primer objetivo de la misma, a saber, identificar de manera automática sentimientos sobre la migración en México. En esta primera fase se buscó, además de una clasificación automática de sentimientos, acercarse a los elementos lingüísticos más representativos de los mensajes xenofóbicos expresados en dos redes sociales mediante la ganancia de la información de estos. De este modo, uno de los resultados del análisis de sentimientos da pie a continuar con la fase 2, al tiempo que establece las bases para investigar las relaciones entre técnicas propias de la lingüística computacional y la lingüística de corpus.

Para dar cuenta de estos objetivos se revisan los resultados de los experimentos planteados en la sección **¡Error! No se encuentra el origen de la referencia. ¡Error! No se encuentra el origen de la referencia.**. Esto es, se verán los resultados de aplicar diferentes técnicas de clasificación automática aplicadas al corpus de Twitter balanceado (Tabla 4), que consiste en un conjunto de datos de 712 tweets manualmente clasificados como xenofóbicos y no xenofóbicos. De igual manera, se revisará la eficacia de dichas técnicas al usar representaciones de los tweets que consideren distintas longitudes de n-gramas. También se presentará una comparación de los resultados obtenidos en esta investigación en relación con antecedentes relevantes (**¡Error! No se encuentra el origen de la referencia. ¡Error! No se encuentra el origen de la referencia.**), pues, como se mostrará, no solo se logró cumplir con el primer objetivo específico de esta investigación, sino que se lograron los mejores resultados para tareas con datos en español entre los experimentos de los que se tuvieron noticias a la finalización de este proyecto.

7.1.1 Resultados de los experimentos de detección automática

Los experimentos se realizaron con los algoritmos Naïve Bayes, Naïve Bayes Multinominal, SMO y Logistic (Tabla 8). Como mencionábamos en la sección de metodología, cada uno de estos clasificadores fue puesto a prueba con mil características vectorizadas, sin atrevesar por un proceso de stemmatización ni lematización ya que en investigaciones anteriores el rendimiento de estas frente a la implementación de n-gramas es menor (Pitropakis et al., 2020). Además, cada uno de los clasificadores se probó tanto con mayúsculas y minúsculas, como con todo el texto homologado a minúsculas, así como con y sin stopwords; no obstante, no se revisarán aquí los resultados para estas pruebas, dado que no obtuvieron resultados significativos.

En cambio, lo que se presenta en la Tabla 9 corresponde a los resultados obtenidos de los experimentos que consideraron cada uno de los cuatro clasificadores con las tres diferentes extracciones de características con n-gramas de distintas longitudes. Tales resultados muestran los valores de la *exactitud*, esto es, el porcentaje de instancias clasificadas correctamente, ya sea que se trate de un tweet xenofóbico o no xenofóbico.

Tabla 9. Resultados de los experimentos de clasificación

	NAÏVE BAYES	*NAÏVE BAYES MULTINOMINAL*	*LOGISTIC*	*SMO*
UNIGRAMAS	47.33%	74.29%	71.34%	70.36%
1-2 GRAMAS	68.53%	79.91%	78.08%	70.36%
1-4 GRAMAS	69.52%	***80.47%***	NA	71.62%

En rasgos generales, se observa que los resultados mejoraron una vez que se implementó la combinación de n-gramas de diferentes tamaños; este incremento se observa con el uso tanto de unigramas como bigramas para todos los algoritmos con excepción de SMO. Nuevamente, estos resultados mejoraron una vez que se utilizó además trigramas y tetragramas para tres de los clasificadores. El mejor resultado, de hecho, se obtuvo con esta representación de características con el clasificador Naïve Bayes Multinominal, el cual alcanzó una exactitud del 80.47%.

A continuación, se presetarán otros métricas, únicamente para el modelo con mejores resultados, para entender cómo funcionó el clasificador. Para ello, es necesario, en primer lugar, detenerse en la distribución de verdaderos y falsos positivos, así como de verdaderos y falsos negativos, misma que se puede revisar en Ilustración 10. Para este trabajo, un tweet positivo, es aquel etiquetado como xenofóbico, dado que es el lenguaje xenofóbico el que es de interés en este estudio; uno negativo, entre tanto, es uno designado como no xenofóbico.

Los tweets que fueron clasificados de mejor manera fueron los no xenofóbicos (54.49%). Esto es esperable, cabe recordar que, si bien se intentó balancear el conjunto de datos utilizado (Tabla 4), la clase negativa estuvo mayormente representada. Los tweets totales que corresponden a la clase no xenofóbico son los aquí identificados como verdaderos negativos (388) y falsos positivos (30), esto es, el 58.7% de los datos totales del corpus balanceado de tweets. Entre tanto, los tweets xenofóbicos del data set son aquellos que la matriz de confusión identifica como verdaderos positivos (185) y falsos negativos (109), es decir, el 41.3% del corpus. En otras palabras, la probabilidad independiente de la clase no xenofóbica es de 0.59, mientras que la de la clase xenofóbica es 0.41.

Ilustración 10. Matriz de confusión del mejor modelo obtenido

	PREDICCIÓN NEGATIVOS	PREDICCIÓN POSITIVOS
CASOS NEGATIVOS	VERDADERO NEGATIVO **54.99%** (388 tweets)	FALSO POSITIVO **4.21%** (30 tweets)
CASOS POSITIVOS	FALSO NEGATIVO **15.30%** (109 tweets)	VERDADERO POSITIVO **25.98%** (185 tweets)

En cambio, una revisión de la eficacia de la clasificación de la métrica de *precision*, que se refiere al porcentaje de tweets que son efectivamente xenofóbicos de entre los clasificados como tales (TP (185) + FP (30)), muestra que este modelo logra que tal métrica alcance un valor de 86%. Aunado a ello, la exhaustividad nos dirá el porcentaje de tweets xenofóbicos sobre los tweets que de hecho corresponden a esta clase (TP (185) + FN (109); es decir, el algoritmo logró clasificar un 63% de todos los tweets correspondientes a la clase xenofóbico. Y una medida balanceada de las dos métricas anteriores, la F1-score (2* precisión + exhaustividad/ precisión + exhaustividad), es del 72%.

7.1.2 Comparación de los resultados de la clasificación con el estado del arte

Para entender los resultados obtenidos, estos se revisan a la luz de los rendimientos obtenidos por los trabajos anteriores mencionados en 4.1 Detección de Discursos Xenofóbicos. En tanto la intención es comprender los hallazgos de este trabajo en un entorno comparable, también se revisarán en este apartado las métricas de exactitud, la precisión, la exhausitividad y la F1-score de los antecedentes, en la medida en que sea posible. Esta información se muestra en Tabla 10 y, por lo que se observa ahí, como por lo que se explicará a continuación, puede afirmarse que el modelo utilizado en esta tesis es sólido y competitivo en tareas semejantes hechas con datos de la lengua española.

Tabla 10. Mejores modelos en la revisión de literatura

	Exactitud	**Precisión**	**Exhaustividad**	**F1-score**
Arcila Calderón et al. (2020)	74.64%	79.88%	72.23%	75.86%
de Paula y Schlicht (2021)	69.79%	89.28%	9.49%	15.47%
Pitropakis et al. (2020)	NA	86%	81%	84%
Plaza-Del-Arco et al. (2020)	71.1%	70.7%	71.3%	70.7%
Este estudio	80.47%	86%	63%	72%

Se decía en la sección **¡Error! No se encuentra el origen de la referencia. ¡Error! No se encuentra el origen de la referencia.**, que el uso de diferentes medidas evaluadoras

dependía de varios factores, entre ellos, los objetivos para los que la clasificación se realizó. Para fines forenses, particularmente dentro del contexto anglosajón (Solan & Tiersma, 2005), la precisión es una medida especialmente importante. En ese sentido, este trabajo, que alcanza en esta medida un valor del 86%, es tan bueno como el de Pitropakis et. al (2020), que utiliza una base de datos de inglés; de entre los trabajos que utilizan recursos en español queda solo debajo de Paula y Schlicht (2021).

No obstante, como también se ha mostrado anteriormente, la diferentes métricas dan pesos diferenciados a las cuatro categorías en la **¡Error! No se encuentra el origen de la referencia.**, por ello es conveniente presentarlas ponderadas y neutralizadas a la luz de otras métricas. Particularmente, en lo que se refiere a la precisión, tiene entre sus desventajas que puede estar influenciada ante un corpus desbalanceado que resulta conflictivo, por ejemplo, cuando se cuenta con un mayor número de instancias de la clase no relevante. En ese sentido, las métricas de la exhaustividad y el F1-score son útiles para ponderar los resultados. Al respecto, observamos que si bien de Paula y Schilicht (2021) obtuvieron la precisión más alta, observan la exhaustividad y el F1-score más bajos (9.49% y 15.47%, respectivamente); el resto de los resultados para ambas medidas ofrecen valores de más del 60%. Es, de hecho, la medida de la exactitud con la que se da balance a estas medidas, y es la exactitud este proyecto la que presenta el rendimiento más alto para las tareas con un conjunto de datos en español alcanzando una calificación de 80.48%.

Así pues, el mejor modelo aquí presentado fue aquel en el que se usó el algoritmo Naïve Bayes Multinomial y que usó en la representación de características a los unigramas, bigramas, trigramas y tetragramas. En esta sección se expusieron los resultados de los experimentos por los que se llegó a esta conclusión, así como una comparación con los resultados de los antecedentes significativos. Hasta aquí, este ejercicio corresponde al primer objetivo específico de la tesis y a la primera fase de la misma. De esta tarea también se desprende la lista de elementos con ganancia de información (que puede consultarse en **¡Error! No se encuentra el origen de la referencia.**). La ganancia de información es una forma de medir la relevancia que tiene un atributo dentro de juego de datos; en otras palabras, se trata de ver qué tan probable es que un atributo (por ejemplo, el adjetivo

“ilegal”) forme parte de la clase de interés en un modelo de clasificación (en nuestro trabajo la clase de interés es mensaje xenofóbico). Por ello, esta lista inicia la fase siguiente, en la que se puede encontrar un análisis guiado por datos en otros dos corpus de redes sociales. Los datos que guían tal análisis son obtenidos, claro, de tal lista; en ese sentido, es usada a manera de la tradicional lista de palabras claves que se hace desde la lingüística de corpus, por lo que será necesario discutir qué tan oportuno es utilizar para ello una herramienta propia de la lingüística computacional (**¡Error! No se encuentra el origen de la referencia.**).

7.2 Descripción semántica de los elementos con ganancia de información obtenidos del análisis de sentimientos

Como ya se mencionó, lo que se verá a continuación son los resultados del análisis de corpus guiado por datos a partir de la lista con ganancia de información (**¡Error! No se encuentra el origen de la referencia.**), correspondiente a la segunda fase de la investigación (Ilustración 1). Entre los resultados del análisis de sentimientos se encuentra la lista de elementos con ganancia de información que se usará en esta sección a modo de la tradicional lista de palabras clave que, desde la lingüística de corpus, se utiliza para aproximarse al *aboutness* de un corpus, es decir, el tema del cual se trata un texto (Grabe y Phillips 1987). Es, por tanto, el punto de partida de un análisis guiado por datos que consta, en primer lugar, de la agrupación por campos semánticos de los primeros cien elementos de dicha lista; se incluye también la consideración de la dispersión de los mismos en dos corpus de redes sociales, que hasta ahora no se han visto en resultados, a saber, el corpus de comentarios a 12 vídeos de YouTube (Tabla 7), y el corpus de Twitter sin balancear (Tabla 4); se compara además con la dispersión dentro del Corpus de Referencia hecho a partir de textos periodísticos del SKE. Lo anterior se hizo con el propósito de ver si tales n-gramas son parte del discurso xenofóbico o más bien del discurso xenofóbico que sucede en Twitter. Cabe mencionar que estos ejercicios se realizan con el objetivo de investigar las relaciones entre técnicas propias de la lingüística computacional y la lingüística de corpus.

En un segundo momento, de entre todos los campos semánticos se seleccionó el de LUGAR para realizar un análisis cualitativo que consistió en la revisión de las concordancias de tres

n-gramas pertenecientes a tal campo. En el proceso de esta búsqueda, se encontraron tres asociaciones semánticas de las que se detallarán los patrones léxicos, gramaticales y semánticos que las conforman, cumpliendo con ello con el objetivo de identificar estos patrones en el contexto lingüístico inmediato de los rasgos prominentes ubicados por el modelo de clasificación automática, así como a las asociaciones semánticas que conformen. Como se verá en (**¡Error! No se encuentra el origen de la referencia. ¡Error! No se encuentra el origen de la referencia.**), el efecto de la activación de ciertos elementos léxicos en estos discursos da pie a hablar también de relaciones semánticas tanto entre las asociaciones encontradas como entre los subconjuntos semánticos que las conforman. Este hallazgo es un argumento a favor de la hipótesis principal de esta tesis, a saber, que el discurso xenofóbico que sucede en ambas redes sociales es una unidad, si bien compuesta por diferentes emisiones protagonizadas por distintos autores. Es posible hablar de la unidad precisamente por el efecto de activación propio de cada elemento léxico que hace posible usos compartidos, sedimentados y culturales.

7.2.1 Descripción de la Lista de elementos con ganancia de información

Esta sección es el inicio de los resultados del análisis de corpus, que, sin embargo, debe comenzar con el análisis de los elementos de la lista de los elementos con ganancia de información aunque esta sea uno de los resultados de los experimentos correspondientes a la sección anterior. La lista completa puede consultarse en Lista de elementos con ganancia de información (sección **¡Error! No se encuentra el origen de la referencia.**). Dado que el mejor modelo de clasificación usó una representación de características que consideraba n-gramas de cuatro longitudes, lo que se verá en el anexo serán estos n-gramas ordenados de acuerdo a su relevancia. En la

Tabla 11 se puede revisar, a manera de ejemplo, el primer n-grama de cada tamaño con su peso o probabilidad de ocurrencia en un tweet xenofóbico.

Tabla 11. N-gramas con ganancia de información por longitud y por peso probabilístico

id	Peso	Atributo	Longitud del n-grama
1	0.17014	nuestro país	2
2	0.16279	vienen	1
24	0.1438	elemento de la	3
43	0.1438	aquí en Tapachula Chiapas	4

La exploración de la lista de elementos con ganancia de información que se considera aquí consiste en la separación por campo semántico de los primeros cien elementos. Entre esos no se consideraron unigramas ni bigramas de categorías gramaticales funcionales, es decir, en el análisis solo se consideró sustantivos, verbos, adjetivos o sintagmas cuyo valor semántico pueda ser deducido sin necesidad de estar en su contexto de uso, de modo que estos n-gramas puedan ser incluidos en una clasificación por campo semántico sin revisarlo en sus concordancias.

Si los elementos con ganancia de información se usan con el propósito de conocer qué cadenas de palabras son representativas del discurso xenofóbico, esta agrupación por campos semánticos sirve para aproximarse a los tópicos de dicho discurso. Al respecto cabe resaltar que en esta clasificación manual por campos semánticos del discurso xenofóbico solamente se encontraron rasgos con un indicio de uso valorativo en uno de los campos aquí descritos. A grandes rasgos, parece indicar más bien campos semánticos referentes a fenómenos migratorios: POLÍTICA, LUGAR, COLECTIVIDAD y MOVIMIENTO.

Es importante recordar también que los corpus que aquí se revisan fueron conformados por las facilidades tecnológicas a las que accedimos, por lo que los hechos sociales específicos de los que hablan no son los mismos; sin embargo, tratan de eventos ocurridos en un periodo homologado por varias características, entre las que sobresalen el ingreso de

personas centroamericanas y caribeñas por medio de la frontera sur de México por caravanas.

A continuación, se presenta una tabla, por cada uno de los rubros, que muestra la distribución en ambas redes sociales y en un corpus de referencia, tanto en sus frecuencias brutas como las normalizadas sobre la base de un millón de palabras. Además, se ofrece una breve descripción que sirva de orientación al lector para comprender el contexto social que referencian los tweets del corpus no balanceado de Twitter.

a) ***POLÍTICA***

Naturalmente, los n-gramas que aparecen en la lista de elementos con ganancia de información contienen referentes a los sucesos migratorios del momento en que se emitieron las opiniones de los usuarios. Si bien es cierto que la planeación de la investigación y las posibilidades económicas y tecnológicas orillaron a que los datos que se recogieron para el análisis de sentimientos (esto es, los que se obtuvieron de Twitter) fueran datos a partir del abril del 2020. Es conveniente retroceder, por lo menos, un par de años para comprender el contexto político y social en el que esas opiniones fueron emitidas.

En México, las caravanas migrantes tomaron fuerza mediática –porque crecieron también en dimensión– a partir de octubre del 2018. El inicio de este año fue de campañas electorales que culminaron con las elecciones a inicios de julio y siguió con la toma de posesión por el mandatario electo –Andrés Manuel López Obrador– para diciembre de ese año. La campaña electoral que encabezó se caracterizó por un discurso contra los dos presidentes anteriores en materia de seguridad, en el que criticaba la militarización para combatir problemas de seguridad internos, que también se aplicó al control de los flujos migratorios (Correa-Cabrera, 2014; Ortega Ramírez et al., 2021). Esta postura –descrita en el documento Proyecto de Nación 2018-2024– consideraba la creación de la Guardia Nacional para promover el retiro paulatino de fuerzas armadas de tareas de seguridad pública; y proponía también un plan de amnistía con el objetivo de lograr la pacificación en México (Ortega Ramirez, 2021). Socialmente, esta posición fue tanto celebrada como criticada. Entre los detractores se encontraban las autoridades de las fuerzas armadas;

quien fuera Secretario de la Defensa Nacional en el sexenio del 2012 al 2018, el general Salvador Cienfuegos, por ejemplo, calificó la propuesta de amnistía como una muestra de populismo.

Sin embargo, tanto esta enemistad como la posición de López Obrador sobre las fuerzas armadas vieron un cambio drástico a partir de agosto del 2018, una vez ganadas las elecciones, después de una reunión del presidente electo con Cienfuegos y el entonces Secretario de Marina, Vidal Francisco Sanz. En noviembre de ese mismo año, legisladores de MORENA, el partido político que fundó y al que pertenece López Obrador, presentan una reforma constitucional para crear la Guardia Nacional. Finalmente dicha ley se promulga en marzo del 2019, los lineamientos de su constitución obedecen a este nuevo acercamiento con las fuerzas armadas, a pesar de las críticas de varios actores políticos (Semple & Villegas, 2019). Posteriormente, a raíz de la amenaza del entonces presidente de EE.UU., Donald Trump, de imponer aranceles como castigo por la migración irregular de México, o que pasa por México (BBC News Mundo, 2019), se confirma el despliegue de la Guardia Nacional a las fronteras Norte y Sur para controlar la migración el 7 de junio del 2019 (Pineda, 2019).

En este contexto se analizan los elementos con ganancia de información que se agruparon bajo este campo semántico denominado POLÍTICA y en el se encuentran n-gramas que referencian instituciones o medidas que se reconozcan en el contexto migratorio actual como restricciones al ingreso de migrantes.

Tabla 12. Distribución de rasgos del campo semántico "política" en dos redes sociales

N-gramas		Youtube		Twitter general no balanceado		Referencia SKE
		Frecuencia	**Frecuencia normalizada**	**Frecuencia**	**Frecuencia normalizada**	**Frecuencia**
POLÍTICA	**Leyes**	249	708.12	19	182.91	**101**
	a la guardia	51	145.04	14	134.78	**0**
	de la guardia nacional	22	62.56	15	144.41	**0**
	de la guardia	1	2.84	19	182.91	**0**
	Balas	23	65.41	5	48.14	0

En la Tabla 12 se puede consultar la distribución de dichos n-gramas en el corpus de Youtube, en el corpus de Twitter no balanceado y en el corpus de referencia. Tal distribución puede leerse tanto por las ocurrencias brutas como por la frecuencia normalizada. Se busca que las frecuencias normalizadas de los corpus de redes sociales sean mayores que las del corpus de referencia, tanto para poder probar la utilidad de la lista de elementos con ganancia de información como guía de un análisis de corpus, como para poder decir que estos elementos son propios del discurso sobre migración que ocurre en redes sociales –es decir, el *aboutness* de los corpus de Twitter y YouTube. Por ejemplo, si bien la mera frecuencia bruta de todos los elementos presentes en dicha tabla muestra que efectivamente son irrelevantes en el corpus de referencia, dada su nula aparición en la mayoría de los casos, es la frecuencia normalizada la que permite saber que el unigrama *leyes* es muy relevante en los dos corpus de redes sociales, pues ocurriría 98 veces por cada millón de palabras en el corpus de referencia, mientras que en el corpus de Twitter ocurriría el doble de veces y en el de Youtube casi el triple.

Es esperable que el unigrama *leyes* sea el único que está presente en el corpus de referencia dado que es, al final de cuentas, un sustantivo común que no se limita al contexto social que interesa. En los corpus de redes sociales, precisamente este n-grama se usa para aludir lineamientos jurídicos que restrinjan los flujos migratorios dentro del país (28) y (29).

(28) 😂😂😂 ahora no suenan taaan descabelladas las **leyes** migratorias de Estados Unidos (corpus Twitter no balanceado)

(29) Violar las **leyes** mexicanas, la soberanía, hoy se nombra “caravanas de migrantes” (corpus Twitter no balanceado)

Una búsqueda a la izquierda de este unigrama en el corpus de comentarios de YouTube advierte del uso de este sustantivo como algo que limita el comportamiento de los migrantes (Ilustración 11). Al respecto, cabe decir que dicho n-grama es acompañado en su contexto inmediato a la izquierda con verbos como *acatar, seguir, respetar* y *cumplir*, y también con sus opuestos *violar, pisotear, quebrantar*.

Ilustración 11.Concordancias para el sustantivo "leyes" ordenadas a la izquierda en el corpus de comentarios de YouTube

	File	Left Context	Hit	Right Context
1	video 5.txt	cientes para que queremos más Tiene que respetar las	leyes	We los Zetas se los van a comer vivos No mames es ob
2	video 6.txt	hazamos a nadie pero también tienen que respetar las	leyes	Monterrey helado..? Naa calor sentí cuando he ido Don
3	video 9.txt	s piensan que están en su país tienen que respetar las	leyes	y no estamos para estar manteniendo gente huevona C
4	video 10.txt	nta que si cruzaste deforma ilegal hay que respetar las	leyes	y respetar a la gente de ese país y sobretodo no exigir
5	video 2.txt	re en el pas k los regresen as no deben de respetar las	leyes	d los pases k agan eso en su pas no en los agenos Vier
6	video 5.txt	Tambien los centro americanos deben de respetar las	leyes	mexicanas pues los mexicanos son muy buenas persor
7	video 5.txt	la misma manera q han querido entrar sin respetar las	leyes	d otros países y cuando uno no tiene para uno pasa ha
8	video 10.txt	hos. "__ Vienen a este pais a la fuerza, sin respetar las	leyes	y ¿así quieren que no los traten cómo criminales? Que
9	video 1.txt	egal ganan unos y ganan otros enseñese a respetar las	leyes	Este paracito activista hablando de derechos humanos?
10	video 10.txt	on d otros países Soy latino pero debemos respetar las	leyes	d otros paises Ya dejen los pasar a usa No los quieren
11	video 3.txt	o!! No es xq sea racista sino q ellos deben respetar las	leyes	de nuestro paìs....vivaa Méxicoo!!" "Mientras les sigan c
12	video 9.txt	irosos donde están las imágenes deberían respetar las	leyes	q México ofrece. Medios vendidos. "Dónde está el víde
13	video 5.txt	 Una cosa es ser hermanos y otra el respetar las	leyes	VAN A DESASTIBILIZAR A NUESTRO PAIS Cuando uno de
14	video 10.txt	er desmanes y humillar al mexicano, al no respetar las	leyes	y haciendo lo que quiere y obvio si viven así es porque
15	video 5.txt	sabees Entran a la fuerza porque no saben respetar las	leyes	de Otros paises. Por eso los Estados Unidos ya no Quir
16	video 10.txt	duro para solventarte, pagar tus servicios, respetar las	leyes	y no evadir nada. Allá no puedes hacer y exigir como L

Entre tanto, en el corpus de referencia los contextos son más variados (Ilustración 12). En primer lugar, el tópico no es únicamente migratorio; en segundo lugar, las leyes no solo sirven para restringir sino para otorgar oportunidades o derechos.

Ilustración 12. Concordancias para el sustantivo "leyes" ordenadas a la izquierda en el corpus de Referencia

	File	Left Context	Hit	Right Context
1	rererenciasp.txt	menino en la region. Primera tesis: La aplicación de las	leyes	de cuotas para las mujeres en cargos de representació
2	rererenciasp.txt	Sólo sirve de preámbulo –de las constituciones, de las	leyes–,	pero luego se desvanece en el aire para dar paso a otro
3	rererenciasp.txt	ciones existentes entre el excepcional desarrollo de las	leyes	respecto a la igualdad de género, y la real desigualdad
4	rererenciasp.txt	cho y pugnó sin descanso por la reforma de dos de las	leyes	esenciales de nuestro ordenamiento jurídico: El Código
5	rererenciasp.txt	ción femenina en el mundo, describe los efectos de las	leyes	nacionales e internacionales sobre el estatuto de la mu
6	rererenciasp.txt	controlar este fenómeno con el endurecimiento de las	leyes	de inmigración y los requisitos para conseguir la reunif
7	rererenciasp.txt	s principios de legislación castellana contenidos en las	Leyes 54	a 61 de Toro, que consolidan la supremacía del hombr
8	rererenciasp.txt	oro analiza el tratamiento de la libertad religiosa en las	leyes	de educación y de libertad religiosa norteamericana, es
9	rererenciasp.txt	entes del matrimonio y de la familia, sobre todo en las	leyes	y la cultura dominante. 8. El matrimonio es un bien soc
10	rererenciasp.txt	n las diversas regiones del mundo. Se constata que las	leyes	no son suficientes para garantizar la igualdad de los de
11	rererenciasp.txt	«derecho al aborto» internacional sosteniendo que las	leyes	restrictivas obligan a las mujeres a ir en busca de práct
12	rererenciasp.txt	ne se publicó en 2011. La diferencia en el trato que las	leyes	e instituciones confieren a hombres y mujeres puede af
13	rererenciasp.txt	país aconfesional. Otra cosa es que la educación y las	leyes	sean terreno vedado a cualquier inspiración religiosa. E
14	rererenciasp.txt	ransito entre las leyes de cuotas para las mujeres y las	leyes	de la paridad de genero, pero que ni unas ni otras han
15	rererenciasp.txt	es de su propia casa, de sus propios afectos. Todas las	leyes	aquí acopiadas, a través de distintos instrumentos juríc
16	rererenciasp.txt	de que la investigación no está sometida más que a las	leyes	que ella se da a sí misma, y que no tiene otro límite qu

El unigrama *balas* parece entenderse como parte de ese discurso de restricción (30), como un recurso que se pide que se use contra migrantes:

(30) Deberían de darles armamento de **balas** para que los regresen (corpus Youtube)

El resto de elementos que encontramos en la Tabla 12. Distribución de rasgos del campo semántico "política" en dos redes sociales (p.106) contienen la referencia a la Guardia Nacional, cuyo contexto de creación y acción se explicó anteriormente. De manera semejante, como sucede con los unigramas *leyes* y *balas*, los usuario de redes sociales reconocen a esta institución que es utilizada para enfrentarse a los migrantes. Este hecho puede ser simplemente referenciado (31), criticado (32), o bien valorado, al grado de pedir dicha intervención (33) y (34).

(31) Agentes **de la Guardia Nacional** dispararon en contra de un automóvil en el que se trasladaban 13 personas migrantes en Chiapas. Hiriendo a dos de ellos y asesinando a uno. La vícitma mortal era de origen cubano (corpus Twitter no balanceado).

(32) Sugiero que en lugar de utilizar **a la Guardia *Nacional*** para reprimir y amedrentar migrantes, se utilice para protección y transporte para el personal médico en zonas de riesgo (corpus Twitter no balanceado).

(33) Estoy indignada con los vídeos de los migrantes pateando y apedreando **a la guardia** civil. No ve muchos comentaros en las redes. Ya no hay pretexto para dejarlos seguir en nuestro territorio. Los gringos los hubieran acribillado… (corpus Twitter no balanceado).

(34) Debemos usar la fuerza **de la GUARDIA NACIONAL**. NO MÁS INVASIÓN EN EL TERRITORIO NACIONAL (corpus Youtube).

b) ***LUGAR***

Dentro del campo semántico denominado LUGAR encontraremos ubicaciones que referencian distintos sucesos migratorios que ocurrieron en las fechas de extracción de los corpus. A decir de las frecuencias normalizadas, se observa que la mayoría de los n-gramas presentados en Tabla 13 son característicos del discurso de las redes sociales. Este es un requisito que se cumple para la mayoría de los n-gramas, salvo para *centro* que es un sustantivo que, de acuerdo con la RAE, cuenta con más 20 acepciones de uso, de modo que, puede suponerse que es un elemento altamente productivo en tanto puede producirse en diferentes contextos discursivos (ASALE & RAE, s. f.), lo que explicaría la alta frecuencia en un corpus que recoge varios temas, como el corpus de referencia.

Tabla 13. Distribución de rasgos del campo semántico "lugar" en dos redes sociales

N-gramas		Youtube		Twitter general no balanceado		Referencia SKE
		Frecuencia	**Frecuencia normalizada**	**Frecuencia**	**Frecuencia normalizada**	**Frecuencia**
	nuestro país	175	**497.67**	29	**259.93**	67

LUGAR	**de nuestro país**	36	102.38	5	48.14	14
	aquí en	145	412.36	21	202.17	5
	aquí en Tapachula	NA	0	6	57.76	0
	aquí en Tapachula Chiapas	NA	0	6	61.52	0
	en Tapachula Chiapas	NA	0	6	61.52	0
	río	22	62.56	21	202.17	13
	calle	54	153.57	19	182.91	7
	suelo	24	68.25	9	86.64	4
	instalaciones	1	2.84	8	77.02	5
	puente	11	31.28	6	57.76	8
	cerrar fronteras	NA	0	4	38.51	0
	a la frontera	29	82.47	11	105.9	0
	centro	159	452.17	30	346.57	492
	nación	18	51.19	21	202.17	36
	USA y	62	176.32	11	105.9	2

El caso contrario a *centro* son los tres n-gramas que contienen *Tapachula Chiapas.* Se trata de cadenas de palabras tan particulares que solo suceden en el corpus de Twitter, a pesar de que este municipio es ubicado como uno de los principales sitios receptores de población migrante en la frontera sur de México (Monitoreo de flujos migratorios en Tapachula y Tenosique, 2022). Ahora bien, este corpus fue formado por semillas entre las que destacan tanto *Tapachula* como *Chiapas*. Estos sustantivos fueron escogidos como semillas, como ya se dijo anteriormente (**¡Error! No se encuentra el origen de la referencia. ¡Error! No se encuentra el origen de la referencia.**), porque una exploración de notas periodísticas destacó a este municipio como un lugar que, en el discurso mediático se vincula con la

recepción de migrantes. No obstante, a propósito de las formas de obtención de los datos, debe recordarse que el corpus de YouTube está hecho de comentarios a doce vídeos, entre estos, ocho hablan del ingreso de migrantes por la frontera sur de México, como aluden sus títulos (Tabla 7). Una búsqueda del término clave *apachula, dio como resultado solamente cinco concordancias (Ilustración 13. Concordancias para el término clave *apachula en el corpus de YouTube), ninguno de ellos tiene que ver con las cadenas de palabras de las que se hablan; entre tanto, una búsqueda de **hiapas* se muestra arrojó 70 concordancias, pero solamente (35) y (36) coinciden con uno de *aquí en* que es n-gramas. De ello se concluyó que se debe a la particularidad de estas cadenas de palabras, y no a la referencia a estos lugares.

	File	Left Context	Hit	Right Context
1	video 1.txt	y las noticias que sabemos se aqui y haya en	tapachula	es que aumento las enfermedades de transmission s
2	video 10.txt	ay testigos q dicen q un carro lo mató en	Tapachula	Mexico pero nadie más nos da razón de el
3	video 5.txt	para todos. Espero y ruego no se queden en	Tapachula	Ello saben k en USA no los ban a
4	video 10.txt	exico. Specially in the border cities like Tijuana and	Tapachula	Chiapas." "Yes, this is what Honduras suffer every da
5	video 10.txt	en chiapas donde ellos pasaron en tecun y despues	tapachula	fue muy fuerte y largo camino para que detengan

*Ilustración 13. Concordancias para el término clave *apachula en el corpus de YouTube*

(35) **aquí en chiapas** donde ellos pasaron tecun

(36) la verdad no saben como se sufre **aquí en chiapas** la pobresa ipocrita gobierno

Ahora bien, lo que se observan en la Tabla 13 son referencias a ubicaciones concretas como demarcaciones geográficas (*río*), o geopolíticas (*Tapachula Chiapas; USA*), que aluden a situaciones migratorias en las fronteras norte (37) y (42) y sur de México (38), (39), (40) y (43) . Son, a su vez, lugares emblemáticos por referenciar la forma de cruzar de Centroamérica a México ya sea que los mencionen explícitamente (38) o no (39) y (40); o bien, de cruzar de México a EE.UU. (37) y (42); o bien por ser percibidos como lugares de alta concentración de poblaciones migrantes (37), (41) y (43).

(37) Pues el Muro ya se lo construyó el Coronavirus, ¿Se acuerdan de las caravanas de inmigrantes? Ni sus luces de ellos. En Cd Juárez ya no se ven correteando para cruzar el *Río Bravo* (Twitter).

(38) Informan que alrededor de 500 elementos de migración son desplegados en la rivera del río Suchiate para frenar caravana migrante de hondureños prevista que llegue en las próximas horas (Twitter).

(39) Echénle huachicol al río y préndanle fuego (YouTube, comentario a video 3, Tabla 7. Composición del corpus de YouTube).

(40) EL EJÉRCITO Y LA GUARDIA NACIONAL DEBERÍAN HACER MÁS PROFUNDO EL RÍO PARA QUE AÚN EN TEMPORADA DE SEQUÍA TENGA BASTANTE AGUA Y NO PUEDAN CRUZAR TAN FÁCILMENTE (YouTube, comentario a video 1, Tabla 7. Composición del corpus de YouTube).

(41) De una vez denles pasaporte para trabajar porque ya se quedaron en mexico pura madre los dejan pasar a usa y no queremos inutiles ya bastante tenemos con políticos (YouTube, comentario a video 5, Tabla 7. Composición del corpus de YouTube).

(42) Ustedes con las balas y nosotros acá en la frontera sin poder cruzar a USA porque los inmigrantes amenazaron con cruzarse a USA y que nos cierran el puente a las 12:15 am son las 4:29 am y aquí seguimos 👰🤬😡a donde vamos a parar #durmiendoenelcoch (Twitter).

(43) En la décima norte aquí en Tapachula, Chiapas, se encuentra una cantidad considerable de migrantes que no aplican el #QuedateEnCasa ellos también deberían obedecer las instrucciones del gobierno (Twitter).

Las ubicaciones geopolíticas aparecen también de manera abstracta por la carga léxica (*nación*) (44) y (45); o por la conjunción de elementos que dan cuenta de estrategias de subjetivación, a saber, adverbios con valores deícticos (*aquí en*) (46) y (47), o por el uso de

pronombres personales que vinculan al hablante con dicha demarcación geopolítica pero también con una colectividad (*nuestro país*) (48) y (49).

(44) Ejemplo, la fuga de cerebros nos afecta como **nación**, el gobierno morenista no genera empleos (Twitter)

(45) Pero porque prefieren ir a morir a otra **nación** y no luchar muriendo combatiendo asu mal gobierno… (YouTube, comentario a video 5…….)

(46) La gente es masa manipulable, **aquí en** México le tiran mucho por las leyes migratorias que implementó cuando entraron los hondureños a la fuerza a México (Twitter)

(47) Con la pena pero **aquí en** México no se pueden quedar (YouTube, comentario a video 4)

(48) Preguntémosle cómo frenará a inmigrantes centroamericanos que de menos un 30% vienen infectados listos para entrar **a nuestro país** a fin de mes!!! (Twitter)

(49) …que tengan respeto ellos nosotros estamos en **nuestro país** y no andamos haciendo males en otros países (YouTube, comentario a video 9).

Como ya se mencionó fue de este campo semántico del que se tomaron los n-gramas para el análisis de concordancias. Algunos de los n-gramas que suceden en la Tabla 13, coinciden con las semillas con las que se buscaron los tweets para conformar el corpus, por lo que se evitó seleccionar uno de ellos para el análisis en concordancias, con el propósito de evitar caer en un razonamiento circular. De modo que los elementos seleccionados son *aquí en, nuestro país* y *de nuestro país*. Estos sintagmas cuentan con el requisito que los marcan, de acuerdo con nuestros propios lineamientos, como propios de discurso xenofóbico que ocurre en redes sociales. A saber, la frecuencia normalizada que presentan es mayor que la del corpus de referencia, además tanto para las cadenas *nuestro país*, como para *aquí en*, se trata de las frecuencias normalizadas más altas para ambas redes sociales. Los detalles de este análisis no se revisarán aquí sino en la suiguiente sección **¡Error! No se encuentra el**

origen de la referencia. Análisis de Asociaciones Semánticas en el Campo Semántico de (p. **¡Error! Marcador no definido.**).

c) ***Colectividad***

Este campo semántico está compuesto por n-gramas que referencian a diferentes colectivos. La mayoría de ellos son actores sociales identificables en el fenómeno de migración que aquí ocupa.

En cuanto a las tendencias en las frecuencias normalizadas comparadas, entre las de las redes sociales y el corpus de referencia, se observa un fuerte contraste. En primer lugar, porque más de la mitad de estas cadenas de palabras no ocurren en el corpus de SketchEngine. Sucede, por otro lado, que el corpus de YouTube tiene la mayor frecuencia normalizada entre los primeros cien elementos con ganancia de información; se trata del unigrama *nosotros*, que sucede ocho veces más que en Twitter y 28 veces más que en el corpus de referencia. Es importante hacer notar que es un elemento con propiedades deícticas que comparte con los elementos más frecuentes del campo semántico anterior de LUGAR: la activación de la primera persona plural.

Tabla 13. Distribución de rasgos del campo semántico "colectividad" en dos redes sociales

N-gramas		Youtube		Twitter general no balanceado		Referencia SKE
		Frecuencia	**Frecuencia normalizada**	**Frecuencia**	**Frecuencia normalizada**	**Frecuencia**
COLECTIVIDAD	**los migrantes porque**	2	5.69	6	57.76	0
	son migrantes	9	25.59	18	173.29	0
	migrantes son	14	39.81	19	182.91	0
	mexicanos y	99	281.54	17	163.66	0
	los mexicanos y	17	65.41	9	48.14	0
	y nosotros	23	65.41	5	48.14	2
	nosotros	571	**1623.83**	19	182.91	59
	a inmigrantes	5	14.22	14	134.78	0
	inmigrantes y	19	54.03	18	173.29	11

	nadie	380	**1080.66**	39	375.45	43
	negros	6	17.06	9	86.64	5
	hijos de	30	85.32	13	125.15	69
	elemento de la	0	0	6	57.76	0
	un elemento de la	0	0	5	48.14	0
	un elemento	0	0	5	48.14	16
	un elemento de	0	0	5	48.14	2
	de migrantes cubanos	0	0	4	38.51	0
	cubanos	21	59.72	14	134.78	0
	su gente	58	164.94	10	96.27	0
	migrantes porque	5	14.22	7	67.39	0

Dado que el discurso que se estudia es sobre la migración centroamericana y caribeña que ingresa a México, y debido también a que se procuró con la extracción de los datos que ese discurso fuera principalmente emitido por mexicanos, es esperable encontrar entre estos colectivos tanto a los mexicanos como centroamericanos y caribeños y sus respectivas nacionalidades. Esta expectativa se cumple para el caso de los mexicanos, quienes se encuentran explícitamente en dos n-gramas–; también es esperable que ese *nosotros* que mencionábamos conceptualice precisamente a los mismos mexicanos. Al respecto, una búsqueda en el corpus de YouTube a través de AntConc de este pronombre en su contexto lingüístico inmediato, específicamente cuando este pronombre es seguido del artículo definido plural en su forma masculina (*nosotros los*) que escogió con el propósito de encontrar quiénes componían esa colectividad, arrojó 132 concordancias, 79 de las cuales se refiere a mexicanos (Ilustración 14).

	File	Left Context	Hit	Right Context
1	video 6.txt	uestro hermoso mexico mal agredecidos Por eso	nosotros los	mexicanos no crecemos por mentalidad para hur
2	video 9.txt	lo que le garantizo es que de la frontera con	nosotros los	mexicanos no van a pasar _____ @Neily Ram
3	video 10.txt	el Camino Los migrantes no se merecen eso que	nosotros los	mexicanos no los apoyemos y los tratemos así
4	video 10.txt	y los que van a salir afectados vamos a ser	nosotros los	mexicanos no los hondureños" Honduras __ __ aj
5	video 10.txt	n q te hayas vendido victor nadamas te digo que	nosotros los	mexicanos no hacemos caravanas ni llegamos a u
6	video 10.txt	el que debe informarse y aprender a leer eres tú	Nosotros los	mexicanos no atacamos a nadien simplemente d
7	video 5.txt	ren hacer ver cómo xenófobos Valeria Valencia a	nosotros los	mexicanos nos tienen entre la espada y la
8	video 6.txt	-x-kFncrJQca6oxA 0 Exelente comentario amigo	nosotros los	mexicanos nos quejamos de los emigrantes en M
9	video 10.txt	lar a la gente hondureña pero como hacerlo si ni	nosotros los	mexicanos nos alludamos . Like si quieres mater
10	video 10.txt	plazados x que los olvidan también sufren x que	nosotros los	mexicanos nos tenemos que preocupar x otra ge
11	video 5.txt	venir a nuestro país a exigirle q los mantenga y	nosotros los	mexicanos tenemos que aguantarlos y lo único q
12	video 5.txt	wQwigUWW-uQ" que se regresen a su país	nosotros los	mexicanos tenemos muchos problemas y hora vc
13	video 5.txt	pasar para ir a estados unidos. @Miller tambien	nosotros los	mexicanos tenemos alma y no le hariamos daño

Ilustración 14. Concordancias para "nosotros los"

El n-grama *(un) elemento de la*, aunque con una baja ocurrencia bruta, es representativo de la red social Twitter, pues no sucede ni en Youtube ni en el corpus de referencia. Junto a *mexicanos,* se trata de la enunciación de actores que se oponen a quienes migran, dado que se trata de la parte constitutiva de una institución encargada de la defensa en México que ha sido utilizada para impedir el ingreso de las caravanas migrantes (Ortega Ramírez et al., 2021; Pineda, 2019; Semple & Villegas, 2019), de la que ya se explicó su contexto en el campo semántico POLÍTICA.

Frente a esto, es decir, en cuanto a los actores sociales que no son mexicanos, sino que ingresan a México, sobresale que, al menos en los primeros cien n-gramas, solo destaca la nacionalidad cubana entre todos los países de los que proviene la migración; así el gentigilicio *cubanos* ocurre en dos n-gramas y es uno de los tres sustantivos que destacan para hablar de tales actores. Dos de estos denominan independiente del contexto social, a quienes llevan a cabo el desplazamiento, esto es, *migrantes* e *inmigrantes*, que aparecen en cuatro y dos n-gramas respectivamente. Además *migrantes* y *cubanos* aparecen juntos en un trigrama que no es relevante para el caso de la red social YouTube. Entre tanto, el resto de cadenas de palabras que incluyen los sustantivos *migrantes* e *inmigrantes* parece que son usados para hablar de un colectivo en el que no es importante mencionar la nacionalidad.

Al menos, eso indica una rápida revisión de las concordancias de los n-gramas que contienen esos sustantivos en el corpus no balanceado de Twitter, entre las que sólo tres n-grama que

muestran un interés explícito por parte de los hablantes por especificar el origen de las personas que migran (50) a (52).

---

(50) Preguntémosle cómo frenara **a inmigrantes *centroamericanos*** que de menos un 30% vienen infectados…

(51) Agentes ministeriales aseguran a 100 **inmigrantes *de origen Hondureño*** *y Guatemalteco*…

(52) **564 migrantes son de Guatemala, 39 de Honduras, 20 de El Salvador, 28 de Nicaragua y uno de Belice**.

El resto de los n-gramas contienen sustantivos que necesitan un análisis más amplio para conocer a qué actores se refiere. Las mismas búsquedas que se hicieron para los n-gramas que sí se describen, se hicieron para *nadie, negros, hijos de,* y *su gente*, pero la variación de sus referentes no era tan determinante, y detenerse en este punto rebasa los objetivos de esta tesis.

d) ***MOVIMIENTO***

Si bien en los campos semánticos anteriores pueden mencionarse referentes a las particularidades de los acontecimientos migratorios que a esta tesis ocupan, MOVIMIENTO es un conjunto semántico que trata de las generalidades del hecho de migrar, a saber, la descripción de los desplazamientos que conforman dicho fenómeno.

En cuanto a la distribución de estos rasgos en los corpus que se revisaron (Tabla 14), se confirma nuevamente, por sus frecuencias normalizadas, que son de relevancia en los discursos de las redes sociales; de hecho, se trata de elementos prácticamente exclusivos. Dentro de esos rasgos, el que más veces sucede en el corpus de referencia, *vienen*, ocurriría casi 15 veces por cada millón de palabras, entre tanto, ocurriría casi 22 veces más en el corpus de Twitter, y 75 veces más en el de YouTube.

Tabla 14. Distribución de rasgos del campo semántico "movimiento" en dos redes sociales

MOIMIENTO	N-gramas	Youtube		Twitter general no balanceado		Referenciaske
		Frecuencia	Frecuencia normalizada	Frecuencia	Frecuencia normalizada	Frecuencia
	vienen	384	**1092.88**	33	317.69	15
	entrar a	208	591.97	18	173.29	2
	llegaron a	10	28.46	7	67.39	2
	se larguen	9	25.59	5	48.14	0
	se queden	71	201.91	1	9.63	0

Ahora bien, los rasgos presentes en la Tabla 14, se diferencian de lo que sucede en los otros campos semánticos en dos aspectos. El primero de ellos, como ya decíamos, se refiere a que las cadenas de palabras de este campo cuentan con una carga léxica que describe la generalidad de los fenómenos migratorios. No obstante, a diferencia de los otros, solo estos elementos contienen rasgos que apuntan a un discurso valorativo, a pesar de que estamos describiendo el discurso xenofóbico en las redes sociales. Específicamente, todos están conjugados en tercera persona del plural y esta persona se enfatiza en un par de ellos al estar antecedidos por el pronombre dativo reflexivo "se", también en tercera persona del plural. Una muestra de ello lo representa el bigrama *se larguen*, pues este verbo, coocurriendo con el reflexivo, en un registro informal y en una situación comunicativa de tensión entre interlocutores, es usado por uno para ordenar abandonar un lugar a otro. Una muestra de las concordancias de este elemento para ambas redes sociales indica que es precisamente la situación que acabamos de describir, la que está ocurriendo (Ilustración 15 **e** Ilustración 16).

	File	Left Context	Hit	Right Context
1	Tw_ge...	MIERDAS de México Hay que sacarlos a patadas de México Que	se larguen	a su país bola de MIERDAS" Son delincuentes,en la
2	Tw_ge...	🤐 A los migrantes deben masacrarlos con plomo grueso ,que	se larguen	a su país bola de MIERDAS @laoctavadigital @GN_
3	Tw_ge...	reña en nuestro país de ilegal Visa humanitaria.... Por Dios que	se larguen	a sus países todos los centro americanos si no les
4	Tw_ge...	ad, son apátridas y se sienten de la realeza. Pobres diablos que	se larguen	del país. Gracias a los jueces en EEUU que tienen
5	Tw_ge...	FADORA y DEMAGOGA....tienen 22 AÑOS de FRACASOS... URGE	SE LARGUEN.🤬	https://t.co/H0BM3wDgwF" #editorial Este es el p

Ilustración 15. Concorancias para el bigrama "se largen" en el corpus de Twiter no balanceado

	File	Left Context	Hit	Right Context
1	video 10.txt	No dan pena ve a venezuela Like si quieres que	se larguen	de México Ps:Estos de badab
2	video 10.txt	n la culpa "Pues esperemos que ya todos tus vecinos	se larguen	de nuestro país ya tenemos bastante delincuencia, a
3	video 1.txt	aja eres pariente de Trump? __ Eso estaría mejor que	se larguen	Los Mexicanos No los queremos y Estados Unidos me
4	video 5.txt	ndo de anbre nobinieran mujeres como marrana Que	se larguen	a usar violencia a su pais LLÉVENSELOS PARA SU CAS
5	video 1.txt	as pares en conejas en México no los queremos que	se larguen	los trabajos que hay en México.son de los mexicanos
6	video 5.txt	__ Si que regresen a sus países Blanca Romero a que	se larguen	mo los queremos aquí @ximena bañuelos lo bueno q
7	video 3.txt	porke su intencion es no kedarse en mexico pues ke	se larguen	por cuba y luego de ahi en balsas a miami e.u. y &qu
8	video 11.txt	n que todo mundo los va a sectar.que.pendjos AQUE	SE LARGUEN	QUITAN TRABAJO A MEXICANOS. LOS EMPLEAN PORQ
9	video 5.txt	Honduras no los tiene así No Lo queremos aquí que	se larguen	Una niña de un añito Por que los periodistas y canale

Ilustración 16. Concordancias para el bigrama "se larguen" en el Corpus de YouTube

Hasta aquí se ha descrito someramente cada uno de los campos semánticos en los que se clasificaron los primeros elementos con ganancia de información. Como ya se habrá advertido, estos campos semánticos podrían estar relacionados unos con otros, de hecho, a esto obedece la descripción tan parca del campo MOVIMIENTO, los verbos de la Tabla 14, así como otros verbos de movimiento co-ocurren frecuentemente con el campo semántico de LUGAR, que es el campo en el que abundaremos en descripciones. Estas últimas se pueden revisar a continuación.

7.2.2 Análisis de Asociaciones Semánticas en el Campo Semántico de LUGAR

En esta sección se exploran los rasgos del campo semántico LUGAR en las redes sociales YouTube y Twitter en su contexto lingüístico inmediato, con la mira a responder dos objetivos de este proyecto, los cuales son 1) identificar los patrones semánticos y gramaticales en los que ocurren los rasgos lingüísticos ubicados por el modelo de clasificación automática, así como, 2) identificar las asociaciones semánticas y las anidaciones en las que participen dichos rasgos en su contexto lingüístico inmediato. Para ello se hablará de los resultados de la revisión de los n-gramas con propiedades deícticas *aquí en, nuestro país* y *de nuestro país* en sus líneas de concordancias.

Cabe señalar que, al ser obtenidos de un campo semántico específico que refiere a lugares, es esperable que estos n-gramas actúen como un set semántico que agrupe a lugares relevantes de los sucesos migratorios que a los hablantes del corpus le interesen. Además de este hecho, en el análisis se encontraron tres asociaciones en que la asociación LUGAR podía aparecer anidada, a saber, MOVIMIENTO, PELIGRO y ACTITUD. La primera de estas, como se ve, coincide con el último de los campos semánticos encontrados entre los elementos con ganancia de información.

(53) MOVIMIENTO → LUGAR

(54) PELIGRO → LUGAR

(55) ACTITUD → LUGAR

En la Tabla 15 se puede observar la distribución de tales asociaciones tanto por su prosodia semántica, como por las redes sociales en las que ocurre. Cada una de estas asociaciones está compuesta por subconjuntos semánticos ubicados en la segunda columna, cuya descripción será dada más adelante.

Tabla 15. Distribución de asociaciones y prosodia semántica por red social para los n-gramas "aquí en", y "[en] nuestro país"

		Twitter		**YouTube**	
Asocación semántica		**Prosodia no negativa**	**Prosodia negativa**	**Prosodia no negativa**	**Prosodia negativa**
Movimiento	Entrada	2	5	1	13
	Estadía	5	7	17	36
	Salida	0	0	2	1
	Retorno	0	3	0	10
Peligro	Amenaza	0	2	0	41
	Protección	1	5	1	17

Actitud	Problemática	5	7	17	10
	Asistencia	3	1	7	38
Total general		**17**	**29**	**48**	**164**

En esta tabla no se consideró relevante presentar las frecuencias normalizadas de las asociaciones, en tanto no se presenta la totalidad de las ocurrencias de las mismas dentro de los corpus sino solo aquellas que suceden con los n-gramas escogidos para el análisis. A propósito de la distribución de la polarización de la prosodia semántica, cabe recordar que la ganancia de información de los elementos se refiere al peso probabilístico que tiene cada uno de ellos de ser parte de una expresión xenofóbica, en ese sentido se observa que, de hecho, aparecen mayormente asociados a la prosodia negativa. Con las categorías, las prosodias positiva y neutra (estas dos se mencionan indistintamente como prosodia no negativa) y negativa de cada uno de los contextos revisados, apuntan a la valoración de los usuarios de redes sociales hacia la migración.

Lo que se verá a continuación será la descripción de cada una de las tres asociaciones con sus respectivos subconjuntos.

7.2.2.1 MOVIMIENTO

La primera de estas asociaciones, MOVIMIENTO, describe los tipos de –valga la redundancia– movimientos que conforman las migraciones humanas. De acuerdo con la metodología seguida, esta descripción obedece a la relación semántica que conforman los n-gramas seleccionados que referencian lugares con el verbo de la cláusula dentro de la que aparecen. Es esperable que, por ello, sea la asociación que más fuertemente restrinja un set semántico de lugares en un corpus temático sobre migración, como de hecho sucede, al ser la más frecuente con 102 ocurrencias.

Compuesta a su vez por cuatro subconjuntos, el primero de ellos describe la *entrada* a un país destino o de paso (56) y (57); *estadía* se refiere a la permanencia en un país que no sea el propio (58) y (59); *salida* hace alusión al hecho de dejar el país de origen (61); mientras que *retorno* es un subconjunto en que se agruparon las descripciones del regreso de una persona, o un grupo de personas, al país propio (61).

(56) Que bueno! Así no **entraran** más sus migrantes a **nuestro pais** (entrada-prosodia negativa)

(57) …cobija a los hermanos inmigrantes q **llegan a refugiarse** a **nuestro país** (entrada-prosodia positiva)

(58) Dicelo a los 70 mil mantenidos que **siguen** en **nuestro pais.** No se quieren ir (estadía-prosodia negativa)

(59) …de la caravana se superen y **logren** *sus sueños* **aquí en *Tijuana (Méxic)*** o en EUA (estadía-prosodia positiva)

(60) pero nosotros también sufrimos tantas muertes adiario y no **dejamos nuestro país** (salida- prosodia negativa)

(61) A balazos hay que **sacar** a estos infelices **de nuestro país**. (retorno-prosodia negativa)

Tabla 16. Sets semánticos que conforman la asociación MOVIMIENTO

	Entrada	**Estadía**	**Salida**	**Retorno**
Sujeto: quien migra +agente +movimiento	Entrar, llegar, pasar, refugiarse, venir,	Atravesar, estar, hacer, lograr (sueños), nacer pasar (por), permanecer, poder + infinitivo, quedar, querer + infinitivo; seguir; tener + inf, vivir	Dejar, inmigrar, salir	largar, regresar, salir

Sujeto: quien migra *-mov* *-agencia*	Ser bienvenido, caber	Encontrarse, haber, ir para, merecer		
Sujeto: desde el país receptor *+ agencia* +movimiento	Permitir, dejar entrar, recibir			expulsar, sacar
Sujeto: desde el país receptor *+ agencia* *-movimiento*		Aceptar, querer, tener + aguantar, tratar, tener esperando		

Revisado, entonces, el contexto lingüístico inmediato de los n-gramas se obtuvieron los cuatro sets semánticos descritos anteriormente, en la tabla Tabla 16 pueden observarse las realizaciones de las frases verbales que se encontraron. En la tabla se considera, de acuerdo con lo que sugieren los datos, que el sujeto gramatical de tales acciones puede a) tener el rasgo de la agentividad b) coincidir con las personas que migran o bien, con las personas del país receptor. La tabla también da cuenta de verbos que son, de hecho, de movimiento; hay otros, sin embargo, que apuntan a la aceptación u obstrucción de tales movimientos.

Tanto lo que sucede en la **¡Error! No se encuentra el origen de la referencia.**, como los ejemplos (56) a (61), dan cuenta de que referenciar hechos migratorios no era la intención de los hablantes de nuestro corpus, sino valorarlos. Tal valoración tiende, como ya se mencionó, a ser una opinión negativa sobre la migración. No obstante, tanto los casos de prosodia negativa, como positiva, usan los rasgos que se encuentran descritos anteriormente.

Se observa que las tendencias de ocurrencia del subconjunto *salida*, en la Tabla 16. Sets semánticos que conforman la asociación MOVIMIENTO, tienen un par de comportamientos de activación distintos a casi todos los subconjuntos: a saber, es, por mucho, el menos frecuente; y tiene, junto con *problemática* (Tabla 18. Set semánticos que conforman la asociación ACTITUD, p. 134), más ejemplos no xenofóbicos que xenofóbicos. De modo que se puede decir que el tema de la *salida* de una persona de su lugar de origen está restringido

de los n-gramas de interés; aunque las pocas veces que sucedió, coincidió que el sujeto gramatical referencia a la persona que migra utilizando verbos con el rasgo de agentividad (inmigrar, salir, dejar).

Ahora bien, el ejemplo de prosodia negativa de *salida* se puede leer en (60), donde se advierte una comparación entre lo que, según se dice, sucede en el país de origen de los migrantes y lo que sucede en el país receptor (punto deíctico desde donde se conceptualiza el hablante); dicha comparación se corona con aquello que se hace ante ello en el lugar receptor, "pero no dejamos nuestro país", evaluando con ello como negativo al acto de migrar.

Entre tanto, los ejemplos con prosodia semántica positiva hacia quienes migran se observan un par de estrategias donde el hablante equipara consigo mismo a las personas migrantes. En (62), ello ocurre gracias a que el hablante directamente se considera en el sujeto gramatical del verbo *inmigrar*; en tanto, en (63), se referencia a quienes salen de un país como *hermanos,* además *nuestros.*

(62) ESO PORTENSE BIEN Y A TRABAJAR QUE A **ESO IMIGRAMOS DE NUESTRO PAÍS** VIVA MÉXICO VIVA AMERICA LATINA (salida-prosodia positiva)

(63) NUESTRO PAÍS SALDRÁ ADELANTE SIN QUE *NUESTROS HERMANOS HONDUREÑOS* **SALGAN DE NUESTRO PAÍS** (salida-prosodia positiva)

Existen también ejemplos de prosodia positiva en los subconjuntos de *entrada* y *estadía*, sin embargo, las estrategias anteriores también ocurren una vez cada una. Ambas suceden dentro del subconjunto de *entrada* como en (64) y (65). En el primero de estos ejemplos sobresale que la migración referenciada no es una internacional como la que nos ha ocupado hasta el momento y más: está hablando sobre el mismo punto deíctico desde el que enuncia (*aquí en Tijuana somos*). En (66) también sucede, aunque no en el contexto lingüístico más inmediato, que el hablante cuenta a los migrantes dentro de una categoría a la que él mismo pertenece, a saber, *seres humanos*.

(64) @MENCIÓN Tu así nos quieres ver, **aquí en *Tijuana somos tierra de emigrantes***, mi familia llegó de Jalisco me siento orgullosa de ser de México. (entrada-prosodia positiva)

(65) Amigos denle amor a la págiana de @MENCIÓN, son una fundación q cobija a ***los hermanos inmigrantes q*** **llegan a refugiarse a nuestro país**. en Zapopan (entrada-prosodia positiva)

(66) Si por mí fuera ayudaría a todos los niños y niñas hondureñ@s, pero tan solo tengo 12 años y no tengo dinero, *y me duele y desepciona como los mexicanos tratan a los demás seres humanos,* solo le pido a Dios que los mexicanos recapaciten y se ponga a pensar que lo que están haciendo esta mal por qué es como discrminar, y lamentablemente lo único que puedo hacer es orarle a dios para que todos los de la caravana se superen **y logren *sus sueños* sea aquí *en Tijuana*** (Méxic) o en EUA (estadía- prosodia positiva)

En otro tipo de patrón sintáctico se encuentran los sujetos gramaticales referenciando a personas no migrantes como agente de dichos movimientos. Al igual que con las estrategias anteriores, reportar tales acciones no es meramente una función referencial, sino mostrar una posición ideológica que, en este caso, es casi exclusiva de la exposición de la disconformidad. En ese sentido, hay realizaciones verbales entre los subconjuntos de *entrada* y *estadía* que dan cuenta de una dinámica de fuerzas a la que se oponen las personas migrantes si quieren entrar o permanecer en el punto deíctico desde donde el usuario habla, se trata de los verbos *aceptar (67)* y (68)*, permitir* (69), así como *dejar entrar (70)*. En estos ejemplos, la referencia al punto deíctico se marca con los n-gramas "(de) nuestro país" y "aquí en"; la dinámica de fuerzas si bien se establece a través de los verbos también hay otras evidencias. En (67) sucede un rechazo de las personas migrantes pese a las expectativas de hablante, contrario a lo que sucede en (69) y (70) donde lo que va en contra de tales expectativas es la entrada o estadía de los migrantes.

(67) *Siempre he pensado que cuando sea grande voy a proteger a todas esas personas migrantes* porque en mi pais colombia hay muchos migrantes de venezuela que vienen a buscar una oportunidad y *anunque* en **nuestro pais** no todos ***los*** **aceptan**

(68) Honduras y estados unidos hay personas amaberes onestas **aquí en *México*** **aseptamos *a cualquier país***

(69) …era una locura **permitir** la migración ilegal a través **de nuestro país**. Bien por ambos.

(70) Nunca renunciaremos al derecho de tomar una decisión con quien nos gustaría vivir juntos en nuestro país… a quienes nos gustaría **dejar entrar** *en el territorio de* **nuestro país** violar una serie de fronteras, no es un derecho humano… Mi pregunta es cómo ser un refugiado y violar las fronteras de cinco o seis países seguros

En (67) y (68), el contexto permite asignarle una prosodia positiva a la anidación de las asociaciones MOVIMIENTO – LUGAR. En uno de tales ejemplos, (67), el hablante está referenciando la postura negativa de algunos de sus conciudadanos, ante la que expresa una actitud crítica. En otras palabras, lo que sucede aquí es un ejemplo de lo que Hoey llamó *priming pasivo*, por lo que el hablante reconoce la postura ideológica contraria a la suya. Mientras que en (68), el hablante se cuenta en el sujeto gramatical al tiempo que evoca un punto deíctico que habita (que inmediatamente codificada como *México*), el hablante reconoce un grupo de personas (honestas y de cualquier *otro* país) ajenas a dicho punto deíctico a las que él –y sus conciudadanos– *aceptan*. Estos dos ejemplos, aunque favorables con la población migrante, dan cuenta de un conocimiento cultural compartido, que sitúa a las personas que no son miembros de un país en medio de un proceso que requiere una *aceptación* más allá de instituciones oficiales, pero igualmente *nacionales.*

Ambos aspectos son, de hecho, recurrentes en la valoración negativa ante la migración en los subconjuntos referidos. De hecho, la noción de que los movimientos migratorios están condicionados, no se limita al uso de tales verbos. Una segunda estrategia que se observa en este mismo sentido es enlistar calificaciones morales, o actividades con tal carga moral

que separa a personas que merecen continuar en el flujo migratorio de las que no. Por ejemplo, en (62) se trata de "portarse bien" y trabajar; en (68) los merecedores son personas amables y honestas.

Retorno es una asociación en la que, al menos en los datos seleccionados para el análisis, se muestra exclusivamente una prosodia negativa. Coincide con *salida* en la descripción de salir de un punto deíctico, pero en esta ocasión se trata de salir del país receptor; en ambos subconjuntos no es necesariamente relevante el punto al que se llegue siempre y cuando se salga de uno. A semejanza de los subconjuntos *entrada* y *estadía*, entre sus rasgos semánticos se encuentra el hecho de que el sujeto es agente y su acción se vincula a un punto deíctico, pero, esta vez, no será necesariamente el hablante el agente, aunque sí será este último alguien que habite el mismo punto deíctico; además más que una dinámica de fuerza donde la población migrante puede ser recibida o no, se expresa una oposición tajante: la acción de los agentes será sacar a los migrantes del territorio que comparte con el hablante, como se muestra en Tabla 16, con el uso de verbos como *expulsar* (71)*,* y *sacar (72)*. En esta asociación, cuando el sujeto gramatical coincide con las personas migrantes se utilizan los verbos *largar* (73)*, salir (74), regresar* (75)**,** sin embargo, aunque con una estrategia distinta en la que quien enuncia demanda tales acciones, se expresa igualmente la oposición a la población migrante.

(71) Deben ser **expulsados** de nuestro país inmediatamente!

(72) Muy bien GN **saquen** *a esa gente* de aquí.

(73) Pues esperemos que ya todos tus vecinos se **larguen** de nuestro país ya tenemos bastante delincuencia

(74) Deberían ponerse a trabajar y **salir** de nuestro país!!!!

(75) que se **regrese** tu gente de honuduras aquí en Estados Unidos no queremos a tu gente

En resumen, lo que sucede en la asociación MOVIMIENTO no es la intención de los hablantes de referenciar flujos migratorios conocidos, sino la de valorarlos, esto es, la de expresar su postura ideológica respecto a aquellos. Los n-gramas estudiados son parte de una estrategia lingüística en la que se marcan puntos deícticos para referir las direcciones de los movimientos, a saber, se considera que cuando se sale de un punto de origen no es un viaje de destino vago, sino que es con la intención de llegar al punto desde el que el usuario escribe, y desde el cual se conceptualiza y valora si tal viaje debería o no suceder. Lo anterior se observa en la codificación de elementos morales tanto en los eventos de prosodia positiva como negativa.

El hecho de que la principal intención dentro de esta asociación no sea referencial, apunta al uso de estrategias en las que el hablante imprime su subjetividad. Si bien ya hemos visto aquí alguna de tales estrategias, veremos que en el resto de asociaciones también ocurren. También veremos en su momento que a menudo estas asociaciones tienden a anidarse entre sí.

7.2.2.2 PELIGRO

Esta asociación está compuesta por dos subasociaciones que juntas contribuyen al discurso de los migrantes vistos como un peligro para el país al que llegan. El subconjunto que se denominó *amenaza* denota, de hecho, dicho peligro (76); en oposición, el subconjunto *protección* expresa las acciones que se deben –a valoración de los hablantes– hacer ante tales amenazas (77). A propósito, vale decir que esta asociación, que es la tercera más frecuente en el corpus con 67 ocurrencias, tiene en su haber dos concordancias de prosodia positiva. Una de ellas pertenece a un hablante hondureño que exhorta a sus connacionales a apreciar su país, aunque por lo tanto, a no salirse de él (78); otra es una crítica a políticas contra migrantes

(76) Nos faltaban **estupidos miserables migrantes golpeando mexicanos aquí en *mexico*** pche gente que sigue ayudando estas lacra (amenaza-prosodia negativa)

(77) Por que dejan entrar a maras a mexico?? **Aquí en *Puebla*** ya **los estamos esperando** perros (protección- prosodía negativa)

(78) porque Honduras es maravilloso y pues lo único que puedo desearles a esas personas que van en la caravana es suerte y que yo se que es por necesidad que se van pero que aprendamos ***a volar [valorar] lo hermoso*** **que es nuestro pais** (protección- prosodia positiva)

(79) … **si [sí] me importa, es el futuro de migrantes y de nuestro país**. No hables por mi. Por favor Miserable y cobarde. Trump le impuso una política migratoria contraria a nuestras leyes y aquí casi se le agradece. MI SE RA BLE. (protección- prosodia positiva)

En la Tabla 17 se observan los sets semántico-gramaticales de los elementos cercanos a los n-gramas de interés. Dado que *amenaza* activa tanto las acciones amenazantes como a los ofensores, y *protección* activa tanto a los protectores como a las acciones para defender ante el peligro, se cuenta con dos sets semánticos para cada subconjunto. En tanto el subconjunto *amenaza* fue más productivo que el de *protección* (46 y 12 casos, respectivamente) era de esperarse además más diversidad entre las elecciones léxicas de los hablantes en el primero de estos campos, por ello fue necesario subdividir en otras categorías sus respectivos sets semánticos. El set semántico correspondiente a las acciones amenazantes fue subdividido en dos categorías. En la primera, los verbos que componen este set describen una acción transitiva dañina y agentiva. En la segunda, los verbos carecen de estos rasgos semánticos pero son usados para escribir una situación perjudicial para el país destino como consecuencia de la llegada de la migración.

Tabla 17. Sets semánticos que conforman la asociación PELIGRO

Amenaza		**Protección**
Ofensores	**amenazas**	**protectores**
Etnias y nacionalidades: judios, españoles, (la gente de) Honduras/hondureños	***Agente (quien migra) + hacer daño*** Adueñar, arruinar, chingar, dañar, desestabilizar, golpear, imponer, invadir,	Mexicanos, Migrantes mexicanos, Ejército, Presidente

Centroamérica/ centroamericanos, **Colectivos:** Gente, invasores, migrantes, inmigrantes, pandillas, personas, calificaciones: miserable, estúpido, ilegales	matar, pisotear, respetar, robar, saquear, tumbar, volar (las leyes) ***Situación*** Andar, aumentar, dejar pasar, ser	

Así pues, entre las acciones del subconjunto amenazas están las que corresponden a un léxico militarista, como *saquear,* o las perífrasis verbales que dirigen la acción agresiva hacia el punto déctico desde el que el hablante se conceptualiza (venir + *invadir/ hacer guerra* + nuestro país/ aquí en). Junto al uso de léxico militarista, como sucedió en MOVIMIENTO, se observa también una estrategia de subjetivación en la que el hablante valora no solo como peligroso el hecho de la inmigración, sino peligroso contra un territorio que habita. O mejor dicho, contra algo de lo que se considera parte, de modo que si bien los ejemplos del léxico belicoso corresponden a metáforas de guerra (Baker & McEnery, 2005; Camargo Fernández, 2021; Taylor, 2009, 2021) –o quizá, representaciones sociales promovidas por otros discurso dominantes– y por lo tanto a expresiones abstractas de lo que es la migración, los usuarios también dieron muestras de amenazas más concretas sobre el daño que representan las personas al país que las recibe. Entre estas amenazas se encuentra el daño hacia un territorio (80) a (82) , la violencia hacia personas de instituciones nacionales (83), o el ignorar los códigos legales (84) a (86).

(80) YA SABE QUE CLASE DE *GENTE MISERABLE* ***A SAQUEADO*** **A NUESTRO PAÍS** Y SI HAY PERSONAS QUE LOS ORGANIZAN

(81) a su país no se vale que *venga a invadir* nuestro pais

(82) No mas hondureños *Vienen a hacer guerra* aquí en *México*

(83) la emigració **aquí en México** para fastidiar al presidente Obrador

(84) …la guardia Nacional con piedras y **pisoteando las leyes migratorias de nuestro país** no esperen que los traten bien.

(85) …es posible que esa gente quiera imponer su ley en nuestro país

(86) A detener a esos invasores **que *no respetan*** *la soberanía* **de nuestro país**

Esta, que es la más productiva subdivisión de las acciones amenazantes, tiene como característica que el sujeto gramatical coincide con la población migrante. En esta el sujeto está marcado como ofensor; tal marca no solo sucede al hacerlos sujetos gramaticales de verbos con valor semántico de daño, sino que también lo es al ser explícitamente codificado en frases nominales. Entre estas frases nominales vemos que se habla de etnias y nacionalidades (ej. *Judíos, Honduras*); también se observa que se marcan colectivos que no necesariamente referencian el punto de origen de los migrantes (ej. *gente* o *personas*) y pueden estar codificando además la amenaza que representan como colectivo (ej. *pandillas, invasores*). Asimismo, se encuentra la presencia de adjetivos que califican a los migrantes con una prosodia negativa (ej. *estúpido, ilegales*).

Como decíamos, también ocurre que los verbos no cuentan con el rasgo de agentividad (**¡Error! No se encuentra el origen de la referencia.**); a partir de estos elementos léxicos se describe una situación en el que la presencia de la población migrante representa un daño para el país que los recibe. En tanto lo que el hablante hace en estos ejemplos es resaltar una situación peligrosa, en esta categoría el sujeto gramatical puede o no coincidir con los migrantes. Lo que sucede en los ejemplos (87) y (88) es la expresión de la valoración negativa sobre el ingreso o la permanencia de personas migrantes. Dicha valoración sucede, en (87), mediante la introducción de la premonición de un peligro vislumbrado; en (88) y (89), entre tanto, se describen peligros inmediatos –igualmente percibidos por los hablantes– como un daño económico (Tenemos que pagar todo) o, el daño que supone una forma de ingreso (Emigrar no es igual a invadir) para concluir con la valoración negativa.

(87) ***Q será*** **de nuestro país** cuando los chikillos estos crezcan.

(88) Tenemos que pagar todo *eso no es nada bueno* para nuestro país

(89) Emigrar no es igual a invadir. Son un peligro para nuestro país

Por otro lado, lo que sucede en el subconjunto *protección* es el contraste con lo visto hasta el momento en esta asociación. Es decir, se marca una "línea de defensa", tanto con el uso de verbos –la mayoría con el rasgo de agentividad que puede ser coincidente con un sujeto perteneciente al país de origen– como con sujetos explícitos en los que recae la responsabilidad de tal protección. Estos están vinculados con un punto deíctico particular –México, la mayoría de los casos–, tanto con la vinculación que da el gentilicio, como por la referencia a las instituciones o actores encargados de la protección (Ejército, presidente).

En esta oposición que supone el subconjunto PROTECCIÓN a AMENAZA, se activa igualmente léxico militarista con verbos como *defender* y *proteger*. Destaca, no obstante, que el sujeto gramatical de la mayoría de las concordancias de estos verbos no coinciden con los actores sociales en los que recae la responsabilidad de la defensa (solo 1 vez se nombra al presidente (90) y otra al ejército (91). Como se vio anteriormente, el léxico militarista sucedía a la par de una estrategia de subjetivización en la que el hablante consideraba que la migración era una amenaza para el punto deíctico desde el que habla y al que pertenece; en esta ocasión el hablante se suma a la defensa de dicho punto deíctico al contarse él mismo en el sujeto gramatical, mayormente codificado como *mexicanos (92)* y (93).

(90) Iniciativa Porfavor presidente si amas nuestro pais Mexico no es posible que tengamos que lidiar…

(91) EL EJERCITO PARA ESO ESTA PARA DEFENDER LA SOBERANIAAA DE NUESTRO PAIS

(92) *Levantense* ***Mexicanos*** a **defender a nuestro Pais** de estos INVASORES CRIMINALES!!!!!!!!!!

(93) Hermanos Mexicanos es URGENTE unirnos y protejer y defender a nuestros Pais!!

Como se concluye hasta ahora, de lo observado en la clasificación en la Tabla 17, se desprende que entre los subconjuntos que conforman esta asociación existe una relación semántica de oposición que está marcada por el punto deíctico del que dan pauta los n-gramas. De una manera semejante a lo que sucede con las asociación de MOVIMIENTO, en esta se utilizan estrategias de subjetivación. Particularmente, marcan a tal punto deíctico como nacionalista, como un espacio circunscrito al territorio mexicano. De modo que la oposición que se aprecia desde los campos semánticos de la tabla 17, es lo extranjero y la ofensa. En términos del Análisis Crítico del Discurso: el Otro.

7.2.2.3 ACTITUD

Esta asociación es la segunda más frecuente en los datos, suma 88 ocurrencias. Aquí, como en el resto de las asociaciones, se observa que la principal intención de los hablantes no es referenciar sino expresar una postura ideológica sobre la migración. Organiza un discurso dicotómico por el que se conceptualiza una situación precaria tanto a quienes migran como al territorio al que pertenecen. Se optó por dividir este discurso en dos sets semánticos. El primero de ellos es *problemática*, y comprende aquellas concordancias en las que existe la descripción de una causa de la migración reconocida por el hablante (94). El segundo es *asistencia* que abarca aquellos modos que el hablante reconoce como una posibilidad de ayuda a las personas migrantes (95).

(94) Aquí en *México* ya tenemos problemas con nuestro presidente pero nos quedamos en nuestro país a seguir luchando

(95) muchos MEXICANOS estan siendo olvidados, estamos **gastando los impuestos de nuestro país** en ellos

En la Tabla 18 se muestra la distribución de los verbos que coocurren con los n-gramas seleccionados. De los subconjuntos, además se destacan dos rasgos semánticos, a saber la

agencia o falta de agencia que corresponden a *asistencia* y *problemática* respectivamente y de manera exclusiva.

Tabla 18. Set semánticos que conforman la asociación ACTITUD

	Problemática	**Asistencia**
-agencia	Aumentar, costar, estar, faltar, fastidiar, haber, hablar de, liberar, llevar en, padecer, palear, pasar en, saber, ser, subir, tapar, tener, trabajar, valer	
+ agencia		apoyar, ayudar, conseguir, dar, dejar de lado, escasear, estar, exigir justicia gastar, guardar, haber, habilitar, mantener, poder, poner a llorar, preocuparse, querer, salir, ser unidos, tener, tratar, ver

Como puede observarse en la Tabla 18, *problemática* es uno de los dos subconjuntos donde hay más concordancias con prosodia no negativa hacia la migración que negativa. Bajo este rubro se registraron dos estrategias de los hablantes. En la primera de ellas se introduce un tema, esta vez el de las causas –negativas todas, de allí el nombre de *problemática*– de la migración con varias realizaciones verbales. Entre estas se encuentran el uso de verbos que denotan existencia como *haber* y *pasar (96) y (97)*; el uso del verbo *estar* para marcar atribuciones (98); y también dentro del uso de frases verbales se encuentran aquellas acciones cuyo valor semántico apunta a una situación concreta más bien negativa (padecer, fallar) (99) y (100).

(96) Hola yo soy hondureña también y me da mucha tristeza pero es la viva realidad de **lo que esta pasando en nuestro pais**

(97) **En nuestro país hay *de masiada corrucion*** el presidente q tensmos es un estadafor duele ver a nuestros hermanos hondureños asi (Youtube)

(98) honduras es un pais pobre nuestros gobiernos son corruptos estamos al borde del colapso **nuestro país está *igual q venezuela*** nesesitamos ayuda

(99) AL PARECER TELEMUNDO ESTA PRESENTA ESTA SITUACION COMO QUE ES JUSTO QUE PASEN SIN IMPORTAR ***LA POBREZA QUE*** **PADECE NUESTRO PAIS**

(100) Talento inmigrado, nomas *eso* **faltaba** ante la fuga de cerebros **de nuestro pais**. (problemática-prosodia negativa, Twitter)

Esta estrategia recuerda lo que pasaba entre los ejemplos del subconjunto *amenaza*, donde se usan frases verbales sin el rasgo de agentividad, pero que contienen la noción de peligro en tanto denotan que la presencia de los migrantes en un país al que no pertenecen es dañino para dicho territorio (87) a (89). En estos otros ejemplo, también se observa que no es importante marcar un agente sino describir las condiciones de los territorios de origen de las personas migrantes.

Se encontró una segunda estrategia con la que se introduce la causa de la migración como tópico, en esta ocasión como una forma de fijar postura. Esta estrategia se usa, de hecho, por personas que se identifican como centroamericanos y tratan de apelar a la comprensión de los mexicanos (101) a (104). En todos los ejemplos citados, los hablantes enuncian el punto deíctico desde donde conceptualizan su posición ideológica y desde la que apelan a la empatía; una estrategia parecida que sucedió en ejemplos del subconjunto *salida* con prosodia no negativa (62) donde los usuarios de redes sociales se contaban como parte de los sujetos migrantes.

(101) **aquí en *Honduras*** no hay trabajo y la corrupción es lo que más se mira aquí y lo peor es aquí en Honduras le suben a todo a los impuestos,la gasolina, al transporte, las medicinas,comida etc

(102) **aquí en *Honduras*** no te vale el estudio mientras no tengas alguna persona de cuello para que tengas un trabajo

(103) **aquí en *Honduras*** no hay tanta seguridad

(104) **Aquí en *Honduras*** te cuesta conseguir trabajo si tenes mas de 35 años y

No obstante, también sucedió el caso contrario, esto es, que se usara para marcar la postura en contra de la migración (94). Sin embargo, la manifestación de una y otra posición coincide en que la oposición se marca desde el punto deíctico desde el que el hablante se enuncia.

Ahora bien, centrándose en las causas de la migración, se observa que todas apuntan a hechos negativos (hambre, pobreza, corrupción, ineficiencia de las autoridades políticas), esto es, se refiere la situación precaria que orilla a la migración. La contraparte de este este discurso está contenida bajo el segundo de los subconjuntos de esta asociación, es decir, *asistencia*. Esto quiere decir que entre ambos subconjuntos hay una relación semántica, el primer de ellos señala los problemas, el segundo, las soluciones; sin embargo, son los usuarios quienes evalúan los eventos migratorios, aunque está vez la opinión está más bien focalizada en el subconjunto *asistencia*.

En la Tabla 18, el set semántico del subconjunto *asistencia* está conformado por varias realizaciones verbales que, en su contexto, referencian tipos de ayuda que el hablante identifica como posibles para socorrer a la población migrante. Posible porque, como se verá, no necesariamente se considera que deban aplicarse, de hecho, como puede observarse en la **¡Error! No se encuentra el origen de la referencia.**, 42 de los ejemplos con esta asociación corresponden a una prosodia negativa frente a nueve con prosodia positiva. Entre estos tipos de ayuda, se encuentra asistencia jurídica (105); o apoyo económico (106) y (107). La mayoría de los ejemplos, no obstante, tratan en general el tema de la ayuda para la población migrante.

(105) Dejemos de ser un país de hipócritas y TAMBIÉN EXIJAMOS JUSTICIA POR LOS MIGRANTES asesinados en nuestro país

(106) si yo mirara que vinieran preparados en alimentos no diría nada pero **aquí en la frontera *México los está mánteniendo***

(107) aparte por estar atendiendolos a ellos muchos MEXICANOS estan siendo olvidados, **estamos gastando los impuestos de nuestro país** en ellos

(108) pienso que *primero* ***deverian ayudar*** **a los de nuestro país** y después a los migrantes

(109) pero por que no nos ayudaron a nosotros aquí en nayarit?

Sin embargo, si bien es cierto que en los ejemplos anteriores hay, entre los n-gramas y su contexto lingüístico inmediato, una referencia a los posibles apoyos para la población migrante, lo que tienen en común es más bien la valoración ante el hecho de ayudar. Se establece una relación que contrasta a un *Nosotros* (mexicanos, "a los de nuestros país", "a nosotros aquí en nayarit") frente a un Otro. A propósito, lo que sucede en (105) es otro caso de *priming pasivo*, en el que el hablante acusa una actitud –particularmente una acción política que es llevada acabo a través de actos de habla: exigir– de ignorar un tipo de injusticia que viven las personas migrantes en México; ello puede apuntar a un conocimiento cultural, o socialmente compartido, sobre la evaluación del merecimiento de este apoyo que ya se había visto en los subconjuntos de *entrada* y *estadía*, cuando "los mexicanos" debían restringir tales acciones a las personas migrantes dependiendo de sus actitudes morales. En ese sentido, igual que en las asociaciones anteriores, los hablantes, mediante la relación que reconocen a cada actor con un punto deíctico al que se pertenece, califican si se es o no merecedor de asistencia.

Hasta ahora, se han identificado patrones semánticos y gramaticales en los que ocurrieron los n-gramas (*aquí en, (en) nuestro país)* y se identificaron tres asociaciones semánticas en las que participaban tales n-gramas; además, se identificaron diferentes subconjuntos semánticos que componen a tales asociaciones. De igual manera, el lector habrá podido observar coincidencias entre tales asociaciones. En la siguiente sección se propone que dichas coincidencias obedecen al efecto del léxico de los n-gramas, particularmente a las propiedades deícticas que también serán explicadas a continuación.

7.2.3 Cohesión intertextual en el discurso xenofóbico de mexicanos en redes sociales

Esta es la última sección en la que se describen los resultados de esta investigación. El objetivo principal de este proyecto fue comprobar si era posible hablar de un solo discurso

xenofóbico en las redes sociales a partir de los datos de dos corpus diferentes, pero con la temática en común en torno a una discusión ideológica.

Esta hipótesis está sustentada por un lado en el ACD y, por otro, en la propuesta de la activación léxica. Entendiendo por discurso formas semióticas de construir aspectos del mundo que pueden ser identificadas en la codificación, por ejemplo lingüística, de posiciones o perspectivas hacia diferentes grupos o actores sociales (Fairclough, 2010). Además, se sugiere que esta postura propia del ACD es operable bajo los lineamientos teóricos y metodológicos de la activación léxica en la medida en que la observación de la cohesión ha sido aplicada también para la intertextualidad de corpus temáticos, es decir, aquellos que no están conformados por las producciones de un solo autor sino por diversidad de ellos (Hoey, 2005, 2017). En otras palabras, se define el discurso xenofóbico como prácticas observables en las dinámicas propias de las redes sociales bajo análisis.

Estas prácticas ya han sido observadas en alguna medida y aquí se termina de explorarlas. La primera evidencia de esas prácticas, o de ese discurso cohesionado, es la repetición de distintos n-gramas (elementos con ganancia de información encontrados en una sola red social mediante una técnica de clasificación automática), en una segunda red social; se suma el hecho de que los usos que les dan los usuarios son los mismos al grado de que conforman las mismas asociaciones semánticas tanto en Twitter como en YouTube. Dentro de los ejemplos de las asociaciones, además, hubo evidencia de priming pasivo, lo que da cuenta de que los hablantes tienen conocimiento de posiciones ideológicas que no son las propias y aunque las ven con recelo, son un indicio de que hay una visión compartida por varios miembros de su comunidad y conocida por otros que no comulgan con ella.

La segunda evidencia a favor de un solo discurso cohesionado tiene que ver con las relaciones semánticas que se establecen entre los subconjuntos de cada una de las asociaciones. A saber, en la asociación semántica de MOVIMIENTO hay relaciones de contraste –esta noción de oposición está presente en las tres asociaciones–; en la de la asociación de PELIGRO, de causa y efecto; y en la asociación de ACTITUD, de problema y solución. Dichas relaciones son posibles, de acuerdo con Hoey, porque para ello está preparado cada ítem léxico y es, de hecho, su efecto. Esto es, cada palabra que usa un

hablante puede formar relaciones cohesivas al tiempo que genera expectativas en el interlocutor. Esta expectación se resume en la suposición de que ambos comparten el mismo conocimiento, es decir, el de que cada palabra que diga uno de ellos es capaz activar otros elementos lingüísticos. Estos elementos son, para el caso particular del estudio de las expresiones ideológicas que aquí nos ocupan, las relaciones semánticas que, de hecho, construyen.

Lo que el lector puede encontrar en esta sección es, precisamente, las propiedades léxicas de los n-gramas estudiados *aquí en* y *(en) nuestro país*, con el propósito de describir cómo es que hacen posible que se formen las relaciones semánticas de las que ya hablamos. Esto nos llevará a describir un uso compartido, tanto en tales n-gramas como en otros elementos del campo de locación de la lista con ganancia de información: a saber, un sentido de cercanía que circunscribe a la patria.

7.2.3.1 Para lo qué está preparado el léxico: relaciones semánticas

Las tres asociaciones descritas anteriormente tienen en común el hecho de que más que hablar sobre migración, la valoran. Esto se realiza a través de diferentes estrategias de subjetivización, que pueden revisarse en la sección anterior; que son posibles tanto por las restricciones del tópico sobre el que versan como porque dichos n-gramas ya están preparados para ello. Esto último es lo que se revisa a continuación. A grandes rasgos, ello sucede porque las tres cadenas de palabras que se escogieron para el análisis en corpus cuentan con propiedades deícticas.

Se ha dicho que encontramos tres asociaciones semánticas en las que se usan los n-gramas *(de) nuestro país y aquí en*. Estas son MOVIMIENTO, PELIGRO y ACTITUD. A propósito de la primera de estas asociaciones, se adelanta que no es sorprendente encontrar nociones de movimiento asociadas a referentes de lugar ni en un corpus sobre migración. Dejando eso de lado, otros hallazgos son relevantes.

MOVIMIENTO es una asociación compuesta por cuatro subconjuntos que describen los flujos migratorios: *salida, entrada, estadía y retorno.* Estos movimientos se dan con relación

a un punto deíctico que, como se mostró, los hablantes tienden a relacionar consigo mismo. A saber:

- *Salida* y *retorno* son dos subconjuntos que se vinculan en una relación de oposición. El primero, que tiene en su haber más ocurrencias de concordancias de prosodia positiva hacia la migración, trata de cómo los migrantes abandonan sus países. El segundo, entre tanto, describe un supuesto deber de los mexicanos por expulsar a los migrantes. Es relevante que el lugar, una vez fuera de México, no parece importar.
- Si bien ambos subconjuntos son los menos frecuentes entre todos los que encontramos (con 3 y 13 ocurrencias, respectivamente), *salida* ocurre aun 10 veces menos. De modo que, por alguna razón, no es relevante para los usuarios de redes sociales en México hablar sobre el porqué los migrantes que ingresan a este país, salen de los suyos. Quizá tal razón sea precisamente la oposición en que, desde el discurso mexicano, se construye un Otro al que se opone.
- *Entrada* y *estadía* son dos subconjuntos que comparten rasgos semánticos importantes, a saber, la noción de dinámica de fuerzas en la que participan agentes y pacientes. Estos rasgos se han encontrado en un discurso que opone a los mexicanos como agentes que obstruyen el ingreso o la permanencia de personas centroamericanas o caribeñas (pacientes). También sucede que lo que opone la resistencia a que los migrantes estén en México no son personas como tales, sino las cualidades morales, esto es, cada migrante debe ser merecedor de estar en territorio mexicano.

PELIGRO está compuesta por los subconjuntos de *amenaza* y *protección*. También es un discurso de oposición, esta vez, sin embargo, es un discurso explícitamente nacionalista en el que los mexicanos forman un grupo frente el inminente peligro del Otro. Para ello:

- Se utiliza léxico militarista.
- Se codifica a los migrantes como amenaza.
- Se construye una noción del deber que tienen los mexicanos para imponerse a las personas migrantes, independiente de las instituciones oficiales que tengan la

atribución de controlar, por las razones sociales o políticas que sean, los flujos migratorios

ACTITUD es igualmente un discurso dicotómico, sus dos subconjuntos son *problemática* y *asistencia*. Aquí se construye el conocimiento más "objetivo" de los hechos migratorios.

- Los n-gramas son utilizados para conceptualizar un territorio al que se ve desde la distancia y por lo tanto con relativa objetividad, y lo que los valida a emitir opiniones. El territorio que se codifica es el de los países de origen de los migrantes, a estos espacios se les caracteriza por motivos negativos como pobreza, corrupción, de modo que se explican las causas de la migración.
- Perfiladas las causas de la migración como problemas, las soluciones componen el segundo set semántico de esta asociación, *asistencia*. En tanto vienen de un espacio, la situación del migrante es precaria y necesitan ayuda. No obstante, los hablantes tienden a construir una noción de merecimiento de dicha asistencia. Por su parte, el merecimiento está condicionado a las cualidades morales de los migrantes.

No solo en este trabajo, ni solo para el caso de México, se ha encontrado que en los discursos sobre migración se construye a quien migra como *Otro*, mientras que las personas del país destino se perciben como un gran *Nosotros* (Bevitori, 2018; Galindo Gómez, 2019; Hartnett, 2019; Islentyeva, 2021; Lawson, 2015; Montali et al., 2013; O'Regan & Riordan, 2018; Salahshour, 2016; Taylor, 2009). Así, los migrantes son, en el mejor de los casos, sujetos a un examen por parte del *Nosotros* para que *determinemos* si *merecen* estar en *nuestro territorio;* en el peor de los casos son peligro inminente de guerra.

El ACD propone la revisión de diferentes formas semióticas para conocer cómo sucede esta construcción. Bajo esta lógica, aquí se revisan cómo a partir de propiedades deícticas que permiten a los hablantes conceptualizarse como más o menos cercanos a un territorios se refuerza esta construcción de Nosotros vs Los Otros, o mejor, cómo se construye la noción de NUESTRO TERRITORIO.

Al haber seleccionado n-gramas con propiedades deícticas, es obligatorio hablar de la reconstrucción mental del espacio que se está haciendo en las expresiones xenofóbicas. Los datos presentados coinciden con varias descripciones del adverbio *aquí* en cuanto a que son

sus propiedades puntuales las que lo llevan a señalar un espacio tan específico (Maldonado, 2013; Sedano, 1994) como un sitio geográfico, ya sea un país (110) y (111) o una ciudad (112).

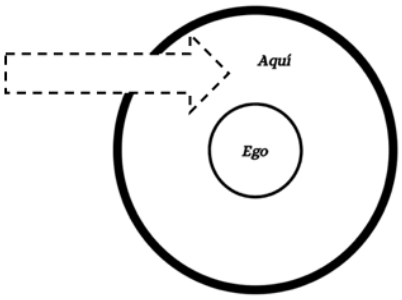
(110) Con la pena pero **aquí en *México*** no se pueden quedar (YouTube)

(111) **…aquí en *Estados unidos*** el gobierno ya no quiere indocumentados (YouTube)

(112) En la décima norte **aquí en Tapachula, Chiapas** se encuentra una cantidad considerable de migrantes

Se puede hablar de un lugar específico en tanto existe la proximidad adecuada en la reconstrucción mental del hablante que se percibe a sí mismo como referente (ego). *Aquí* establecería un punto a una distancia tal que demarque bien el país o la ciudad del que habla, y al que, si bien al mismo tiempo ocupa, no se confunde con el hablante; esta mirada con mayor o menor distancia entre el hablante y lo que refiere es la subjetividad. Lo dicho se representa en Ilustración 17 (imagen adecuada a partir de lo expuesto en Maldonado 2013), la posibilidad de que *aquí* participe en construcciones que denoten una trayectoria hacia el espacio señalado.

Ilustración 17. Aquí



Entre tanto, sucede algo parecido con los n-gramas *(de) nuestro país*. En este caso, el efecto de subjetividad lo añade el adjetivo posesivo de primera persona. Tal efecto es una manera

de expresar distancia o cercanía. En primer lugar, se observa que cuando *país* está antecedido por un determinante definido que no marque la posesión es porque ya sea el contexto del hablante, o el que él ofrece lingüísticamente, deja claro de qué país está hablando. Probablemente el propio, o desde el que se hable como en (113).

(113) Trabajo y empleo femenino en Chile 1880-2000. Su aporte al desarrollo del país desde la economía doméstica (corpus de referencia)

Entre tanto, cuando antecede al sustantivo *país,* se puede establecer la distancia entre el país referido por el hablante y el hablante mismo. En (114) el hablante necesita codificar el posesivo porque no está en su país. La cercanía o distancia, no obstante, puede ser también emocional, esto es, el hablante usará el posesivo para imprimir subjetividad como en (115). Dicha subjetividad la entendemos como la marca de la distancia con respecto del mismo hablante que se establece a sí mismo como punto de referencia.

(114) Mi forma migratoria está por vencer y el aeropuerto de **mi país** está cerrado, qué debo hacer para regularizar. (Twitter)

Tal subjetividad, además, hará posible la expresión de la postura ideológica cuando haya un tópico de interés social y algún posesivo antecediendo a *país*. Como en (115), referente al régimen político, y (116) referente a un posible enfrentamiento entre alguien perteneciente a un país, frente alguien que visita tal país, o bien, ante el disgusto de la sola presencia de un extranjero (117).

(115) La izquierda acabó con **mi país**, Venezuela!! (Twitter)

(116) Nadie puede *venir a insultarte* en **tu país**. (Twitter)

(117) A los migrantes deben masacrarlos con plomo grueso, que **se larguen a su país** bola de MIERDAS (Twitter)

Una acotación necesaria es que, si bien en los primeros cien elementos de la lista de elementos con ganancia de información (apéndice **¡Error! No se encuentra el origen de la referencia.**) solo apareció el adjetivo posesivo para primera persona del plural, la

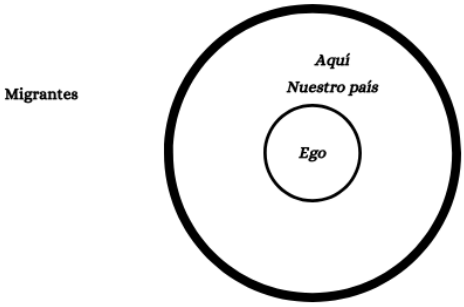
vinculación que se hace entre este y el discurso xenofóbico es, por un lado, meramente proporcional pues aun en los corpus utilizados existen los casos contrarios (118). Por otro, vemos que el discurso xenofóbico se puede manifestar con el resto de los adjetivos del paradigma como se demuestra en (116) y (117).

(118) denle amor a la página de @fm4pasolibre, son una fundación q cobija a los hermanos inmigrantes q **llegan a refugiarse** a **nuestro país** (Twitter)

Ahora bien, no solo es relevante el sustantivo *país*, sino que se deb hablar además de un set semántico de TERRITORIO. Llamado así, en primer lugar, porque, como lo deja claro la mirada subjetiva del hablante, no están hablando de cualquier país al emitir una opinión –xenofóbica para el particular interés de este trabajo– sino que están hablando de un espacio con el que hay una cercanía emocional. En segundo lugar, porque la delimitación de este espacio también queda determinada por el uso del adverbio *aquí*, cuya función referencial al lugar existente es menos clara, pero permite entrever la distancia con respecto al hablante. Y en tercer lugar, porque a partir de estas dos razones, y también de otros datos que se aprecian en (Tabla 13. Distribución de rasgos del campo semántico "lugar" en dos redes sociales, p. 110), se ha observado que, en general, se hace referencia a una demarcación geopolítica, que, en la mente de los hablantes es una causa para oponerse a otros que no son pertenecientes a esa demarcación.

Dicha noción está activada por el factor social de un sentimiento patriótico que activa a su vez, el uso de recursos como *aquí* o *nuestro (+ país)*, en la medida en que son recursos preparados para expresar tal sentimiento. De ese modo, sugerimos que la percepción que hace el hablante de la migración es como sucede en Ilustración 18. aquí/ nuestro país. Ego comparte el espacio con sus connacionales, pertenecen de hecho a ese territorio; el migrante, que *viene* de otro espacio (especificar cuál no es relevante) es, por tanto, el Otro.

Ilustración 18. aquí/ nuestro país



Así, se considera que otros dos elementos del campo semántico de LUGAR (Tabla 13. Distribución de rasgos del campo semántico "lugar" en dos redes sociales, p. 110) deben formar parte de dicho conjunto: nación y suelo. Suponiendo con ello que estas locaciones deben incluirse en TERRITORIO, por la misma razón de que la codificación de la primera persona del plural en la expresión de posturas xenofóbicas responde a la manifestación de un sentido de pertenencia que, en última instancia, se basa en la oposición entre el Otro y Nosotros.

Suelo se considera indicador del sentido patriótico porque es parte del himno de México (119). De hecho, entre las concordancias de los corpus de ambas redes sociales, se hallaron referencias a este símbolo patrio (120), en el corpus de los comentarios de YouTube se encontró textualmente esta parte del himno en 3 de 12 videos.

(119) Mas si osare un extraño enemigo // profanar con su planta tu suelo

(120) La agresión de los migrantes a la guardia nacional del KKS ayer en Chiapas es, es una profanación contra México por parte de un extranjero, dónde quedo: **Más si osare un extraño enemigo profanar con su planta tu suelo…** (Twitter)

Con esta constante referencia al himno mexicano, observamos otra evidencia de que el discurso xenofóbico contra migrantes es parte un conocimiento culturalmente compartido, ahora además está vinculado con la identidad mexicana históricamente construida. Junto con esta identidad, está presente la idea de que el territorio mexicano es algo que se

defiende, pero también es un deber que se extrapola a otras naciones, como sucede en (121) a (126).

(121) Mexicanos es hora de **defender** *nuestra nación* (Youtube)

(122) Nuestra Guardia Nacional Cumple con Su Deber…. **Protejer** *Nuestra Nación* (Youtube)

(123) Que clase de gobierno tenemos que no puede **proteger** a nuestro suelo.

(124) Trump es inocente por **defender** *su nación* (Twitter)

(125) El solo quiso **proteger** a la nación (Twitter)

(126) Que se vayan q **luchen** en su nación aquí no son bienvenidos

Para este punto es importante mencionar que para estos sustantivos, *nación* y *suelo,* solo se activa la asociación semántica de PELIGRO. Esta distribución lleva a la conclusión de que la cuestión nacionalista es un rasgo que organiza el discurso antimigrantes. En otras palabras, varios de los subconjuntos del resto de asociaciones también deben entenderse dentro de este marco. Con ello, además, se prueba que pueden establecerse relaciones semánticas en el conjunto de los dos corpus de redes sociales, de modo que puede hablarse de un discurso xenofóbico cohesionado.

Así, habría una relación de ejemplificación entre las subconjuntos de *entrada* y *estadía* con *amenaza*. Esto es, el mero hecho de que estén o permanezcan migrantes en nuestro territorio es una amenaza. Este sentido tiende a suceder cuando, en las concordancias de estos subconjuntos, las realizaciones verbales reservan el rasgo de agencia para los migrantes. Junto a esto, el resto del contexto lingüísticos hay marcas de contraexpectativa a las acciones que se atribuye a los migrantes; como en (127) *no tienen porque*; o en (128) que se califique a la entrada de *ilegal*. Obsérvese además que junto a la contraexpectativa ocurre la amenaza: la entrada y la estadía de personas migrantes es en sí misma amenazante.

(127) *No tienen porque entrar* de esa forma a **nuestro país** y menos exigiendo derechos que no les corresponden

(128) …es que la mayoría de *los migrantes entraron* **nuestro país** ilegalmente y ya cuando están aquí exigen…

A propósito de expectativas, si un usuario introduce la noción de amenaza, es esperable para sus interlocutores que la referencia a la respuesta a tal amenaza, la *protección*, ocurra en algún momento. Si bien es esto lo que sucede con PELIGRO, también ocurrió dentro del subconjunto de MOVIMIENTO, y de hecho con los mismos subconjuntos de *entrada* y *estadía*. Como se observa en (129), el ingreso va en contra de las expectativas, o de los deseos del hablante que se ubica dentro del punto deíctico *México*. Dice su postura abiertamente: *estoy en contra*.

(129) Yo estoy en contra de los hondureños que están aquí en México

Ya se vio, cuando se describía MOVIMIENTO, que en estos dos subconjuntos está presente una dinámica de fuerza por medio de la cual *los mexicanos* dificultan el paso o la permanencia de los migrantes en su territorio. En el mismo sentido, pero con un grado mayor de severidad, en el subcampo de *retorno* hay una oposición tajante que lleva a expulsar a los migrantes. Sin detenerse aquí en lo que se puede leer en la sección 7.2.2.1 MOVIMIENTO (p. 121), se defiende que las categorías de *retorno, entrada* y *estadía* guardan, en ocasiones, una relación de ejemplificación con *protección*. También es relevante observar cómo estas expresiones de protección están presentes en general en el discurso xenofóbico, incluso a pesar del intento de establecer tres categorías discretas para explicarlo.

Particularmente, estas expectativas están presentes por la demarcación de un espacio nacionalista. Se considera además que la distribución de los subconjuntos de MOVIMIENTO son evidencia del discurso nacionalista. *Salida* apenas ocurre porque no interesa *lo propio* de los migrantes; *retorno* presenta una prosodia negativa casi total porque expulsar migrantes es una cuestión patriótica; y en un sentido similar, las concordancias de *estadia* y *entrada* pueden presentar tanto la contraexpectativa a los flujos migratorios (y con ello el

disgusto) o dinámica de fuerza que pone estándares morales para decidir quién continúa su paso y quién no.

El campo semántico de ACTITUD es un discurso dicotómico que categoriza las causas de la migración como propias de una situación precaria, por un lado, al tiempo que conceptualiza, con ello, a la población migrante como necesitada de asistencia, por otro. Como se vio en (101) a (104), la mayoría de las concordancias que se clasificaron bajo uno de estos subconjuntos, *problemática*, corresponden a centroamericanos trayendo al foco de la conversación en redes sociales la situación desfavorecida que hay en sus países. Nótese que sucede entonces la implementación de la subjetivización que se grafica en la Ilustración 17, que no en la Ilustración 18. Se remarca que esta es una asociación poco frecuente entre mexicanos, y de hecho, es coherente con la hipótesis recién propuesta según la cual lo que sucede con el subconjunto de *salida* no interesa. Y más: no interesa en la medida en que no sucede en el territorio mexicano. Al respecto, se ejemplifica con lo que sucede en (130), donde se prioriza que sean las personas que sí son del territorio, quienes merecen la asistencia, tal como se sugiere en la Ilustración 18. aquí/ nuestro país.

(130) Disculpa pero estamos dejando de un lado a la gente de **nuestro país *México*** para darle prioridad a alguien ajeno.

Si bien este ejemplo también funciona como *primig pasivo* –esto es, una persona mexicana referencia una situación que normalmente es dicha por personas no mexicanas además con el objetivo de expresar la posición ideológica contraria– también es una muestra de la expectativa de la que se estado hablando, es decir, de la activación de un sentimiento nacionalista. Específicamente, el subconjunto de *asistencia* se comporta como una especie de defensa, de la labor a hacerse contra la amenaza migrante.

Hasta aquí se muestran los resultados de una investigación guiada por datos. Como ya se mencionaba, se trató de un trabajo multidisciplinar que pretendió revisar la constitución lingüística del discurso xenfóbico mexicano en redes sociales. Un estudio como este, cuya pretensión es no solo es dar cuenta de un elemento lingüístico sino del corpus donde proviene, necesariamente enfrenta la cuestión de otros campos de análisis, algunas de estas posibilidades se discutirán en la siguiente sección.

8 Discusión

Anteriormente se expuso un panorama de diversas posiciones teóricas con una hipótesis común, según la cual *las ideas* que nos hacemos sobre otras personas son parte de un proceso cognitivo (Fairclough, 2010; Fiske et al., 2009; Massey, 2008; van Dijk, 2016; Wang et al., 2011; White et al., 2009). Desde la psicología social y la sociología, el prejuicio y la segregación social han sido explicados como consecuencia del tal proceso al involucrar emociones, positivas o negativas, asociadas con grupos particulares (Fiske et al., 2009; Harris & Fiske, 2006; Massey, 2008). Según el ACD, tales ideas serían expresadas en discursos, que en términos generales, se conocen como las formas semióticas desde las cuales se construye un aspecto social; tanto la producción como la constante repetición de tales discursos sería parte de ese proceso cognitivo (Fairclough, 2010; van Dijk, 2016). Entre tanto, desde la psicolingüística se ha aportado evidencia de que existe un componente cognitivo en el prejuicio, al medir la relación entre dos elementos léxicos usados para describir a grupos minoritarios (White et. al. 2009; Wang et. al. 2011). En este tipo de estudios, uno de tales elementos es *priming* y se espera que un participante reaccione ante otro lexema que es *target*, los resultados muestran que cuando entre estos dos elementos hay una relación semántica incongruente con un estereotipo, el tiempo de reconocimiento es mayor.

Este estudio se suma a la perspectiva de los enfoques descritos. Para hacerlo, combina herramientas metodológicas de la lingüística computacional y la lingüística de corpus; de esta última, además toma la teoría de la activación léxica (Hoey, 2005) que se complementa con las nociones de subjetividad y deixis trabajadas desde la lingüística cognitiva. Este enfoque multidisciplinar obedece a la necesidad de probar si es posible hablar de un solo discurso xenofóbico en redes sociales. Esta hipótesis nace del concepto de discurso que aporta Fairclough (2010) esto es, discurso es aquel que se constituye a partir de la repetición o acumulación de formas semióticas (que pueden ser lingüísticas). Estas formas semióticas se toman aquí como repeticiones de ciertas codificaciones lingüísticas, más precisamente, activaciones léxicas presentes en asociaciones semánticas que se acumulan hasta construir un discurso intertextual observable en dos diferentes corpus de redes sociales. De este modo, la propuesta de análisis se articula, principalmente, en los principios metodológicos

de la lingüística de corpus que asigna la correspondencia entre elementos lingüísticos con un tipo de discurso solo una vez que se considera la medición de la frecuencia de un rasgo dentro de las fronteras de un corpus.

Así pues, lo que el lector encontrará en este capítulo será una discusión sobre lo oportuno de integrar estas diferentes perspectivas para comprobar la hipótesis. El análisis de sentimientos es la herramienta que se utilizó donde el papel del investigador es más reducido, es decir, es una herramienta automática. Como se sabe, las técnicas de la lingüística de corpus son más bien semiautomáticas e involucran fuertemente la presencia del juicio humano, más aún lo necesitan los estudios críticos y los estudios de la lingüística cognitiva. Por esto, interesa principalmente argumentar que el uso de una herramienta de la lingüística computacional, como lo es un análisis de sentimientos, para un estudio como este resultó funcional dadas las características de la comunicación que permite una de las redes sociales estudiadas, y que aportó una ventaja sobre la lista de palabras clave, pues desde un primer momento dirigió el análisis a cadenas de palabras. Una vez discutido lo anterior, se argüirá que todo este análisis multidisciplinar fue útil para encontrar un solo discurso xenofóbico y se presentan tres tipos de evidencia. No obstante, se hablará también de las limitaciones de tal implementación. Finalmente se discute qué significan nuestros hallazgos en relación con lo que ya había sido encontrado en materia de análisis de discursos sobre migración en estudios críticos desde la lingüística de corpus y cómo pueden ser incorporados nuestros hallazgos en futuras investigaciones.

8.1 Sinergía de metodologías de cómputos de la lengua

Aquí se sostiene que realizar un modelo de detección de sentimientos xenófobos es una técnica oportuna cuando se busca un análisis en redes sociales tanto por las características de las redes sociales, como por la necesidad de analizar un fenómeno que es intertextual. Como se verá, esta técnica puede guiar análisis cualitativos sobre elementos lingüísticos que ocurran en corpus diferentes de momentos históricos compartidos.

En este trabajo, el análisis intertextual se procuró a partir de la sistematización de datos producidos en comunicaciones reales de redes sociales. Tales datos están circunscritos a dos redes sociales y a discusiones que se desarrollaron en ellas sobre migración

centroamericana y caribeña en su paso o estadía por México sucedidas de octubre del 2018 a noviembre del 2020, periodo en que sucedieron caravanas migrantes masivas con un fuerte interés mediático.

Como ya hemos dicho, la primera etapa del procesamiento de los datos consistió en un análisis de sentimientos. Esta técnica, si bien constituye un producto en sí misma, se usó aquí como punto de partida de nuestra investigación. En tanto producto, se construyó un modelo sólido y competitivo en comparación con otros proyectos semejantes para tareas en español (Tabla 10. Mejores modelos en la revisión de literatura, p. 100). En tanto punto de partida, se comprobó que la lista de elementos con ganancias de información, obtenida del mejor modelo de clasificación logrado, es una herramienta útil para centrar el análisis en elementos lingüísticos específicos a modo de la tradicional lista de palabras clave de la lingüística de corpus.

Esto se debe a dos razones. La primera de ellas es que las condiciones de obtención de datos de un corpus como el de Twitter no hacen viable la posibilidad de uso de esta herramienta. La segunda está limitada al caso particular de esta investigación y se refiere a la ventaja que otorgó ante la posibilidad de trabajar con n-gramas como rasgos lingüísticos característicos del lingüístico xenofóbico dado que nuestro objetivo versaba sobre las asociaciones semánticas.

Así pues, en primer lugar consideramos un acierto iniciar un análisis de corpus para redes sociales desde un análisis de sentimientos dadas las preferencias de la lingüística de corpus por la construcción de corpus representativos con la finalidad de dar respuesta al comportamiento de ciertos rasgos lingüísticos en su contexto de uso. Así, la representatividad es entendida como la capacidad de una muestra de incluir el rango completo de variabilidad de una población (Biber, 1993). Por eso, para los principios metodológicos de esta disciplina es tan importante –en la medida de lo posible– contar con un número de documentos representativos del tipo de discurso que se esté estudiando, al tiempo que dichos documentamos contengan una ocurrencia representativa del rasgo de interés (Biber, 1993; Gabrielatos, 2007). Ambas consideraciones respaldarán que el corpus empleado sea representativo de la población a estudiar.

Obsérvese que se necesitan tres elementos para este análisis: el rasgo lingüístico, el documento o texto y, finalmente, el corpus. Tanto las formas de comunicación que hacen posible algunas redes sociales –entre las que se encuentra Twitter–, como las técnicas de obtención de datos de las mismas, dificultan la definición del documento o texto y con ello, incluso, la definición de la población a estudiar. En otras palabras, ¿cuál es la situación comunicativa en la que contextualizamos los mensajes xenofóbicos que obtuvimos de Twitter? Para YouTube, se extrajeron comentarios a videos noticiosos sobre migración, en ese sentido los mensajes xenofóbicos de esta red social son una respuesta, forman una conversación; el conjunto de respuestas a cada video es un documento, lo que posibilita que el rango de variabilidad de un rasgo sea medido dentro del documento y entre documentos, y más importante, que se cuente con la situación comunicativa completa en la que sucede el rasgo de interés.

Entre tanto, el corpus de tweets se compiló a través de una base de datos que no se construyó específicamente para esta investigación, y que recogió datos únicamente con un criterio temporal y otro geográfico, necesarios pero insuficientes, en tanto no son los criterios situacionales que fijan la comunicación en la que ocurrieron dichos tweets. De esta base, se buscaron tweets a partir de semillas, ello quiere decir que se excluyó todo aquello que no tuviera tales semillas, y que pudieron haber sido respuestas a un tweet sobre migración, o parte de un hilo sobre migración, por ejemplo.

Cabe mencionar que se ensayó una alternativa a esta base de datos. Se construyó un listener propio de este proyecto y se ehcó a andar en algunas ocasiones en que la discusión pública versaba sobre el tema de migración. Se experimentó con coordenadas para las frontera norte y sur del país, al haber detectado discusión sobre la estadía de migrantes en esas zonas del país; también se intentó con palabras clave o hashtags de la discusión que eran parte de los temas de moda del momento. De este modo, se extrajeron varios datos, y luego de revisar alrededor de 20 mil y, al no obtener ninguno sobre migración, se optó por la base de datos COVID (*COVID-19 México*, 2020).

Con los datos obtenidos de este modo, ¿representatividad de qué tipo de texto tenemos? ¿cómo le pueden hacer frente las herramientas de la lingüística de corpus a esta situación?

Esta disciplina puso sobre la mesa la importancia de la representatividad al momento de hacer análisis crítico del discurso, y, en la medida en que las redes sociales crezcan en relevancia como parte de las discusiones públicas, debe resolverse cómo abordar los estudios de estas discusiones sin dejar de lado la necesidad de construir bases de datos representativas.

Al contrario del planteamiento de la lingüística de corpus, que necesita la situación comunicativa completa, el planteamiento de un análisis de sentimientos es el de *aprender* a detectar mensajes, xenofóbicos en para el caso de este proyecto. Para los sistemas de aprendizaje supervisado, como se usó en este trabajo, basta un conjunto de datos considerados xenofóbicos y otro conjunto considerado de datos no xenofóbicos. Estos conjuntos son revisados por clasificadores que aprenderán, a grandes rasgos, la posibilidad de cada una de las características aparecidas en ambos conjuntos –o lo que para el caso de lingüística de corpus hemos llamado "rasgo"– de ser parte de un mensaje xenofóbico. Particularmente, el algoritmo Naïve Bayes Mulinomial, clasificador del mejor modelo de este trabajo (Tabla 9. Resultados de los experimentos de clasificación, p. 98), considera la frecuencia de una característica en un conjunto de datos de entrenamiento, y "castiga" a aquellos rasgos que sucedan menos. Mientras el clasificador decidirá, con base en el cálculo anterior, si tal mensaje pertenece a la clase de interés o no, la lista con ganancia de información asigna a todas las características encontradas en el data set el peso probabilístico de formar parte de la clase de interés.

Tomando lo anterior en cuenta, se cuenta con una herramienta que permitió contrarse en una serie de rasgos, particularmente n-gramas, sobre los cuales desarrollaríamos las etapas más cualitativas de la investigación. A modo de la lista de palabras clave, la lista con ganancia de información dio n-gramas clave de los cuales se revisó su ocurrencia entres corpus target (YouTube y Twitter). Se recuerda que el corpus de Twitter para esta etapa de la investigación no fue el mismo que se utilizó para el análisis de sentimientos. Si bien en aquella etapa se trató de balancear las clases para no afectar la probabilidad de los elementos de la case xenofóbica que ocurre menos (Tabla 4. Corpus usados por momentos de la investigación, p. 78), para hacer el análisis tanto de corpus, como el crítico, sí se debe mencionar que lo que

se considera xenofobia ocurre menos en Twitter de acuerdo a los datos presentados en relación con lo que no es xenofobia en una discusión pública sobre migración.

Ahora bien, la segunda razón por la que se evalua como un acierto la decisión de iniciar nuestro análisis de corpus con un análisis de sentimiento se ubica dentro del contexto de esta investigación. A diferencia de la lista de palabras clave utilizada en la lingüística de corpus, de los atributos empleados como clave resultó una lista de n-gramas, ya que el mejor modelo obtenido en el análisis de sentimientos utilizó una representación de características que incluía n-gramas de cuatro longitudes. Esto otorgó una ventaja sobre la lista de palabras clave, dado que nuestro interés radicó precisamente en la asociación semántica que involucra palabras contiguas. Así, con una lista de n-gramas clave, fue posible elegir tres cuyo valor semántico apunta a propiedades deícticas y, con ellas, a la subjetividad del hablante, a saber, *aquí en, nuestro país* y *en nuestro país*. En otras palabras, este punto de partida nos permitió perfilar las hipótesis del análisis hacia las valoraciones xenófobas que sobresalen en las redes sociales.

El objetivo de la lista de palabras clave en un estudio tradicional de la lingüística de corpus, cuando hace estudios críticos del discurso, es el de superar conclusiones impresionísticas según las cuales se la asignen rasgos lingüísticos –escogidos a priori por el investigador– a un discurso ideológico (McEnery & Hardie, 2012; Stubbs, 1997). El hecho de sustituir tal lista por la de los elementos con ganancia de información de un modelo que compitió con otros y que fue evaluado por cuatro métricas, cumple con tal demanda.

No obstante, debe mencionarse que el procedimiento de análisis de sentimientos se inició mediante una clasificación manual de los tuits objeto de estudio. Esto quiere decir que la decisión de lo que sería lenguaje xenófobo y no xenófobo quedó a cargo del ojo humano y podría objetarse la subjetividad de quien decidió. Al respecto, se recuerda que este paso es necesario en un proceso de aprendizaje supervisado en el cual, como su nombre lo indica, los algoritmos clasificadores “aprenden” las etiquetas a clasificar de un input que el investigador conoce. En el caso de esta investigación, se consideró fijar criterios para esta división, además de que el proceso no quedó a cargo de una sola persona sino que se contrastó la división de dos revisoras bajo el estadístico de kappa, mismo que mostró una

buena concordancia entre ambas divisiones (ver capítulo **¡Error! No se encuentra el origen de la referencia.¡Error! No se encuentra el origen de la referencia.**, p. **¡Error! Marcador no definido.**).

Por otro lado, es cierto que este trabajo pudo haberse realizado omitiendo, en este paso, la observación humana, pues existen alternativas para un análisis de sentimientos que no necesitan esta participación, es decir, mediante procesos de aprendizaje no supervisado (Dalal & Zaveri, 2011). De este modo, podría aprovecharse "la neutralidad" de los algoritmos. Ahora bien, la cuestión de la absoluta neutralidad en un trabajo, entre cuyos objetivos está el de ser crítico, no parece ni posible ni deseable, por lo que el punto a problematizar sería, más bien, en qué momentos de la exploración debería suceder la participación más subjetiva del investigador.

En ese sentido, otra opción podría ser la de aprovechar los resultados de otras investigaciones de enfoque lingüístico para construir diccionarios que busquen términos o incluso construcciones lingüísticas –usando etiquetadores sintácticos– previamente encontrados como xenofóbicos. Al respecto también está el hecho de que el análisis del discurso –hecho desde diferentes disciplinas, pero también desde diferentes corpus con diversas lenguas y referentes a variados contextos sociales– ha evidenciado que ciertos *tópicos* tienden a repetirse en los discursos xenofóbicos contra el ingreso de migrantes a los Estados Nación.

Un enfoque parecido al que proponemos lo hacen Plaza del Arco et al. (2020), quienes construyeron un lexicón con un enfoque mixto –basado tanto en diccionarios como en corpus– para obtener términos ofensivos en español. Los resultados de estos autores coinciden con los hallazgos de esta tesis respecto a la importancia de los n-gramas para detectar lenguaje de odio, pues en español se suele utilizar expresiones; así, para la clase xenófoba encontraron bigramas compuestos de adjetivos negativos antecediendo a sustantivos despectivos referentes a nacionalidades ("malditos sudacas", "putos moros", "putos sudacas", "malditos árabes"). Cabe mencionar, que el lexicón que construyeron los autores, tanto para el lenguaje xenófobo como para el misógino, que también es de interés en sus experimentos, está basado solo en uno de los campos semánticos que arrojó nuestra

lista de elementos con ganancia de información, esto es, *colectividad*. A propósito también se destaca que los n-gramas que se encontraron no son términos en sí ofensivos, salvo, muy probablemente, *negros* e *hijos de*.

Se debe hablar aquí de las posibilidades que da la sinergía entre lingüística computacional y lingüística de corpus. Ya se dijo que ante los nuevos formatos de comunicación donde la población objetivo sea difícilmente circunscrita a sus fronteras, las técnicas del PLN auxilian a modo de corpus piloto para lograr representatividad y las generalizaciones correspondientes. Pues bien, también la lingüística de corpus, en tanto pretende dar cuenta tanto del comportamiento de los rangos lingüísticos como de los corpus en sí (o situaciones de habla en sí), puede ayudar a mejorar las herramientas que el PLN usaría, por ejemplo, en un análisis de sentimientos. Una discusión más abundante al respecto se verá en el siguiente apartado donde se discuten los resultados que respaldan la hipótesis, así como los alcances explicativos de los resultados.

8.2 ¿Un solo discurso xenofóbico?

Previamente se expuso que la teoría de la activación éxica permite explicar la constitución de un discurso a fuerza de la repetición de expresiones con matices políticos (ver seciones **¡Error! No se encuentra el origen de la referencia. ¡Error! No se encuentra el origen de la referencia.**, p. **¡Error! Marcador no definido.**y **¡Error! No se encuentra el origen de la referencia. ¡Error! No se encuentra el origen de la referencia.**, p. **¡Error! Marcador no definido.**). Se explicó también que tanto tales expresiones, como la constitución del discurso, tienen como base la restricción que condiciona la elección léxica. En cuanto a las expresiones, la constante ocurrencia de dos elementos léxicos refuerza en el hablante las relaciones gramaticales, semánticas y pragmáticas que de esa combinación emanen. Ello quiere decir que el hablante encontrará natural y accesible la producción de tal ocurrencia, pero también que la esperará cuando el primero de tales elementos provenga de algún interlocutor. En cuanto a la constitución de un discurso vale recordar la noción de cohesión textual que comparte con el concepto de activación la expectación, esta vez a nivel textual; en otras palabras, la expectativa va más allá de la palabra inmediata y de la cláusula. Para Hoey, si la explicación de la activación léxica servía para explicar la cohesión textual, debía

servir también para un corpus intertextual; suposición que se exploró en este trabajo, además, en las expresiones ideológicas de un discurso ideológico común en un corpus intertextual.

A propósito de lo anterior, se presenta a continuación tres tipos de evidencia encontrada para argumentar que es posible hablar de un discurso xenofóbico en un corpus intertextual. La primera de las evidencias es que los elementos con ganancia de información para el lenguaje xenofóbico encontrados para el caso del corpus de Twitter, ocurren –en su mayoría– también en el de YouTube, con una proporción importante de acuerdo con su frecuencia relativa frente a la ocurrencia de un corpus de referencia. Ahora bien, debe advertirse que esta comparación de la ocurrencia de los n-gramas con ganancia de información dio como resultado tres tipos de n-gramas:

1) aquellos que son exclusivos de Twitter. Se trata de n-gramas que tienen algún lexema en común con las semillas con las que se hizo la búsqueda, es decir, que son muy particulares de los eventos que se recogieron al momento de hacer el levantamiento de datos.
    a) (ej. “aquí en Tapachula Chiapas”);
2) aquellos compartidos por ambas redes sociales . La gran mayoría de los n-gramas caen en este tipo; y la mayoría de las excepciones caen, por otro lado, en el punto anterior.
    a) (ej “aquí en”)
3) y por último, aquellos que si bien fueron relevantes dentro del conjunto de datos “xenofóbico” de la red social Twitter y tiene una ocurrencia alta en YouTube, no son representativas de ninguna red social pues ocurren “casi” las mismas veces en el corpus de referencia.
    a) (ej. “centro”)

Estas tres distinciones hablan del valor de la triangulación de los datos con diferentes herramientas. Al respecto, es interesante que se encontraran n-gramas del tipo 3, esto es, que no sean más relevantes en Twitter en comparación con el corpus de referencia. Considérese que estos n-gramas son un producto del análisis de sentimientos, para el cual

se alimentó al modelo con dos conjuntos de datos, ambos de un corpus balanceado de Twitter. Aunque el hecho más relevante es, como ya se dijo, que la mayoría de los n-gramas son representativos no solo de tal red social, sino también de YouTube. Obsérvese cómo este hecho muestra, por un lado, la utilidad de haber mezclado las herramientas de la lingüística computacional y la lingüística de corpus como de hecho se hizo, al tiempo que aporta evidencia en favor de la hipótesis de un discurso xenofóbico compartido entre usuarios de redes sociales. El hecho de haber utilizado un corpus de Twitter balanceado para el análisis de sentimientos, y luego otro de la misma red social, pero sin balancear, no afectó a ese hecho; pero pudo haber repercutido en la ocurrencia de n-grama del tipo 3.

La segunda evidencia a favor de nuestra hipótesis general es la ocurrencia de los n-gramas en contexto gramaticales y semánticos parecidos, o en otras palabras, la presencia de asociaciones semánticas comunes en ambas redes sociales. Como se mencionó, al estudiar n-gramas en contextos se encontraron las asociaciones semánticas MOVIMIENTO, PELIGRO y ACTITUD. En tanto asociaciones tienen las siguientes características:

1. Cada asociación está compuesta por diferentes conjuntos semánticos. Estos subconjuntos semánticos se deben, sobre todo, a las diferentes propiedades léxicas de los verbos con lo que co-ocurren nuestros n-gramas, y pueden consultarse en la **¡Error! No se encuentra el origen de la referencia.** para MOVIMIENTO, en la **¡Error! No se encuentra el origen de la referencia.**, para PELIGRO y en la **¡Error! No se encuentra el origen de la referencia.** para ACTITUD.
2. Entre conjuntos hay relaciones semánticas que las cohesionan, para empezar, como parte de la misma asociación. A grandes rasgos, se trata de relaciones de oposición entre Nosotros y los Otros, oposición marcada por la propiedad deíctica latente en los n-gramas de análisis. Al respecto, se observó que, en las asociaciones, algunos subconjuntos se activaron con ciertos referentes como sujetos gramatical, aquellos que podrían ser asignados a los mexicanos –a un Nosotros–, contra aquellos que podrían ser asignados a los migrantes en tanto Otros. Se subrayan un par de ejemplos: a) en el caso de MOVIMIENTO, observamos que el subconjunto *salida* está restringido completamente a suceder con sujetos gramaticales que refieren a

migrantes; b) sucede igual en el caso de PELIGRO, vemos que en el subconjunto *amenaza;* con la excepción de que se activa a su vez una codificación explícita del sujeto, referente de tal amenaza.

3. Las tres asociaciones tienden a anidarse entre sí, esta vez cohesionadas como parte de un discurso de xenofobia que gira en torno a una noción nacionalista del lugar que comparten mexicanos y migrantes. La *entrada*, por ejemplo, se percibe como *amenaza*, pero también la *entrada* se concibe como sujeta al merecimiento, merecimiento que está codificado también en la *asistencia* dando cuenta de que, en términos generales, los migrantes (y algunos mexicanos) son mentados en una relación ordinal para, valga la redundancia, ordenar quién merece las ventajas del territorio nacional, tanto el acceso a él como a sus recursos.

Adicionalmente, se encontró una tercera evidencia: la ocurrencia del fenómeno *priming pasivo*. Este fenómeno da cuenta del conocimiento de una postura ideológica xenófoba desde una perspectiva crítica, es decir, sin asumirla necesariamente como propia. El priming pasivo indica que el discurso está presente en el contexto cultural de los usuarios, aunque no lo reproduzcan de manera activa. Tal fenómeno evidencia que lo que se encuentra en las redes sociales es una conversación en torno a la migración en la que los hablantes pueden identificarse ideológicamente, pero también con respecto a su adscripción al territorio en disputa. En 160 tenemos ejemplos del conocimiento de los hablantes en torno a esa necesidad de justificar que una persona pueda salir de su país para entrar en otro, entre tanto, el hablante está fuera del territorio en disputa, a saber, México, y se sitúa más bien desde uno de los territorios ajenos. Desde ahí intenta establecer un diálogo de comprensión. Visto como un fenómeno intertextual, que da cabida a una conversación, el mensaje contra la migración de (132) podría verse como una respuesta a (131). En (133), entre tanto, un hablante se posiciona desde el Nosotros, aunque fuera de México, y reconoce esa noción de merecimiento pero desde una postura crítica.

(131) **aquí en *Honduras*** no hay trabajo y la corrupción es lo que más se mira aquí y lo peor es aquí en Honduras le suben a todo a los impuestos,la gasolina, al transporte, las medicinas,comida etc (Comentario YouTube)

(132) **Aquí en *México* ya tenemos problemas con nuestro presidente** pero nos quedamos en nuestro país a seguir luchando (ejemplo priming pasivo pero del discurso no xenofóbico)

(133) *Siempre he pensado que cuando sea grande voy a proteger a todas esas personas migrantes* porque en mi pais colombia hay muchos migrantes de venezuela que vienen a buscar una oportunidad y *anunque* en **nuestro pais** no todos ***los* aceptan**

Esta característica se debe a la intertextualidad, pero también a las limitaciones metodológicas que se enfrentaron cuando se fue posible prescindir en estos corpus de los comentarios de personas no mexicanas. Al respecto, vale advertir que en este trabajo no se profundizó sobre cómo la conversación en redes sociales puede afectar la cohesión del discurso de interés.

Con o sin análisis conversacional, estos datos contraponen dos perspectivas sobre un fenómeno. En este estudio fue posible detenerse en él gracias a la categoría de priming pasivo; en contraparte, los antecedentes que se reportaron lo buscaron al comparar corpus con diferentes tipologías textuales o diferentes actores sociales (sección 4.2.1 Punto de partida: construir un corpus, p. 57). Al respecto, se considera emblemático el trabajo de Galindo (2019), pues además es de los pocos estudios encontrados que trataran el análisis del discurso mediante un cómputo de datos sobre el discurso mexicano. Lo interesante de su trabajo es que versa sobre la migración mexicana hacia EE.UU. y que contrapone el discurso de prensa tanto de ese país como el de la prensa en México. La autora resaltó que en los periódicos estadounidenses sobresalen palabras como *delincuente, drogas, detenidos, violentos*, mismas que no están en la prensa mexicana donde más bien se encuentran términos como *paisanos, derechos, connacionales,* y *víctimas*. Estas diferencias, tanto en el trabajo de Galindo como en el que aquí se presenta, abren la interrogante de si estas diferencias en el discurso mexicano se deben a la tipología textual o al cambio de perspectiva en tanto esta vez quienes migran “somos nosotros”. Esta última hipótesis va en concordancia con la importancia de las propiedades deícticas de los n-gramas con los cuales se inició el análisis.

En resumen, gracias a estas tres evidencias es que se afirma que estamos ante la convergencia de opiniones de distintos emisores hacia una narrativa xenofóbica común. A pesar de lo que mensajes provienen de distintas cuentas y usuarios en ambas redes sociales, el análisis realizado muestra que estas emisiones individuales se entrelazan y refuerzan mutuamente para construir una narrativa cohesionada. Esta narrativa comparte los mismos elementos léxicos, las mismas asociaciones semánticas y el sentido de territorio nacional, lo que sugiere una ideología xenofobia subyacente compartida.

8.3 Discusión de los antecedentes y de caminos próximos

En el apartado anterior, se mencioba que esta investigación dio como resultado tres tipos de evidencia que sustentan la hipótesis de un discurso xenófobico cohesionado, común en las redes sociales al discutir el temas de las migraciones centroamericanas y caribeñas en México. No obstante, se considera que la coincidencia de estos resultados con lo observado en la revisión de literatura constituye un cuarto tipo de evidencia.

De acuerdo con las asociaciones encontradas en los corpus de redes sociales construídos para este proyecto, el discurso xenofóbico mexicano se puede articular en una noción nacionalista de territorio; en contraste, el nacionalismo también es un tópico que se ha encontrado en antecedentes, pero también sobresalen al respecto tópicos como ciudadanía e identidad (Bevitori, 2018; Galindo Gómez, 2019; Guerra Salas & Gómez Sánchez, 2017; Hartnett, 2019; Islentyeva, 2021; Lawson, 2015; Montali et al., 2013; O'Regan & Riordan, 2018; Taylor, 2009). Esta noción activa asociaciones de amenaza hacia lo que no es propio del territorio, en esta investigación; entre tanto ya se había encontrado la representación del ingreso de migrantes a un territorio nacional como invasión (Baker & McEnery, 2005; Camargo Fernández, 2021; Taylor, 2009, 2021), o incluso bajo metáforas que codifican agua o líquido (Ferreira et al., 2017; Salahshour, 2016; Turnbull, 2018), con la idea subyacente de que las migraciones son fenómenos incontrolables y, otra vez, amenazantes. A su vez, en repetidas ocasiones la entrada de personas migrantes a los países destino tienden a cuantificarse en grandes proporciones apoyando la representación anterior (Baker et al., 2008; Baker & McEnery, 2005; Fotopoulos & Kaimaklioti, 2016; Gabrielatos & Baker, 2008;

Salahshour, 2016; Taylor, 2009; Turnbull, 2018). Dadas las carácterísticas del fenómeno de la migración, es esperable encontrarse descripciones de los flujos migratorios y, de hecho, se han encontrado bajo el tópico movimiento (Baker et al., 2008; Baker & McEnery, 2005; Gabrielatos & Baker, 2008; Montali et al., 2013; Taylor, 2009), que en esta investigación se encuentra fuertemente asociado a prosodias semánticas negativas.

Dadas las coincidencias anteriores entre los trabajos de la revisión de literatura, y en concordancia con los requerimientos del ACD de situar socialmente al discurso (van Dijk, 2016), es que se sugiere una discusión de los antecedentes por las características sociohistóricas fijadas en los corpus de estudio. Es por lo menos interesante que a pesar del gran contraste entre el contexto histórico y social del que emana de estos corpus y el contexto de los antecedentes revisados, los resultados coincidan tanto. Además de las diferencias entre los contextos, hay diferencias también entre los actores y las tipologías textuales. Cabe preguntarse, entonces, ¿a qué obedece que en diferentes países, con diferentes gobiernos, y aun entre diferentes actores sociales, los discursos xenofóbicos coincidan?

En el momento en el que se inició la investigación, se decidió no revisar el discurso de las élites. Sin embargo, ante estas coincidencias se pone sobre la mesa una alternativa al planteamiento de este trabajo, esto es, la de rastrear el discursos de actores políticos mexicanos a modo de tercer corpus. Al respecto se advierte que, a la hora de extraer datos de la base de datos COVID de la UNAM, no se conservaron los datos de los usuarios, no obstante esta es una posibilidad para cuando es un dato relevante, como sería precisamente el de dar el peso de un actor de élite en una discusión pública.

Tanto el ACD como la lingüística de corpus subrayan la necesidad de contextualizar el discurso a estudiar. Este diseño, no obstante, requiere una relación entre investigación teórica, en la que se identifican parámetros para elegir textos, e investigación empírica, por ejemplo, al probar un corpus piloto como sugiere Biber (1993), o bien en aprovechar los resultados de las investigaciones para futuras decisiones en la creación de tales corpus. Precisamente, los resultados observados a partir del procesamiento de la lista de elementos con ganancia de información pueden continuarse en otras investigaciones.

Se mencioba que la comparación entre dos corpus target y uno de referencia de los elementos con ganancia de información dio como resultados tres tipos de n-gramas. La contraposición de los n-gramas tipo 1 y 2 sugieren la existencia de elementos comunes al discurso xenofóbico (tipo 2), así como la particularidad del discurso dependiendo de su tipología textual, del actor social que pronuncie el discurso y de las circunstancias sociohistóricas (tipo 1). Esta distinción recuerda a los conceptos de colocación estacional, y colocación consistente usada en la investigación diacrítica (H. Baker et al., 2017; P. Baker et al., 2008)(Baker et al. 2008, Baker et al. 2017).

Por lo tanto, una investigación futura podría plantearse precisamente indagar en qué medida estos n-gramas pueden ser entendidos como análogos a dichas clases de colocaciones tal como se han empleado en estudios diacríticos. En una investigación en este tenor, deberían considerarse variables tales como la tipología textual, el actor social que emita el discurso y las circunstancias sociohistóricas en donde se esperaría que se difieran los n-gramas tipo 1, mientras los n-gramas tipo 2 señalarían elementos comunes. Elementos que, de acuerdo a los visto en la sección **¡Error! No se encuentra el origen de la referencia. ¡Error! No se encuentra el origen de la referencia.** (p. **¡Error! Marcador no definido.**), deberían ser consistentes a lo largo del tiempo, e independiente de tipologías y actores.

Ahora bien, los hallazgos de este trabajo pueden ayudar a redireccionar experimentos de corte cuantitativo como los de los análisis de sentimientos. Al respecto se retoma el trabajo de Plaza del Arco et. al (2020), autores que hicieron un trabajo que profundizó en la importancia de dirigir las características que distinguen el discurso de odio. Sin embargo, a raíz de la comparación de la metodología de frente a los resultados de esta tesis se resalta que los autores construyeron su lexicón solamente con aquellas palabras que designaran despectivamente a los referentes de personas migrantes. Estos referentes están contenidos en el campo semántico *colectividad*, que se obtuvieron del procesamiento de los primeros 100 elementos de la lista de palabras claves y sugiere una hipótesis interesante en tanto una categoría del lenguaje de odio puede estar caracterizado por los insultos a su población objetivo; no obstante, el mismo análisis de la lista de elementos con ganancia de información sugieren otros campos semánticos –a saber *lugar, política* y *movimiento*– que

se concluyó que son parte del discurso xenofóbico en tanto se usan como recursos evaluadores. Con ello en mente, un análisis de sentimientos basado en lexicón podría extenderse al uso de tales campos semánticos.

En resumen, los resultados de esta investigación aportan evidencia sólida para argumentar la existencia de un discurso xenofóbico cohesionado y compartido entre usuarios de diferentes redes sociales al discutir temas de migración centroamericana y caribeña en México. Esto se sustenta en tres tipos de hallazgos: la recurrencia de n-gramas clave en ambos corpus de redes sociales, la presencia de asociaciones semánticas comunes que articulan una noción nacionalista del territorio, y la identificación de un fenómeno de priming pasivo que revela el carácter intertextual de dicho discurso. Adicionalmente, la coincidencia entre estos resultados y los de estudios previos sobre discursos anti-migratorios en diversos contextos constituye un cuarto tipo de evidencia que refuerza la propuesta de que nos encontramos ante una narrativa xenofóbica compartida, más allá de las diferencias contextuales.

Estos hallazgos abren nuevas vías de investigación, como el análisis diacrónico de la evolución de este discurso utilizando los conceptos de colocaciones estacionales y consistentes, así como la exploración de cómo el discurso de élites políticas puede haber influido en su configuración. Nuestros resultados pueden ayudar a redireccionar experimentos de corte cuantitativo como los análisis de sentimientos.

9 Conclusiones

En concordancia con el ACD, y con el objetivo general de este trabajo, esta investigación concluye que abordar el fenómeno de la xenofobia en los espacios digitales requiere una mirada multidimensional como la que aquí se utilizó. A grandes rasgos, se puede decir que la xenofobia que se encontró en dos redes sociales tiene un fuerte anclaje en procesos cognitivos, y aun emocionales, que condicionan la manera en que se percibe y se conceptualiza a grupos sociales, en este caso, a los migrantes. A manera de conclusión, en este capítulo se recuerda las fases que nos llevaron a esta conclusión y se reflexiona en torno a la necesidad de investigar los discursos de odio contra personas en redes sociales.

Anteriormente se mencioba que la discriminación y segregación social, de la que la xenofobia es solo una manifestación, había sido explicada desde varios enfoques entre los cuales nuestra investigación sería solo un aporte. A saber, el de la comprobación en evidencia intertextual de que las ideas en torno al hecho de la migración tienen un fundamente psicolingüístico. En ese sentido, el análisis de sentimientos y posteriormente de corpus ha revelado patrones y tendencias generales que reflejan esa postura de rechazo, hecho que sugiere la existencia de una activación léxica y semántica subyacente que condiciona y reproduce este tipo de discurso.

Para llegar a esta conclusión, como se ha dicho a lo largo de todo el documento, fue necesario un enfoque multidisciplinar. Este nació luego de sopesar las posibilidades explicativas de la teoría de la activación léxica (2005) para un estudio crítico del discurso sobre migración; de considerar las contribuciones que la lingüística de corpus ha hecho al ACD (Baker et al., 2008; Stubbs, 1996, 1997); así como de valorar las posibilidades que ofrecen las técnicas del PLN en un contexto social mediado cada vez más por comunicaciones en línea. Sobresalió, entonces, que ninguna de esas disciplinas se entiende en la individualidad del dato, sino en su constante repetición.

Fue así como se planteó una primera fase que mirara el discurso desde su cara más general. Es decir, aquella que permitió mirar cuantitativamente la existencia de patrones lingüísticos. En esta primera etap, fue oportuno realizar un análisis de sentimientos, y se construyó un detector de tweets xenofóbicos. Este modelo mostró ser competitivo en relación con otros

experimentos con tareas semejantes para datos en español. Posteriormente, se utilizó el enfoque de la activación léxica que proporcionó un marco teórico, en combinación con la semántica cognitiva, para comprender cómo los discursos ideológicos se manifiestan a través de la elección y activación de ciertas asociaciones semánticas.

En esta investigación, en contraste con algunos de los antecedentes revisados, se reconoció el papel de las redes sociales en la discusión pública en torno a las migraciones centroamericanas y caribeñas en su paso o estadía por México. En tanto las redes sociales ganen adeptos como plataformas de comunicación, se vuelve relevante investigar en qué medida son usadas para amplificar y multiplicar discursos, ideologías u opiniones.

Relacionado con lo anterior, a la luz de las conclusiones de esta investigación, pero también de otras que han mostrado que las redes sociales son cámaras de eco de lenguaje de odio, vale cuestionarse las posibilidades democráticas de estas plataformas. Una de las conclusiones de este trabajo es que es tan necesaria la existencia de marcos legales que permitan la diversidad de voces y opiniones que estas tecnologías hacen posibles, como apremiante la necesidad de comprender cómo se reproducen los discursos ideológicos en estos contextos. Incluso vale la pena preguntar cómo influyen en las redes sociales dinámicas políticas que incluyen en las interacciones en línea.

En tanto México es un país receptor de población migrante, este trabajo es un aporte para comprender el debate entre la aceptación y el rechazo de la población migrante por los mexicanos. En la medida en que estos discursos influyan en la forma en que los migrantes son percibidos, reconocidos y tratados es que resulta crucial analizar y comprender los diferentes discursos. Con estos análisis pueden encontrarse patrones, tendencias y construcciones discursivas que promuevan la xenofobía o la inclusión.

Cabe destacar que, al momento de redactar estas líneas, México se ostenta como uno de los países más violentos del mundo (Human Rights Watch, 2023). Para los migrantes que ingresan por la frontera sur de nuestro país, esto se traduce en quedar en medio de una situación legal incierta, por un lado, o de suma violencia, por otro. En cuanto a su situación frente al Estado Mexicano es necesario aclarar que, durante su gobierno, el presidente López Obrador movilizó más de 31 mil soldados para control migratorio, alcanzando la cifra

más alta en la historia de detención contra migrantes (Guillén, 2023; Human Rights Watch, 2023); se han establecido puntos de revisión y control migratorio que, a pesar de haber sido declarados como inconstitucionales en 2022, siguen operando; las condiciones que atraviesan los migrantes en instituciones estatales son de hacinamiento e insalubridad, muestra de ello es el incendio en el que murieron 40 personas migrantes en un centro de detención migratorio de Ciudad Juárez (Human Rights Watch, 2023). En medio de este contexto es que se necesitan trabajos que cuestionen los discursos que justifican las situaciones de riesgo y vulnerabilidad que las personas extranjeras puedan tener al entrar al país.

No obstante, aquí se consideró más importante cuestionarse la lógica de la legitimidad de estos discursos. Se planteó para ello un marco teórico según el cual la segregación social y el discurso que la legitime o promueva tendrían ambos sus orígenes en la cognición, es decir, a procesos de los que todos somos susceptibles. Se propousó un marco metodológico multidisciplinar que indague esa cognición en dos corpus intertextuales.

El aporte central de esta investigación radica en vincular discursos xenófobos con la coocurrencia de determinados elementos lingüísticos –inicialmente léxicos– que el hablante reproduce de manera naturalizada y que el oyente también espera encontrar, ¿y cómo se desmantela la xenofobia si cree natural? Precisamente, creo que abordar esta investigación como un esfuerzo por ampliar la mirada multidisciplinar que sugiere que segregar socialmente es el resultado de procesos que involucran tanto nuestra cognición como nuestra dimensión emocional avanza hacia dicho objetivo.

10 Referencias

Anón. 2020a. «COVID-19 México». Recuperado 29 de abril de 2024 (http://www.miopers.unam.mx/covid/).

Anón. 2020b. «Digital: Mexico». *DataReportal – Global Digital Insights*. Recuperado 29 de abril de 2024 (https://datareportal.com/reports/digital-2020-mexico).

Anón. 2021. «DETOXIS- IberLEF». *DETOXIS-IberLEF 2021*. Recuperado 28 de abril de 2024 (https://detoxisiberlef.wixsite.com/website).

Anón. 2023. «HUHU». *IberLEF 2023*. Recuperado 28 de abril de 2024 (https://sites.google.com/view/huhuatiberlef23/huhu).

Anthony, Laurence. 2022. «AntConc (v.4.0.3).»

Arcila Calderón, Carlos, David Blanco-Herrero, y María Belén Valdez Apolo. 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*. doi: 10.5477/cis/reis.172.21.

ASALE, RAE-, y RAE. s. f. «centro | Diccionario de la lengua española». *«Diccionario de la lengua española» - Edición del Tricentenario*. Recuperado 20 de julio de 2023 (https://dle.rae.es/centro).

Baker, Helen, Tony McEnery, y Andrew Hardie. 2017. «A Corpus-based Investigation into English Representations of Turks and Ottomans in the Early Modern Period» editado por M. Pace-Sigge y K. J. Patterson. *Lexical Priming: Applications and Advances* 79(79):41-66. doi: 10.1075/scl.79.

Baker, Paul. 2004. «Querying Keywords: Questions of Difference, Frequency, and Sense in Keywords Analysis». *Journal of English Linguistics* 32(4):346-59. doi: 10.1177/0075424204269894.

Baker, Paul, Costas Gabrielatos, Majid Khosravinik, Micha\l Krzyżanowski, Tony McEnery, y Ruth Wodak. 2008. «A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press». *Discourse & society* 19(3):273-306.

Baker, Paul, y Tony McEnery. 2005. «A Corpus-Based Approach to Discourses of Refugees and Asylum Seekers in UN and Newspaper Texts». *Journal of Language and Politics* 4(2):197-226. doi: 10.1075/jlp.4.2.04bak.

Basile, Valerio, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco

Manuel Rangel Pardo, Paolo Rosso, y Manuela Sanguinetti. 2019. «SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter». Pp. 54-63 en *Proceedings of the 13th International Workshop on Semantic Evaluation*. Minneapolis, Minnesota, USA: Association for Computational Linguistics.

BBC News Mundo. 2019. «Trump anuncia aranceles de un 5% para todas las importaciones desde México "hasta que se resuelva el problema de la inmigración ilegal"». *BBC News Mundo*.

Bengfort, Benjamin, Rebecca Bilbro, y Tony Ojeda. 2018. *Applied text analysis with python: Enabling language-aware data products with machine learning*. O'Reilly Media, Inc.

Bevitori, Cinzia. 2018. «Crossing Boundaries: Investigating "Fair" in British Parliamentary Debates on Im/migration». *Textus* (1/2018). doi: 10.7370/89450.

Biber, Douglas. 1993. «Representativeness in Corpus Design». *Literary and Linguistic Computing* 8(Literary and Linguistic Computing).

Bird, Steven, Ewan Klein, y Edward Loper. 2009. *Natural Language Processing with Python*. editado por J. Steele.

Bryson, Bill. 2010. *Neither Here, Nor There: Travels in Europe*. Vol. 11. Random House.

Butler, Judith, y Nancy Fraser. 2016. *¿Redistribución o reconocimiento?: un debate entre marxismo y feminismo*. Madrid: Traficantes de Sueños.

Camargo Fernández, Laura. 2021. «El nuevo orden discursivo de la extrema derecha española: de la deshumanización a los bulos en un corpus de tuits de Vox sobre la inmigración». *Cultura, Lenguaje y Representación* 26:63-82. doi: 10.6035/clr.5866.

Campos-Domínguez, Eva. 2017. «Twitter y la comunicación política». *El Profesional de la Información* 26(5):785. doi: 10.3145/epi.2017.sep.01.

Caribe, Comisión Económica para América Latina y el. 2019. *Desarrollo y migración: desafíos y oportunidades en los países del norte de Centroamérica*. CEPAL.

Claridge, Claudia. 2007. «Constructing a corpus from the web: message boards». Pp. 87-108 en *Corpus linguistics and the web*. Brill.

Corpora & applied linguistics, dir. 2022. *Corpus linguistics and applied linguistics research 2022 Dr Charlotte Taylor 9 November*.

Correa-Cabrera, Guadalupe. 2014. «Seguridad y migración en las fronteras de México: diagnóstico y recomendaciones de política y cooperación regional». *Migración y desarrollo* 12(22):147-71.

Dalal, Mita K., y Mukesh A. Zaveri. 2011. «Automatic Text Classification: A Technical Review». *International Journal of Computer Applications* 28(2):37-40. doi: 10.5120/3358-4633.

Damiris, Niklas, y Helga Wild. 1997. «The Internet: A New Agora?» Pp. 307-17 en *An Ethical Global Information Society*, editado por J. Berleur y D. Whitehouse. Boston, MA: Springer US.

van Dijk, Teun A. 2016. «Análisis Crítico del Discurso». *Revista Austral de Ciencias Sociales*.

Dobrić Basaneže, Katja, y Paulina Ostojić. 2021. «Migration Discourse in Croatian News Media». *Medijska Istraživanja* 27(1):5-27. doi: 10.22572/mi.27.1.1.

Fairclough, N. 2003. *Analysing discourse: textual analysis for social research*. Routledge.

Fairclough, Norman. 2010. «A dialectical–relational approach to critical discourse analysis in social research». en *Critical Discourse Analysis*. Routledge.

Ferreira, Luciane C., Catarina Valle Flister, y Cassio Morosini. 2017. «The representation of refuge and migration in the online media in Brazil and abroad: a Cognitive Linguistics analysis». *Signo* 42(75):59-66. doi: 10.17058/signo.v42i75.11217.

Fiske, Susan T., Karen E. Rosenblum, y Toni-Michelle C. Travis. 2009. *Social beings: A core motives approach to social psychology*. Wiley New York.

Foster, Derek. 1996. «Community and Identity in the Electronic Village» editado por D. Porter. *Internet Culture* 23-37.

Fotopoulos, Stergios, y Margarita Kaimaklioti. 2016. «Media Discourse on the Refugee Crisis: On What Have the Greek, German and British Press Focused?» *European View* 15(2):265-79. doi: 10.1007/s12290-016-0407-5.

Fowler, R. 1991. *Language in the News: Discourse and Ideology in the Press*. London; New York: Routledge.

Frank, Eibe, Mark Hall, y Ian Witten. 2016. *The WEKA Workbench. Online Appendix for «Data Mining: Practical Machine Learning Tools and Techniques»*. Fourth Edition. Morgan Kaufmann.

Fraser, Nancy. 2000. «De la redistribución al reconocimiento. Dilemas de la justicia en la era postsocialista». *New left review* 1:126-55.

Gabrielatos, Costas. 2007. «Selecting Query Terms to Build a Specialised Corpus from a Restricted-Access Database». *ICAME Journal* (31).

Gabrielatos, Costas, y Paul Baker. 2008. «Fleeing, Sneaking, Flooding: A Corpus Analysis of

Discursive Constructions of Refugees and Asylum Seekers in the UK Press, 1996-2005». *Journal of English Linguistics* 36(1):5-38. doi: 10.1177/0075424207311247.

Galindo Gómez, Sandra Eugenia Galindo. 2019. «Las palabras importan: representación de los inmigrantes mexicanos en periódicos de México y Estados Unidos». *Migraciones Internacionales* (10):14.

Géron, Aurélien. 2022. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media, Inc.

Guerra Salas, Luis, y Ma Elena Gómez Sánchez. 2017. «La cobertura de las migraciones en la prensa de los países hispanohablantes (2016)». *Revista Nebrija de Lingüística Aplicada a la Enseñanza de Lenguas*.

Grabe, William, y Martin Phillips. 1987. «Aspects of Text Structure: An Investigation of the Lexical Organisation of Text». *Language* 63(1):200. doi: 10.2307/415427.

Guillén, Beatriz. 2023. «López Obrador reconoce un aumento del 31% de la migración irregular en la frontera de México y EE UU |». El País México.

Harris, Lasana, y Susana Fiske. 2006. «Dehumanizing the Lowest of the Low: Neuroimaging Responses to Extreme Out-Groups». *Psychological Science*, 847-53.

Hartnett, Sabina. 2019. «Willkommenskultur: A Computational and Socio-Linguistic Study of Modern German Discourse on Migrant Populations». *Transit* 12(1). doi: 10.5070/T7121043491.

Hoey, Michael. 2005. *Lexical Priming: A New Theory of Words and Language*. London; New York: Routledge/AHRB.

Hoey, Michael. 2017. «Cohesion and Coherence in a Content-specific Corpus» editado por M. Pace-Sigge y K. J. Patterson. *Lexical Priming: Applications and Advances* 79:3-40. doi: 10.1075/scl.79.

Honeybone, Patrick. 2005. «J.R. Firth». Pp. 80-86 en *Key Thinkers in Linguistics and the Philosophy of Language*, editado por C. Routledge. Edinburgh: University Press.

Huerta, Amarela Varela, y Lisa McLean. 2019. «Caravanas de migrantes en México: nueva forma de autodefensa y transmigración». *Revista CIDOB d' Afers Internacionals* 163-85. doi: 10.24241/rcai.2019.122.2.163.

Human Rights Watch. 2023. «México: Eventos de 2023». en *Informe Mundial 2024*.

Ibarretxe-Antuñano, Iraide, y Javier Valenzuela. 2012. *Lingüística cognitiva*. Barcelona: Anthropos.

IOM UN Migration. 2022. *Monitoreo de flujos migratorios en Tapachula y Tenosique,*

Ronda 2 (Abril 2022).

Islentyeva, Anna. 2021. *Corpus-Based Analysis of Ideological Bias: Migration in the British Press*. 1.ª ed. New York: Editorial Panel: IVACS.

Johnson, M. 1997. «Embodied meaning and cognitive science». Pp. 148-75 en *Language beyond Postmodernism: Saying and Thinking in Gendlin's Philosophy*, editado por D. M. Levin. Chicago: Northwestern University Press.

Johnson, Mark. 1990. *The Body in the Mind: The Bodily Basis of Meaning, Imagination, and Reason*. Chicago, IL: University of Chicago Press.

KhosraviNik, Majid. 2009. «The Representation of Refugees, Asylum Seekers and Immigrants in British Newspapers during the Balkan Conflict (1999) and the British General Election (2005)». *Discourse & Society* 20(4):477-98. doi: 10.1177/0957926509104024.

Lawson, Michelle. 2015. «Life in the Ghetto: How the Media Represent British Lifestyle Migration to France». *JOMEC Journal* 0(7). doi: 10.18573/j.2015.10003.

Londoño Zapata, Oscar Iván. 2013. *Discurso en sociedad. Entrevista a Teun A. van Dijk*. Ibagué, Colombia: Ediciones Unibagué.

Maldonado, Ricardo. 2013. «Niveles de subjetividad en la deixis. El caso de aquí y acá1». *Anuario de Letras. Lingüística y Filología* 1(2):283-326. doi: 10.1016/S0185-1373(13)70258-1.

Marcoccia, Michel. 2004. «On-line polylogues: conversation structure and participation framework in internet newsgroups». *Journal of pragmatics* 36(1):115-45.

Massey, Douglas S. 2008. «La racialización de los mexicanos en estados unidos: estratificación racial en la teoría y en la práctica». *Migración y Desarrollo* (10):65-95.

McEnery, Tony, y Andrew Hardie. 2012. *Corpus linguistics: method, theory and practice*. Cambridge ; New York: Cambridge University Press.

Merino García, Rafael dir. 2012. *6 - 6 - Multinomial Naive Bayes - Un ejemplo resuelto .mp4*.

Montali, Lorenzo, Paolo Riva, Alessandra Frigerio, y Silvia Mele. 2013. «The Representation of Migrants in the Italian Press: A Study on the *Corriere Della Sera* (1992–2009)». *Journal of Language and Politics* 12(2):226-50. doi: 10.1075/jlp.12.2.04mon.

Müller, Enrique. 2015. «Alemania facilita la llegada de refugiados sirios a su territorio». *El País*.

O'Regan, Veronica, y Elaine Riordan. 2018. «Comparing the Representation of Refugees, Asylum Seekers and Migrants in the Irish and UK Press: A Corpus-Based Critical Discourse Analysis». *Journal of Language and Politics* 17(6):744-68. doi: 10.1075/jlp.17043.ore.

Ortega Ramírez, Adriana Sletza, Luis Miguel Morales Gámez, Adriana Sletza Ortega Ramírez, y Luis Miguel Morales Gámez. 2021. «(In)seguridad, derechos y migración. La Guardia Nacional en operativos migratorios en México». *Revista IUS* 15(47):157-82. doi: 10.35487/rius.v15i47.2021.699.

Partington, Alan. 1998. «Patterns and Meanings. Using Corpora for English Language Research and Teaching». *International Journal of Corpus Linguistics* 6(1). doi: 10.1075/ijcl.6.1.08rom.

de Paula, Angel Felipe Magnossão, y Ipek Baris Schlicht. 2021. «AI-UPV at IberLEF-2021 DETOXIS Task: Toxicity Detection in Immigration-Related Web News Comments Using Transformers and Statistical Models». doi: 10.48550/arXiv.2111.04530.

Pérez Paredes, Pascual, Pilar Aguado Jiménez, y Purificación Sánchez Hernández. 2016. «Constructing immigrants in UK legislation and Administration informative texts: A corpus-driven study (2007–2011)». *Discourse & Society* 28. doi: https://doi.org/10.1177/0957926516676700.

Pineda, Perla. 2019. «Despliegue de la Guardia Nacional en la frontera sur inicia el lunes: Marcelo Ebrard». *El Economista*.

Pitropakis, Nikolaos, Kamil Kokot, Dimitra Gkatzia, Robert Ludwiniak, Alexios Mylonas, y Miltiadis Kandias. 2020. «Monitoring Users' Behavior: Anti-Immigration Speech Detection on Twitter». *Machine Learning and Knowledge Extraction* 2(3):192-215. doi: 10.3390/make2030011.

Plaza-Del-Arco, Flor-Miriam, M. Dolores Molina-González, L. Alfonso Ureña-López, y M. Teresa Martín-Valdivia. 2020. «Detecting Misogyny and Xenophobia in Spanish Tweets Using Language Technologies». *ACM Transactions on Internet Technology* 20(2):12:1-12:19. doi: 10.1145/3369869.

Poletto, Fabio, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, y Viviana Patti. 2021. «Resources and Benchmark Corpora for Hate Speech Detection: A Systematic Review». *Language Resources and Evaluation* 55(2):477-523. doi: 10.1007/s10579-020-09502-8.

Reyes Vázquez, Juan Francisco, y María Inés Barrios de la O. 2019. «El comportamiento de los usuarios de Twitter respecto al tema de la Caravana Migrante a través del Sentiment Analysis (SA), 2019».

Rico-Sulayes, Antonio. 2018. *Authorship Attribution on Crime-Related Social Media:*

Research on the darknet in forensic linguistics. Aracne Editrice. Roma, Italia.

Roman, Victor. 2019. «Algoritmos Naive Bayes: Fudamentos e Implementación». *Ciencia y Datos*. Recuperado 16 de noviembre de 2022 (https://medium.com/datos-y-ciencia/algoritmos-naive-bayes-fudamentos-e-implementaci%C3%B3n-4bcb24b307f).

Romero-Vega, Raúl R., Oscar M. Cumbicus-Pineda, Ruperto A. López-Lapo, y Lisset A. Neyra-Romero. 2021. «Detecting Xenophobic Hate Speech in Spanish Tweets Against Venezuelan Immigrants in Ecuador Using Natural Language Processing». Pp. 312-26 en *Applied Technologies*, *Communications in Computer and Information Science*, editado por M. Botto-Tobar, S. Montes León, O. Camacho, D. Chávez, P. Torres-Carrión, y M. Zambrano Vizuete. Cham: Springer International Publishing.

Salahshour, Neda. 2016. «Liquid Metaphors as Positive Evaluations: A Corpus-Assisted Discourse Analysis of the Representation of Migrants in a Daily New Zealand Newspaper». *Discourse, Context & Media* 13:73-81. doi: 10.1016/j.dcm.2016.07.002.

Sanguinetti, Manuela, Fabio Poletto, Cristina Bosco, Viviana Patti, y Marco Stranisci. 2018. «An Italian Twitter Corpus of Hate Speech against Immigrants».

Schrötei, Melani, Marie Veniard, Charlotte Taylor, y Andreas Blätte. 2019. «A Comparative Analysis of the Keyword Multicultural(Ism) in French, British, German and Italian Migration Discourse». 60.

Sedano, Mercedes. 1994. «Evaluation of Two Hypotheses about the Alternation between Aquí and Acá in a Corpus of Present-Day Spanish». *Language Variation and Change* 6(2):223-37. doi: 10.1017/S0954394500001654.

Semple, Kirk, y Paulina Villegas. 2019. «México aprueba una Guardia Nacional de sesenta mil elementos que, según sus críticos, es más de lo mismo». *The New York Times*.

Sinclair, John. 2004. *Trust the Text: Language, Corpus and Discourse*. 0 ed. editado por R. Carter. Routledge.

Solan, Lawrence M., y Peter M. Tiersma. 2005. *Speaking of Crime: The Language of Criminal Justice*. Chicago, IL: University of Chicago Press.

Stubbs, Michael. 1996. *Text and Corpus Analysis*. Oxford: Blackwell.

Stubbs, Michael. 1997. «Whorf's children: Critical comments on critical discourse analysis (CDA)». *British studies in applied linguistics* 12:100-116.

Stubbs, Michael. 2001. «Computer-assisted Text and Corpus Analysis: Lexical Cohesion and Communicative Competence». en *The Handbook of Discourse Analysis*. Oxford:

Blackwell Publishers.

Tanner, Eliza. 2001. «Chilean Conversations: Internet Forum Participants Debate Augusto Pinochet's Detention». *Journal of Communication* 51(2):383-403. doi: 10.1111/j.1460-2466.2001.tb02886.x.

Taylor, Charlotte. 2009. «The Representation of the Inmigrants in the Italian Press». *UNIVERSITÀ DEGLI STUDI DI SIENA DIPARTIMENTO DI SCIENZE STORICHE, GIURIDICHE, POLITICHE E SOCIALI*.

Taylor, Charlotte. 2014. «Investigating the Representation of Migrants in the UK and Italian Press: A Cross-Linguistic Corpus-Assisted Discourse Analysis». *International Journal of Corpus Linguistics* 19(3):368-400. doi: 10.1075/ijcl.19.3.03tay.

Taylor, Charlotte. 2021. «Metaphors of Migration over Time». *Discourse & Society* 32(4):463-81. doi: 10.1177/0957926521992156.

Taylor, Charlotte, y Anna Marchi. 2018. *Corpus Approaches to Discourse: A Critical Review*. Abingdon New York: Routledge.

Turnbull, Judith. 2018. «Migration – What Is in a Word».

Vasquez, Augusto Cortez, Luzmila Pro Concepción, Oswaldo Rojas Lazo, y Roberto CalmetAgnelli. 2013. «Categorización de Textos mediante Máquinas de Soporte Vectorial».

Vázquez Meneley, Sergio. 2019. «Entre el discurso y la realidad, análisis sobre la cooperación migratoria entre México y el Triángulo Norte de Centroamérica» editado por A. C. Cabrera García, G. Rodríguez Albor, y I. Blanco Rangel. *Migraciones internacionales en el siglo XXI: un análisis desde una perspectiva crítica* 175-203.

Vindell, Juan José. 2021. «Kappa de Cohen en R». Recuperado 3 de octubre de 2022 (https://rpubs.com/VINDELL2981/kappa).

Wang, Lei, Qingguo Ma, Zhaofeng Song, Yisi Shi, Yi Wang, y Lydia Pfotenhauer. 2011. «N400 and the activation of prejudice against rural migrant workers in China». *Brain research* 1375:103-10.

Waseem, Zeerak, y Dirk Hovy. 2016. «Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter». Pp. 88-93 en *Proceedings of the NAACL Student Research Workshop*. San Diego, California: Association for Computational Linguistics.

White, Katherine R., Stephen L. Crites Jr, Jennifer H. Taylor, y Guadalupe Corral. 2009. «Wait, what? Assessing stereotype incongruities using the N400 ERP component».

Social Cognitive and Affective Neuroscience 4(2):191-98.

11 Apéndices

11.1 Lista de elementos con ganancia de información

Características extraídas: 1-4 gramas

ID	PESO	ATRIBUTO	LONGITUD DEL N-GRAMA
1	0.17014	nuestro país	2
2	0.16279	vienen	1
3	0.15872	aquí en	2
4	0.15872	quieren	1
5	0.15872	porque los	2
6	0.15428	cubanos	1
7	0.15428	realidad	1
8	0.14937	porque no	2
9	0.14937	río	1
10	0.14937	problema	1
11	0.14937	deberían	1
12	0.14937	pinches	1
13	0.14937	@m_ebrard	1
14	0.14937	#QuedateEnCasa	1
15	0.14937	son migrantes	2
16	0.14937	migrantes son	2
17	0.14937	calle	1
18	0.14937	suelo	1
19	0.14937	mexicanos y	2
20	0.14937	cerrar	1
21	0.14937	nosotros	1
22	0.1438	permitir	1
23	0.1438	décima norte	2
24	0.1438	elemento de la	3
25	0.1438	décima	1

26	0.1438	@SEDENAmx	1
27	0.1438	la décima	2
28	0.1438	aquí en Tapachula	3
29	0.1438	la realidad	2
30	0.1438	los mexicanos y	3
31	0.1438	la décima norte	3
32	0.1438	de nuestro país	3
33	0.1438	son unos	2
34	0.1438	urgente	1
35	0.1438	intento	1
36	0.1438	instalaciones	1
37	0.1438	elemento de	2
38	0.1438	inmigrantes y	2
39	0.1438	amotinan	1
40	0.1438	puente	1
41	0.1438	Y a	2
42	0.1438	elemento	1
43	0.1438	aquí en Tapachula Chiapas	4
44	0.1438	cerrar fronteras	2
45	0.1438	a la frontera	3
46	0.1438	centro	1
47	0.1438	nación	1
48	0.1438	entrar a	2
49	0.1438	leyes	1
50	0.1438	nadie	1

51	0.1438	dejen	1
52	0.1438	negros	1
53	0.1438	a la guardia	3
54	0.1438	a inmigrantes	2
55	0.1438	hijos de	2
56	0.1438	a nuestro	2
57	0.1438	no les	2
58	0.1438	en Tapachula Chiapas	3
59	0.1438	culpa	1
60	0.13723	un elemento de la	4
61	0.13723	USA y	2
62	0.13723	de manera	2
63	0.13723	un elemento	2
64	0.13723	palos	1
65	0.13723	resuelven su	2
66	0.13723	@lopezdoriga	1
67	0.13723	de la guardia nacional	4
68	0.13723	un elemento de	3
69	0.13723	le de	2
70	0.13723	de migrantes cubanos	3
71	0.13723	debería	1

72	0.13723	los migrantes porque	3
73	0.13723	de la guardia	3
74	0.13723	para que no	3
75	0.13723	los de	2
76	0.13723	trabajan	1
77	0.13723	balas	1
78	0.13723	anda	1
79	0.13723	y nosotros	2
80	0.13723	tipo	1
81	0.13723	tiempo	1
82	0.13723	pero la	2
83	0.13723	llegaron a	2
84	0.13723	que se les	3
85	0.13723	su gente	2
86	0.13723	tal de	2
87	0.13723	agredan	1
88	0.13723	sistemas	1
89	0.13723	@M_OlgaSCordero @SSalud_mx	2
90	0.13723	plena	1
91	0.13723	se larguen	2
92	0.13723	migrantes porque	2
93	0.13723	sociales	1
94	0.13723	dar	1
95	0.13723	cuidado	1
96	0.13723	se quedan	2
97	0.13723	@HLGatell	1
98	0.13723	no hacen	2
99	0.13723	no están	2
100	0.13723	podido	1
101	0.13723	poder cruzar	2

102	0.13723	a la guardia nacional	4
103	0.13723	se larguen a	3
104	0.13723	agredan a	2
105	0.13723	masiva	1
106	0.13723	su situación migratoria	3
107	0.13723	estos migrantes	2
108	0.13723	por qué	2
109	0.13723	suelo mexicano	2
110	0.13723	migrantes cubanos	2
111	0.13723	de la chingada	3
112	0.13723	sólo	1
113	0.13723	las leyes	2
114	0.13723	@SSalud_mx	1
115	0.13723	riqueza	1
116	0.13723	resuelven	1
117	0.13723	migrantes a la guardia	4
118	0.13723	migrante hondureño	2
119	0.13723	son migrantes son	3
120	0.13723	frenar	1
121	0.13723	con tal de	3
122	0.13723	las próximas	2
123	0.13723	las balas	2
124	0.13723	chingada	1

125	0.13723	intento de	2
126	0.13723	a un elemento de	4
127	0.13723	atención a los	3
128	0.13723	hacer algo	2
129	0.13723	igual	1
130	0.13723	Quiero	1
131	0.13723	usar	1
132	0.13723	hora de	2
133	0.13723	y que no	3
134	0.13723	quedan	1
135	0.13723	ello	1
136	0.13723	ayer en	2
137	0.13723	vacunas	1
138	0.13723	próximas	1
139	0.13723	río Suchiate	2
140	0.13723	el puente	2
141	0.13723	gente en	2
142	0.13723	aquí en México	3
143	0.13723	y aquí	2
144	0.13723	No es	2
145	0.13723	Piedras	1
146	0.13723	imigrantes	1
147	0.13723	Parece	1
148	0.13723	en campaña	2
149	0.13723	inmigrantes ilegales	2
150	0.13723	Piedras Negras	2
151	0.13723	incendio	1
152	0.13723	No son	2
153	0.13723	la chingada	2
154	0.13723	ladrones	1
155	0.13723	que las	2

156	0.13723	desmadre	1
157	0.13723	emigración	1
158	0.13723	descontrolada	1
159	0.13723	larguen	1
160	0.13723	Muchos	1
161	0.13723	contra México	2
162	0.13723	las autoridades mexicanas	3
163	0.13723	larguen a	2
164	0.13723	Se amotinan	2
165	0.13723	países de	2
166	0.13723	Estoy	1
167	0.13723	en nuestro país	3
168	0.13723	a un elemento	3
169	0.13723	vándalos	1
170	0.13723	unas	1
171	0.13723	dicen	1
172	0.13723	Negras	1
173	0.13723	en suelo mexicano	3
174	0.13723	en suelo	2
175	0.12463	no se	2
176	0.11138	Tulum	1
177	0.11138	Caravana	1
178	0.10905	estos	1
179	0.10796	Victoria	1
180	0.10767	en las	2
181	0.10767	gente	1
182	0.10614	ElUniversal https	2
183	0.10614	viajaban	1
184	0.10614	ElUniversal	1

185	0.10614	ElUniversal https //t	3
186	0.10614	vía ElUniversal https	3
187	0.10614	vía ElUniversal	2
188	0.10614	vía ElUniversal https //t	4
189	0.10224	los migrantes en	3
190	0.10224	personas migrantes	2
191	0.10224	llegar a	2
192	0.10224	llegar	1
193	0.10012	Caravana migrante	2
194	0.10012	municipio de	2
195	0.10012	salió	1
196	0.09786	mujer salvadoreña	2
197	0.09786	migrante salvadoreña	2
198	0.09786	en el municipio	3
199	0.09786	en esta	2
200	0.09786	a los migrantes en	4
201	0.09544	de Tulum	2
202	0.09544	Una	1
203	0.09544	legal	1

204	0.09544	asesinato de	2
205	0.09544	asesinato	1
206	0.09544	sábado	1
207	0.09544	el municipio de	3
208	0.09544	en el municipio de	4
209	0.09544	viajaban hacinados	2
210	0.09544	agentes de	2
211	0.09544	tres	1
212	0.09544	en Tulum	2
213	0.09368	de inmigrantes	2
214	0.09368	delincuentes	1
215	0.09279	la mujer salvadoreña	3
216	0.09279	y niños	2
217	0.09279	otra	1
218	0.09279	#Tulum	1
219	0.09279	Ciudad	1
220	0.09279	la mujer	2
221	0.09279	crisis	1
222	0.09279	cultura	1
223	0.09279	pena	1
224	0.09279	de origen hondureño	3
225	0.09279	trato	1
226	0.09279	detener	1
227	0.09279	Salazar	1
228	0.09279	2021	1
229	0.09279	sur de	2

230	0.09279	Elementos	1
231	0.09279	camioneta	1
232	0.09279	vez	1
233	0.09279	origen hondureño	2
234	0.09279	25	1
235	0.09279	mujer migrante	2
236	0.08985	avance	1
237	0.08985	tiempos	1
238	0.08985	comunidad	1
239	0.08985	carretera	1
240	0.08985	temporal	1
241	0.08985	#CDMX	1
242	0.08985	condiciones	1
243	0.08985	frontera de #EaglePass	3
244	0.08985	1	1
245	0.08985	migrantes viajaban	2
246	0.08985	23	1
247	0.08985	de #EaglePass	2
248	0.08985	este sábado	2
249	0.08985	en la frontera de	4
250	0.08985	hace unos	2
251	0.08985	migrante de origen hondureño	4
252	0.08985	migrante de origen	3
253	0.08985	#ÚLTIMAHORA	1
254	0.08985	la violencia	2
255	0.08985	#EaglePass	1

256	0.08985	la frontera de #EaglePass	4
257	0.08985	la crisis	2
258	0.08985	mató	1
259	0.08985	Oaxaca	1
260	0.08985	a mujeres	2
261	0.08985	viajaban en	2
262	0.08985	agentes de la	3
263	0.08985	Elementos de	2
264	0.08985	El presidente	2
265	0.08985	Migratoria	1
266	0.08985	que salió	2
267	0.08985	de Trump	2
268	0.08985	digno	1
269	0.08985	salvadoreña madre	2
270	0.08652	alimentos	1
271	0.08652	a mujeres y niños	4
272	0.08652	asesinada en	2
273	0.08652	a otro	2
274	0.08652	visa humanitaria	2
275	0.08652	visa	1
276	0.08652	abrazo	1
277	0.08652	Victoria Salazar	2
278	0.08652	agua y	2
279	0.08652	a mujeres y	3
280	0.08652	por 60 días	3
281	0.08652	una camioneta	2

282	0.08652	municipales	1
283	0.08652	murió	1
284	0.08652	mujeres y niños	3
285	0.08652	estancia legal	2
286	0.08652	mujeres y	2
287	0.08652	estatales	1
288	0.08652	más de 200	3
289	0.08652	tenía	1
290	0.08652	padre	1
291	0.08652	peligro	1
292	0.08652	en frontera	2
293	0.08652	niña	1
294	0.08652	niños y	2
295	0.08652	frente al	2
296	0.08652	fueron colocados	2
297	0.08652	migrantes viajaban hacinados	3
298	0.08652	manos de	2
299	0.08652	la Casa del Migrante	4
300	0.08652	la localidad	2
301	0.08652	migrantes Casa del Migrante	4
302	0.08652	localidad	1
303	0.08652	junto a	2
304	0.08652	junto	1
305	0.08652	mes	1
306	0.08652	mexicanos en	2
307	0.08652	migrantes Casa del	3

308	0.08652	migrantes Casa	2
309	0.08652	migrante salvadoreña madre	3
310	0.08652	el caso de	3
311	0.08652	el caso	2
312	0.08652	el asesinato de	3
313	0.08652	de llegar	2
314	0.08652	de Victoria	2
315	0.08652	de llegar a	3
316	0.08652	proteger	1
317	0.08652	que viajaban	2
318	0.08652	sentido	1
319	0.08652	común	1
320	0.08652	su estancia	2
321	0.08652	su estancia legal	3
322	0.08652	cerca	1
323	0.08652	síntomas	1
324	0.08652	colocados	1
325	0.08652	de policías	2
326	0.08652	del 2021	2
327	0.08652	el asesinato	2
328	0.08652	policías de	2
329	0.08652	por 60	2
330	0.08652	durante	1
331	0.08652	primera	1
332	0.08652	pobre y	2
333	0.08652	detectó	1
334	0.08652	por la pandemia	3
335	0.08652	por lo que	3
336	0.08652	por policías	2

337	0.08652	primaria	1
338	0.08652	presidente @lopezobrador_	2
339	0.08652	por ser	2
340	0.08652	Saltillo	1
341	0.08652	legal en	2
342	0.08652	S	1
343	0.08652	18	1
344	0.08652	Estación Migratoria	2
345	0.08652	#migración	1
346	0.08652	#justiciaparavictoria	1
347	0.08652	60 días	2
348	0.08652	#Huixtla	1
349	0.08652	Los migrantes viajaban	3
350	0.08652	Estación	1
351	0.08652	CDMX	1
352	0.08652	Esta	1
353	0.08652	#Tijuana	1
354	0.08652	La caravana	2
355	0.08652	#Camargo	1
356	0.08652	Fue	1
357	0.08549	tenemos	1
358	0.08175	y no	2
359	0.07641	policías	1
360	0.07622	ilegales	1
361	0.07622	por los	2
362	0.07622	entrar	1
363	0.07622	posible	1
364	0.07622	a su país	3
365	0.07622	si no	2
366	0.07622	vamos	1
367	0.07622	vamos a	2
368	0.07551	nuestro	1

369	0.07519	la guardia	2
370	0.07146	pero	1
371	0.06839	https //t	2
372	0.06839	https	1
373	0.06813	la guardia nacional	3
374	0.06786	//t	1
375	0.0659	a su	2
376	0.06205	México https //t	3
377	0.06205	México https	2
378	0.06158	aquí	1
379	0.06068	salvadoreña	1
380	0.05721	vía	1
381	0.05555	los inmigrantes	2
382	0.05496	-	1
383	0.05414	mujer	1
384	0.05366	guardia	1
385	0.05206	son	1
386	0.05073	no	1
387	0.04882	mil	1
388	0.04882	quien	1
389	0.0444	tiene	1
390	0.04259	niños	1
391	0.04019	de México	2
392	0.03917	inmigrantes	1
393	0.03795	así	1
394	0.03768	si	1
395	0.03028	los	1
396	0.02969	que no	2
397	0.02693	se	1
398	0.02668	años	1
399	0.02533	es	1

400	0.02232	El	1
401	0.02038	migrante	1
402	0.01988	en el	2
403	0.01937	los migrantes	2
404	0.01862	país	1
405	0.00885	en	1