



REPORTE DE CONVERSACIÓN DE MIGRACIÓN Y XENOFOBIA

GUATEMALA

1 DE ABRIL 2020 A 1 DE ABRIL 2021



#XENOFOBIAZERO
ALZA TU VOZ CONTRA LA DISCRIMINACIÓN



ÍNDICE

1. INTRODUCCIÓN.....	3
2. DATOS Y METODOLOGÍA.....	4
3. LA CONVERSACIÓN DE XENOFOBIA	4
3.1 EVENTOS MÁS RELEVANTES EN LA CONVERSACIÓN DE XENOFOBIA	4
4. CLASIFICACIÓN DE LA CONVERSACIÓN DE XENOFOBIA	9
4.1 MODELO LDA	9
4.2 RESULTADOS	10
5. CUANTIFICACIÓN DE XENOFOBIA POR NACIONALIDAD	12
6. CONCLUSIONES Y RECOMENDACIONES DE POLÍTICA	13
ANEXO 1. NOTICIAS Y XENOFOBIA	14
ANEXO 2. ESTIMACIÓN A PARTIR DE LDA	15

1. INTRODUCCIÓN

El presente informe muestra un análisis de la conversación en línea de xenofobia y lenguaje de odio hacia la población migrante en Guatemala. Para realizarlo, se construyó una base de datos que recoge todas las publicaciones de xenofobia y lenguaje de odio en línea hacia la población migrante hondureña, salvadoreña, venezolana y nicaragüense entre el 1 de abril de 2020 y el 1 de abril de 2021. Durante este periodo se generaron un total de 1.559 mensajes de este tipo que tuvieron un alcance de alrededor de 20.427.664 usuarios. Esto incluye retweets y compartidos.

El mes con mayor cantidad de publicaciones con contenido xenófobo fue mayo. Sin embargo, no hubo un evento o conversación específica que agrupara este tipo de publicaciones. La conversación con contenido xenófobo estuvo centrada principalmente en contra de personas provenientes de El Salvador, apuntando a la condición de estos como migrantes ilegales. El primer hallazgo es que la cantidad de comentarios de xenofobia en línea es relativamente baja en comparación a otros países de la región. Al separar las publicaciones de xenofobia por la nacionalidad a que iban dirigidas, se encuentra que el 47% estaban orientadas a población migrante en general, 22% a migrantes salvadoreños, 20% a población nicaragüense, 11% a población venezolana y 2% a población hondureña.

Durante los brotes más importantes de xenofobia durante el período analizado, las publicaciones con contenido de xenofobia se agruparon principalmente en las caravanas de migrantes provenientes de El Salvador y de Honduras. Por otra parte, también rechazaron que las personas migrantes opinaran sobre eventos políticos del país y se hacen llamados explícitos a que vuelvan a su país de origen.

Para estudiar a profundidad los distintos componentes de la conversación de xenofobia, se utilizó un modelo de clasificación de Machine Learning (LDA). Este permite clasificar las principales temáticas dentro de toda las publicaciones de xenofobia. Los resultados indican que hubo dos temáticas principales que se diferenciaron del resto. 1. El rechazo de la participación política de las personas migrantes para hacerse partícipe del debate político en Guatemala. 2. Los insultos y la responsabilización de la población hondureña, venezolana y salvadoreña sobre la situación política de sus países.

Por último, a partir de los hallazgos de este informe se sugieren las siguientes recomendaciones de política:

- 1.Complementar este análisis con otras fuentes de información. La cantidad de comentarios de xenofobia en línea es relativamente baja en comparación a otros países de la región y esto puede responder a diferentes fenómenos.
- 2.Considerar que una idea frecuente detrás de los mensajes de xenofobia mostraban rechazo hacia el hecho de que las personas migrantes opinaran sobre eventos políticos en el país.

El resto del informe sigue de la siguiente manera: En la sección 2 se describen los datos y la metodología utilizada para procesarlos. En la sección 3 se presentan los eventos más importantes en la conversación de xenofobia durante el periodo. En la sección 4 se clasifican y analizan los tópicos dentro de la conversación de xenofobia. En la sección 5 se estudia la nacionalidad a la que van dirigidas las publicaciones de xenofobia. En la sección 6 se concluye y se proponen unas recomendaciones de política.

2. DATOS Y METODOLOGÍA

Los datos utilizados en este informe provienen de la información pública de las redes sociales Facebook, Twitter, Instagram y Youtube en Guatemala entre abril de 2020 - abril de 2021. Para obtener y filtrar los datos se realizó un ejercicio de Web Scraping a través de la plataforma Meltwater. En este sentido, se organizaron y filtraron todos los posts públicos asociados a migración en Guatemala. La organización se llevó a cabo a través de una selección de palabras claves, frases y asociaciones que permiten capturar la conversación de migración. Estas palabras y frases clave captan las conversaciones sobre distintas poblaciones migrantes y fueron construidas en conjunto con el equipo local de la OIM en el país y pueden encontrarse en el Anexo 1. En el caso de Guatemala, las poblaciones migrantes que fueron incluidas en el análisis fueron las poblaciones venezolana, nicaragüense, hondureña y salvadoreña por ser las mayores comunidades de migrantes en el país.

Una vez capturada y delimitada la conversación de migración, el segundo paso fue cualificar los mensajes para poder identificar las publicaciones de discriminación hacia las poblaciones migrantes. Para esto, se capturó la conversación de discriminación a partir de palabras, frases y expresiones que mostraban una actitud xenófoba hacia la población migrante. Estas palabras, frases y expresiones también fueron trabajadas con el equipo local de OIM. Además, para asegurarse de capturar solo publicaciones de discriminación, se eliminaron todas las publicaciones que denunciaban el comportamiento xenófobo. Los datos finales corresponden a la totalidad de publicaciones en línea que muestran actitudes de discriminación, rechazo o lenguaje de odio hacia alguna de las poblaciones migrantes en Guatemala durante el período de estudio. En las siguientes secciones se procede a analizar los contenidos de estas publicaciones.

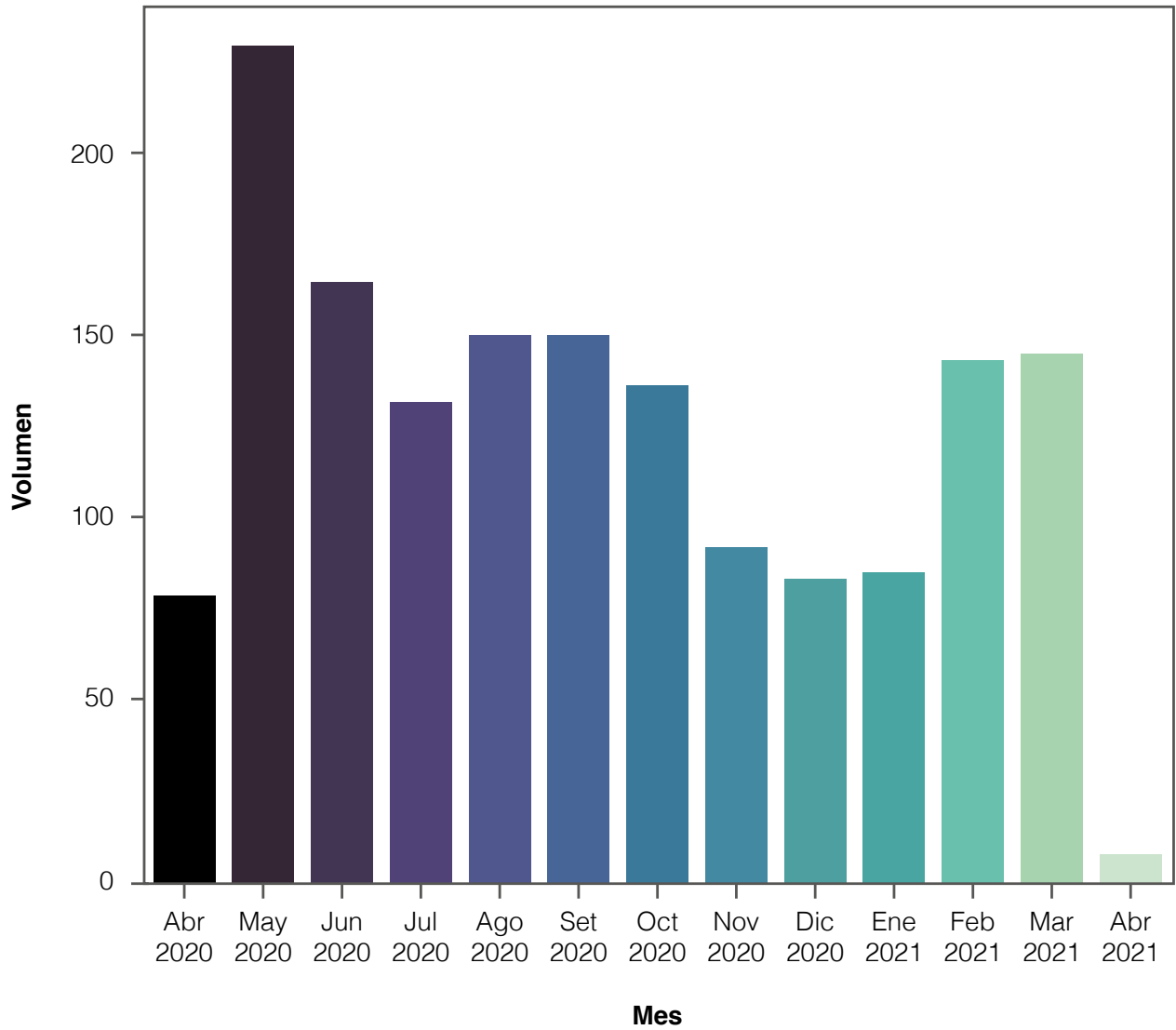
3. LA CONVERSACIÓN DE XENOFOBIA

En esta sección se estudia el comportamiento de la conversación de xenofobia en línea entre el 1 de abril de 2020 y el 1 de abril de 2021. Durante este periodo se generaron un total de 1.596 mensajes de xenofobia en línea. Con el propósito de explicar este comportamiento, primero se discuten los principales eventos que incidieron en esta conversación; segundo, se estudian los tópicos más relevantes dentro de esta conversación; y tercero, se discute la composición por nacionalidad dentro de la conversación de xenofobia.

3.1 EVENTOS MÁS RELEVANTES EN LA CONVERSACIÓN DE XENOFOBIA

Una forma de entender los impulsores de la publicaciones de xenofobia es a partir de los eventos en donde más se generaron ese tipo de publicaciones. Durante el año de estudio, se encontró que hay tres meses que contienen publicaciones de xenofobia por encima del promedio del período. El gráfico 3.1 muestra que los meses con mayor incidencia de xenofobia en línea fueron mayo, junio y agosto de 2020. A continuación se hace una descripción de la conversación de xenofobia durante estos meses.

Gráfico 3.1 Volumen de Publicaciones de Xenofobia por Mes.



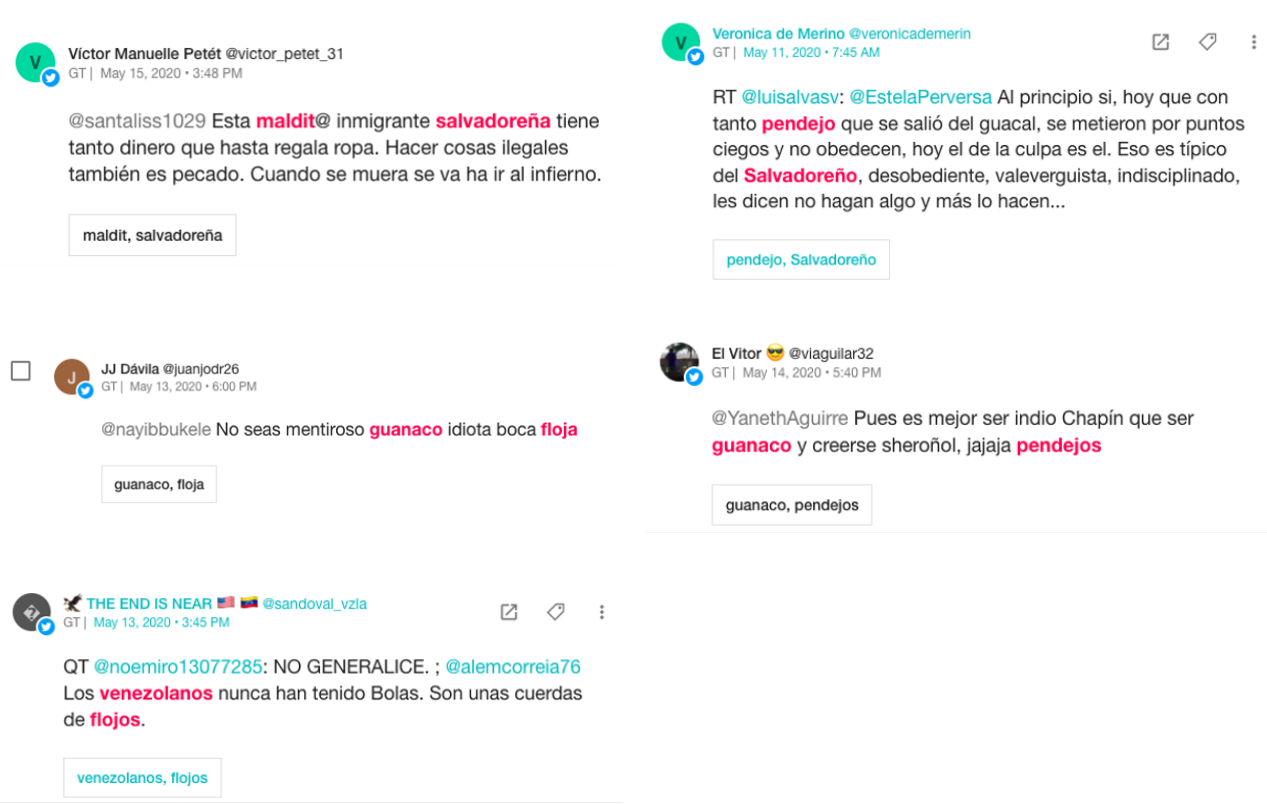
Meltwater (Construcción Propia), 2021.

Mayo 2020

Durante el mes de mayo se registró el mayor número de posts con contenido xenofóbico, alcanzando el total de 229 menciones. Esta conversación estuvo centrada principalmente en contra de migrantes provenientes de El Salvador. Un ejemplo de esto puede observarse en el gráfico 3.2, en el primer, segundo, cuarto y quinto mensaje, que van dirigidos a esta población. Estas publicaciones van acompañadas de adjetivos negativos, en donde se califica a las personas salvadoreñas como flojas e indisciplinadas.

Si bien hubo menciones a otras nacionalidades, como puede apreciarse en el tercer mensaje, el cual relaciona a personas venezolanas con la flojera, la mayor cantidad de mensajes con contenido xenofóbico estaban dirigidas a la población proveniente de El Salvador.

Gráfico 3.2 Ejemplo Mensajes Mayo de 2020

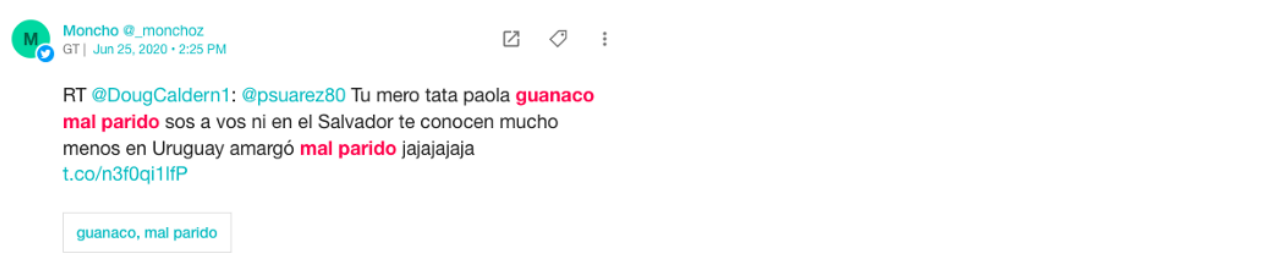


Junio 2020

A diferencia del mes de mayo, en junio puede apreciarse que los mensajes con contenido xenofóbo se abren a más nacionalidades, sin tener un sesgo tan fuerte hacia la población de El Salvador, como se vio anteriormente. Además, se encontraron publicaciones que ligan a personas nicaragüenses con hechos delictivos, como puede observarse en la tercera imagen del gráfico 3.3.

Por otra parte, se aprecia que hay dos mensajes que expresan de manera explícita la expulsión de migrantes de Guatemala, como está manifestado en el segundo y cuarto mensaje del gráfico. El segundo se refiere a una de las tantas caravanas de migrantes provenientes de Honduras que buscaban llegar a Estados Unidos. El otro post alude a una persona de origen nicaragüense.

Gráfico 3.3 Ejemplo Mensajes Junio de 2020

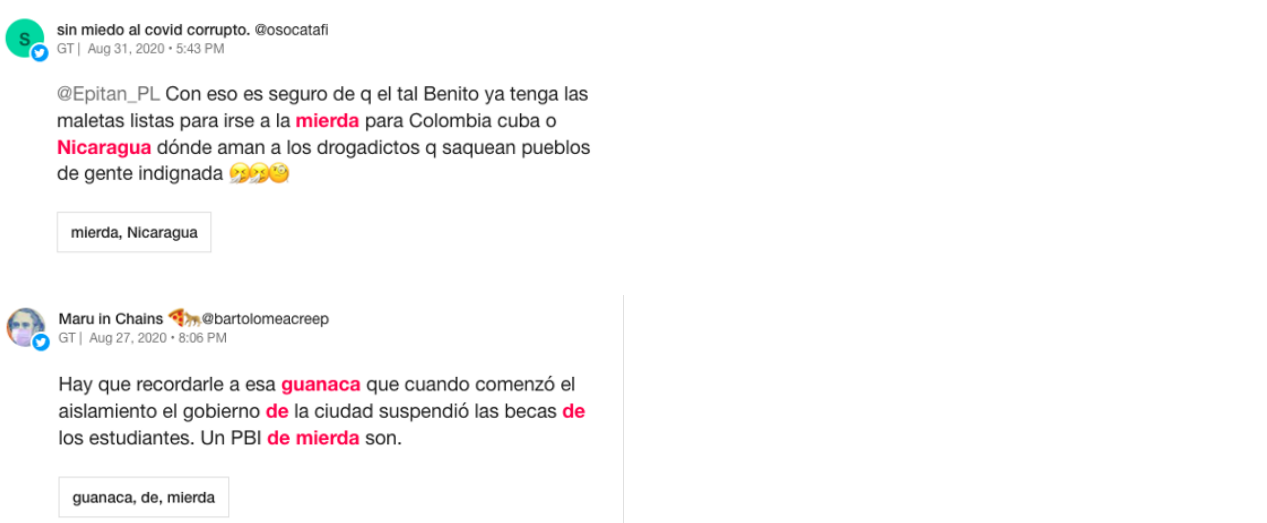


Agosto 2020

Las publicaciones de este mes fueron variadas en cuanto a las nacionalidades a las que se dirigieron los mensajes con una línea discursiva xenofóbica. En primer lugar, el segundo, cuarto y quinto mensaje del gráfico 3.4 se relacionan con publicaciones insultando a las personas de El Salvador. Estos posts van acompañados de insultos y apodos peyorativos hacia la población salvadoreña.

Por otra parte, también hubo publicaciones hacia otras nacionalidades, como puede apreciarse en el primer y tercer mensaje de la gráfica. En estos se llama a que ciertas personas se vayan “para la mierda” a otros países, como Colombia, Cuba o Nicaragua. En la tercera publicación se hace alusión a las personas de Perú y de Venezuela.

Gráfico 3.4 Ejemplo Mensajes de Xenofobia Agosto de 2020



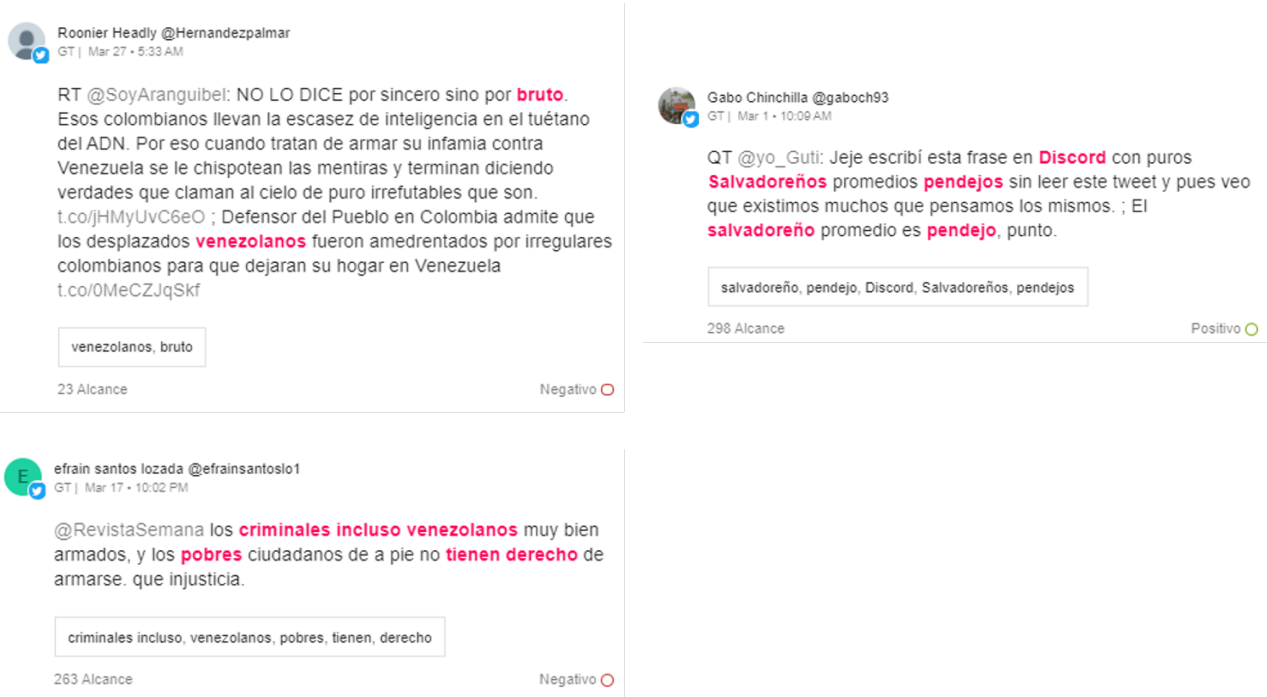


4.2 RESULTADOS

Los resultados de la metodología propuesta sugieren que la conversación de xenofobia estuvo organizada en cuatro tópicos diferentes. El primer tópico estuvo relacionado con generalizaciones y rechazo a la población migrante. El segundo tópico tuvo que ver con temáticas políticas que relacionaban a las personas migrantes con sus respectivos contextos, manifestando el rechazo de su participación en discusiones políticas y pidiendo su expulsión. El tercer y cuarto tópico no comprenden más del 5% de los términos usados en la conversación capturada, esto implica que no son conversaciones frecuentes. El gráfico 4.6 del anexo muestra la forma en que se distribuyen los tópicos dentro de esta conversación.

El tópico que más agrupó publicaciones de xenofobia (Rechazo y Ofensas a Migrantes) tuvo que ver con ofensas generales dirigidas a la población migrante de El Salvador, Nicaragua y Venezuela. Los insultos que fueron utilizados para referirse a estas poblaciones fueron “pobre”, “maldito”, “pendejo”, “puta” y “ladrón” u otros sinónimos. La mayoría de estas publicaciones mostraban un sentimiento de rechazo motivado por estereotipos y generalizaciones sobre la población migrante. En el gráfico 4.7 se pueden observar algunos mensajes que comprende este tópico.

Gráfico 4.7: Mensajes dentro del tópico 1



El segundo tópico que más agrupó publicaciones consiste en mensajes con connotación política expresados al interior de la conversación de xenofobia (tópico 2). Un ejemplo de esto son los mensajes que comparan la situación política del país respecto al caso venezolano y al caso salvadoreño, y como las personas migrantes u otras nacionalidades no deberían de opinar sobre la situación política del país al que migran. Entre otras cosas estos mensajes también piden la expulsión de la población migrante. En el gráfico 4.8 se pueden observar las principales publicaciones que contiene el tópico en cuestión.

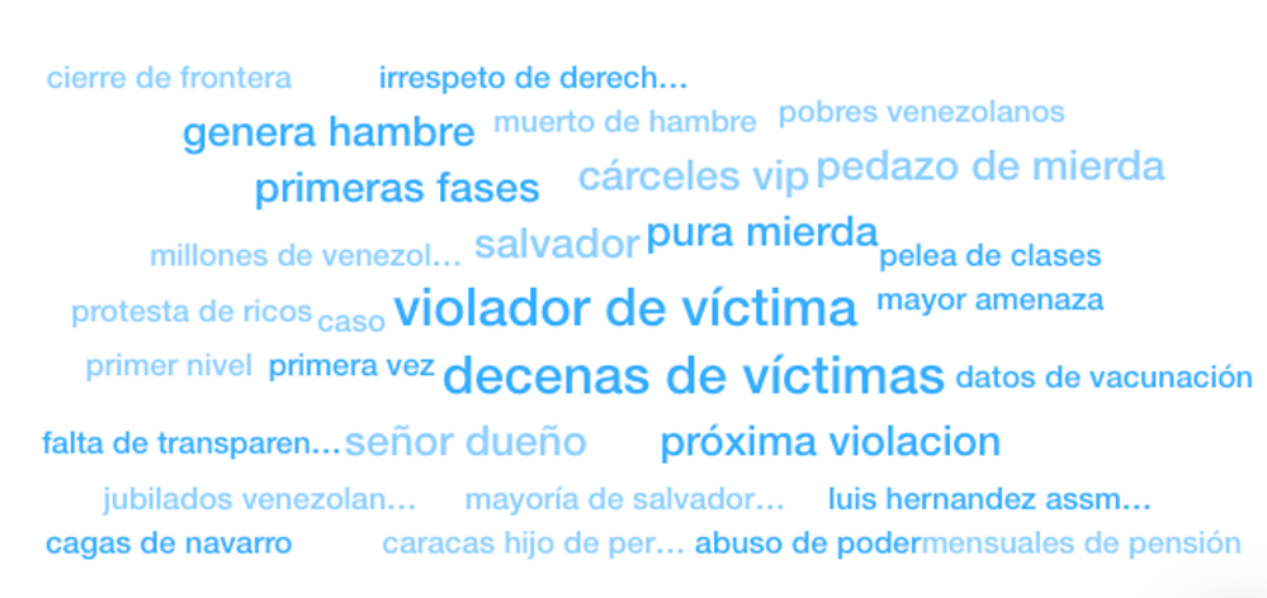
Gráfico 4.8: Mensajes dentro del tópico 2



Palabras y frases más frecuentes en las publicaciones de xenofobia

Además de los tópicos dentro de la conversación de xenofobia, un aspecto interesante a detallar es la forma que tienen los comentarios de xenofobia. Es decir, si existen algunas expresiones o palabras que sean usadas recurretemente dentro del lenguaje de odio y discriminación. El gráfico 4.9 es una nube de palabras que clasifica las palabras según la frecuencia con la que fueron usadas a lo largo del periodo de análisis. Entre mayor sea el tamaño de la frase o palabra dentro de la nube, fue más alto el número de veces que se usó dentro de la conversación.

Gráfico 4.9 Nube de palabras de la Conversación de Xenofobia



Meltwater, 2021.

Las frases más utilizadas fueron “violador de víctimas”, “decenas de víctimas” y “pura mierda”. Un hecho interesante es que gran parte de los mensajes supone que las poblaciones migrantes hacia Guatemala son víctimas de un régimen político turbulento, los mensajes se encargan de responsabilizar a las personas migrantes por estos hechos y rechazan su participación y opinión política respecto a Guatemala.

Otras palabras situadas en las periferias de la nube mencionan temas habituales de administración pública como la “lucha de clases”, “genera hambre”, “mayoría de salvador” y “abuso de poder”, varias de estas

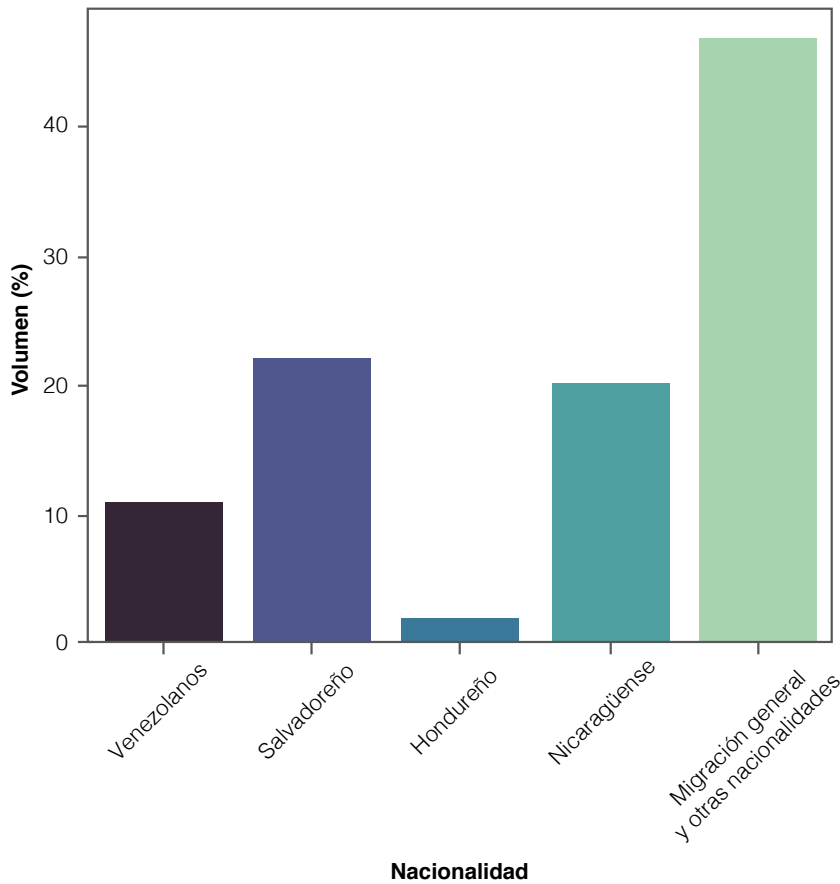
publicaciones relacionan a las personas migrantes con problemáticas básicas de Guatemala o temas de seguridad y violencia, a pesar de que estos mensajes son los que menos se concentran en la nube, muestra cómo la llegada de nuevos migrantes es percibida por los guatemaltecos.

5. CUANTIFICACIÓN DE XENOFOBIA POR NACIONALIDAD

Por último, un elemento importante en el análisis de la conversación de xenofobia en línea es entender la población a la que van dirigidos. En específico, dentro de Guatemala existen migrantes de distintas nacionalidades, siendo las poblaciones migrantes hondureña, nicaragüense, salvadoreña y venezolana las más importantes. En esta sección se estudia la distribución de las publicaciones de xenofobia por nacionalidad a la que van dirigidos.

Mediante la aplicación de análisis de texto y, particularmente expresiones regulares, es posible establecer la nacionalidad a la que van dirigidas las publicaciones de xenofobia. La forma de hacer esto es mediante la búsqueda de palabras o frases que denoten nacionalidad dentro del mensaje de xenofobia. Cabe resaltar que algunas publicaciones discriminatorias no expresan directamente la población a que van dirigidas; estas son clasificadas como “Migración general y otras nacionalidades”. En este sentido, la distribución calculada es una cota inferior de los mensajes de discriminación publicados hacia cada nacionalidad.

Gráfico 5.1 Volumen de Xenofobia por Nacionalidad



Meltwater (Construcción Propia), 2021.

El gráfico 5.1 muestra el porcentaje de mensajes que se refirieron a migrantes de nacionalidad salvadoreña, hondureña, venezolana o nicaragüenses. Salta a la vista que los mensajes xenófobos hacia la población migrante en general corresponde a un 47% del total de publicaciones capturadas, 22% a migrantes salvadoreños, 20% a población nicaragüense, 11% a población venezolana y 2% a población hondureña.

Gráfico 5.2 Ejemplo Mensajes de Xenofobia por Nacionalidad



Meltwater, 2021.

Un ejemplo de estos resultados se observa en el gráfico 5.2, así como en la sección 3.1 donde se aprecia que los mensajes producidos durante los picos de mayor volumen en la conversación se encuentran dirigidos en su mayoría en contra de migrantes en general, por sobre el resto de otras nacionalidades.

6. CONCLUSIONES Y RECOMENDACIONES DE POLÍTICA

El análisis de las publicaciones de xenofobia y lenguaje de odio en línea arroja algunos hallazgos que pueden ser utilizados para mitigar los riesgos asociados al incremento de la xenofobia. El primer hallazgo es que la cantidad de comentarios de xenofobia en línea es relativamente baja en comparación a otros países de la región. Más aún, no se encontraron brotes importantes de xenofobia durante el período estudiado. La mayoría de los mensajes de discriminación van dirigidos hacia la población migrante en general y no hacia una nacionalidad específica. Además, rechazan que esta población opine sobre los eventos políticos del país.

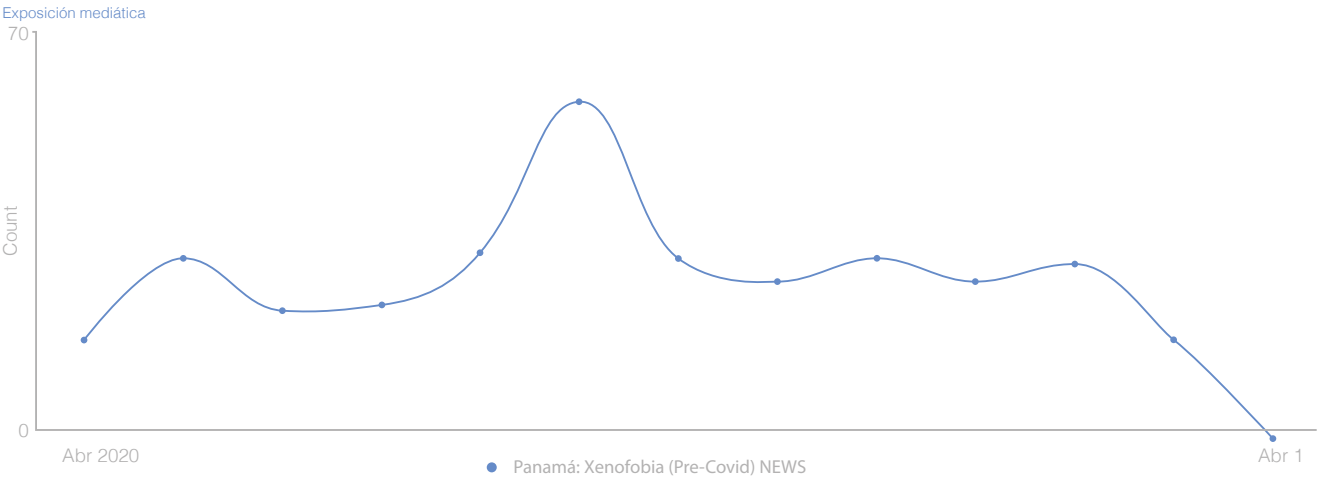
Recomendaciones

- Para tener una mejor comprensión de la xenofobia y discriminación en Guatemala debe completarse este análisis con otras fuentes de información. La cantidad de comentarios de xenofobia en línea es relativamente baja en comparación a otros países de la región y esto puede responder a diferentes fenómenos.
- Una parte de los mensajes de xenofobia encontrados mostraban rechazo hacia el hecho de que los migrantes opinaran sobre eventos políticos en el país. Este es un fenómeno que ha sido encontrado en otros países como Colombia y Perú. Las iniciativas para atender el aumento de la xenofobia deben tomar esto en cuenta.

ANEXO 1. NOTICIAS Y XENOFOBIA

Durante el periodo estudiado, pudimos evidenciar que se han generado 9979 publicaciones sobre migrantes, con un promedio de 27 publicaciones diarias. Sin embargo, no todas estas son catalogadas como xenófobas. Luego de realizar una nueva regla, se pudo apreciar que solo un 1,5% del total de publicaciones fueron con rasgos xenófobos. Es decir, 346 publicaciones, lo que implica un promedio de 0,96 publicaciones diarias.

Gráfico de volumen por mes, publicaciones xenófobas en medios.



Meltwater, 2021.

El peak que se genera en el mes de Septiembre responde a bandas de venezolanos que se dedican al robo de computadores y otras piezas de automóviles.. Otros ejemplos de noticias con rasgos xenófobos son:

Nuestro país, incluidos malandrines | La Prensa Panamá

Desdén, rencor. Pena y hasta vergüenza de habitar este pago. Así percibe este psicólogo y educador la expresión de alta densidad "este país ... la que se ha nacido. 'Este país' puede ser Ticolandia o **Nicaragua**. Alguien admite que no quiere ensuciarse con tanto truhan callejero, sobre

Residentes de La Peña en Darién se sienten amenazados

Los propios residentes y autoridades locales nos han hecho llegar sus testimonios, luego de vivir momentos de angustia frente a las ... de angustia frente a las acciones de **vandalismo** y violencia de los **migrantes** que permanecen en el campamento de **refugiados** de La Peña, en

Doce migrantes haitianos son procesados en Panamá por vandalizar un albergue

La defensa de los imputados anunció un recurso de apelación cuya audiencia se realizará el 12 de agosto, de acuerdo con la información ... en el albergue de La Peña luego de que un grupo de **migrantes amenazarán** con prender fuego al lugar si no los dejaban continuar su viaje.

Imponen medida cautelar a haitianos migrantes por vandalismo

El Ministerio Público MP informó que la Fiscalía Regional de Darién logró la medida cautelar de detención provisional para 12 migrantes ... albergue para migrantes **Migrantes** en Darién dan ultimátum al gobierno y **amenazan** con protestas Panamá podría expulsar a **migrantes** por actos

Colombiano que mató a su cuñada, saldrá de prisión en el año 2058. Lo absolvieron de un tercer delito

El atroz femicidio de la chiricana Leticia Lalyre Serracín en La Locería ya tiene un culpable
 70 años de edad, en la actualidad tiene tan solo 32. Este sujeto **colombiano** es el único responsable de la muerte Leticia Lalyre Serracín, en



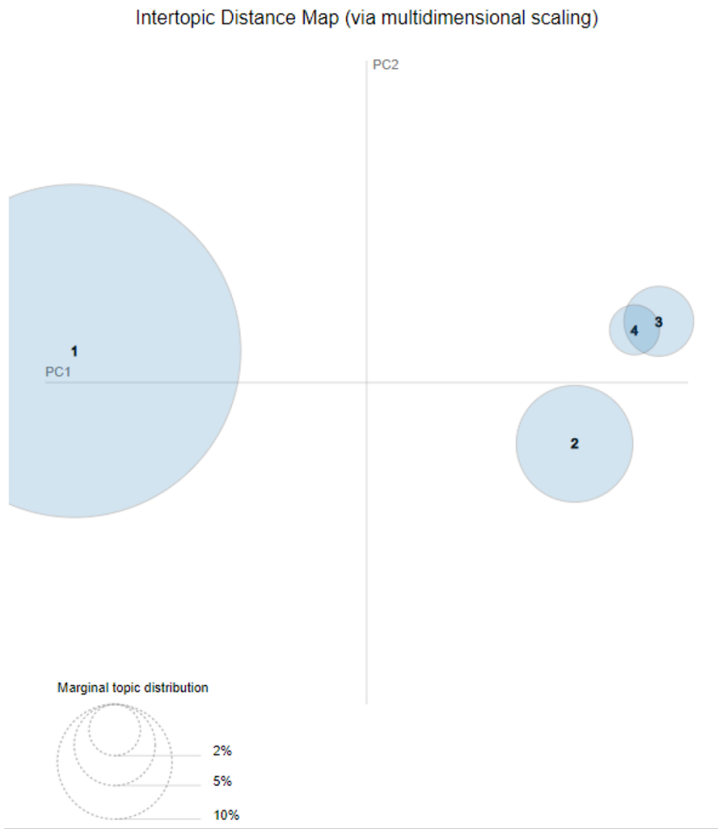
'Colombiana' presa por crimen y tentativa en El Chumical

El Juez de Garantías, Frank Torres Ruiz, decretó la medida cautelar personal de detención provisional a Leydi Guevara, de 31 años y de ...
 Elida Mc Gittenns, respectivamente. El hecho se dio cuando la '**Colombiana**', supuestamente, ordenó que los sicarios conocidos como 'Russo' y



ANEXO 2. ESTIMACIÓN A PARTIR DE LDA

Gráfico 4.6: Distribución general de tópicos dentro de la conversación de xenofobia

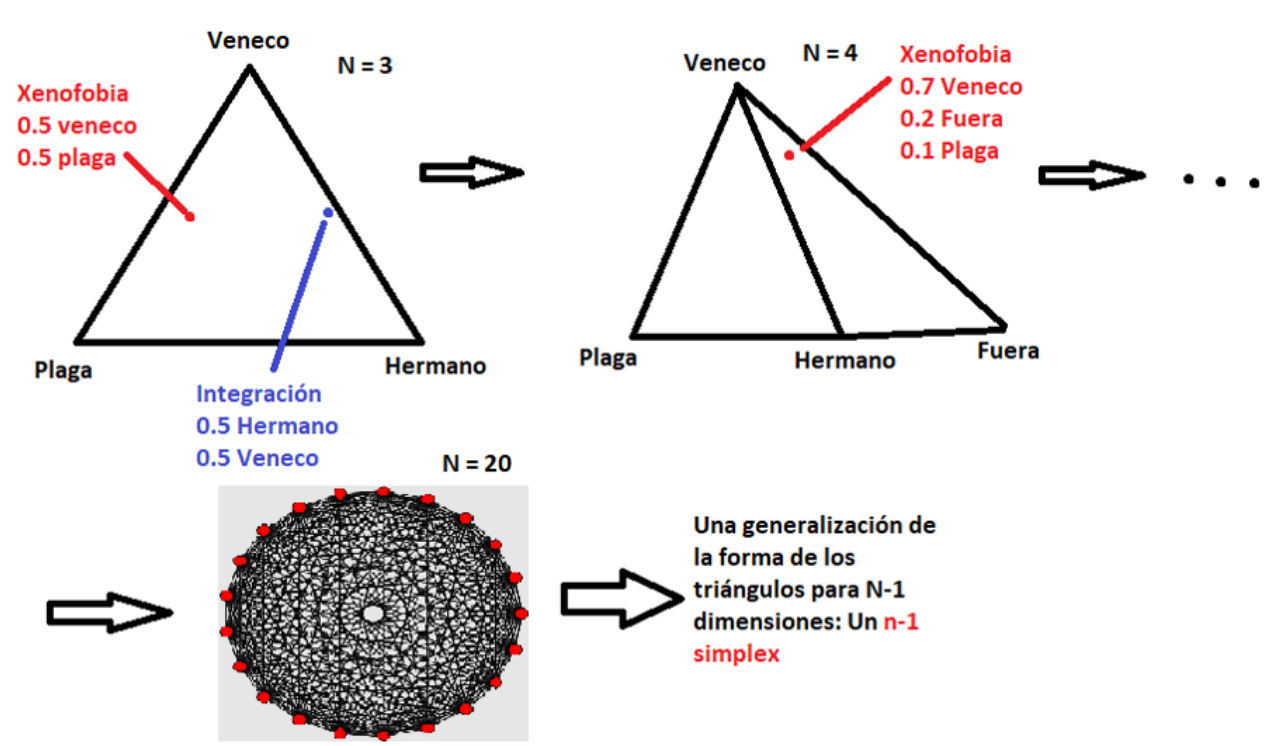


Fuente: Creación propia usando pyLDAvis. Los números representan el orden de los tópicos en términos de número de publicaciones. Es decir, el tópico 1 es el que agrupó más publicaciones y el tópico 4 el que menos. La interpretación del contenido de cada tópico es subjetiva al investigador. Esto quiere decir que, si bien el modelo propone la organización de unos tópicos basadas en las palabras que contienen cada Tweet, es trabajo del investigador estudiar las publicaciones que entraron dentro de cada grupo para determinar de qué se trata el tópico.

Una vez identificados los tópicos se puede calcular qué tan diferentes son entre ellos a partir del prior de Dirichlet. Esto es una métrica que nos permite entender qué tan separados o mixtos están los tópicos entre ellos usando como referencia los documentos en su interior. La distribución de Dirichlet suele representarse como un triángulo equilátero.

Como muestra la gráfica 4.1, a la derecha se encuentra un triángulo en donde cada uno de sus vértices es un tópico que puede aparecer en la conversación capturada a través de Meltwater. Los puntos que están al interior de este triángulo son los documentos o publicaciones que resultan de esta captura. Se puede observar como estos puntos están agrupados en el centro del triángulo, esto quiere decir las publicaciones están distribuidas de forma similar entre tópicos y que entre ellas existen diferencias reducidas (A esto le llamamos un alto prior de Dirichlet; $\alpha > 1$). Por su parte, el triángulo ubicado a la izquierda describe una situación en donde las publicaciones pertenecen con mayor probabilidad a un tópico específico y están separadas fuertemente entre ellas (un bajo prior de Dirichlet; $\alpha < 1$). Existen entonces, por su parte, palabras que pertenecen a varios tópicos (se conoce como el prior de Dirichlet sobre los términos; β); LDA permite establecer cuales son las palabras más usadas por tópico y con esa información entender la clasificación que está realizando sobre las publicaciones. Sin embargo, esto genera un problema de visualización, la idea de poseer una combinación lineal de tópicos por palabra supone que, si el diccionario es lo suficientemente grande, dejemos de visualizar ese triángulo equilátero en donde cada esquina es una palabra (véase gráfico 4.2).

Gráfico 4.2: Clasificación de tópicos con un diccionario creciente

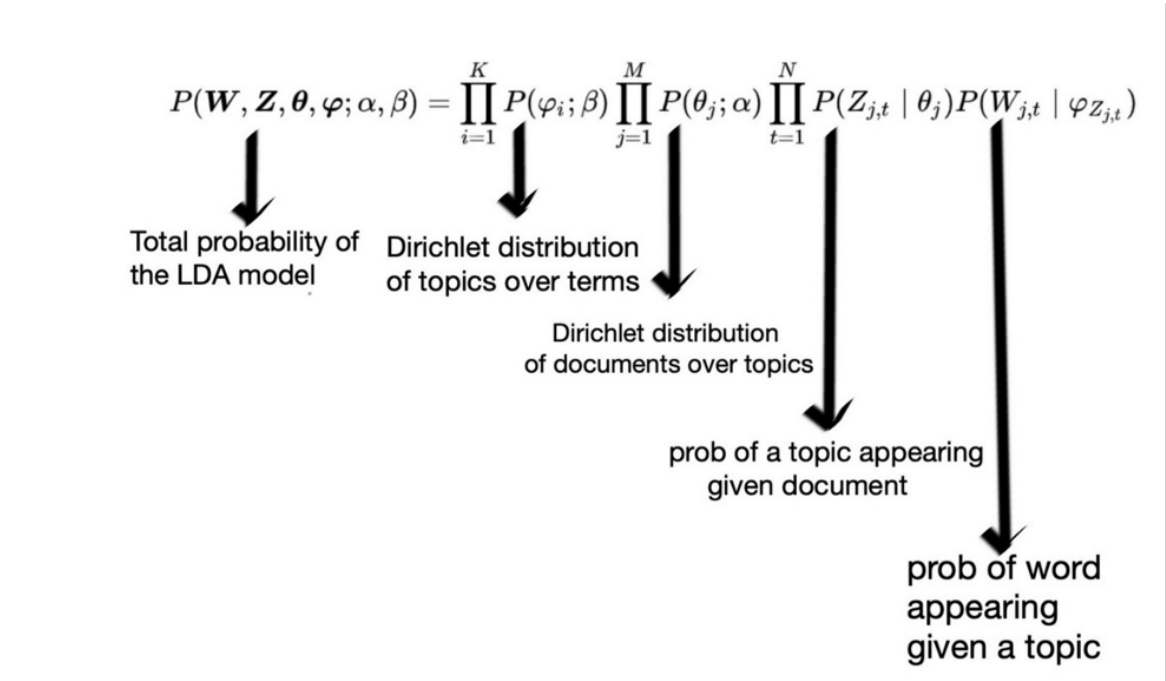


El gráfico 4.2 muestra que sucede con la clasificación de los tópicos una vez el número de palabras totales que constituyen el diccionario crecen. En él se puede encontrar como cada tópico se sitúa en un espacio de términos y como la combinación lineal de esos términos genera un tópico. Por ejemplo, el tópico de xenofobia cuando $N = 3$, se compone de forma equitativa por las publicaciones de 3 palabras que contienen solo “veneco” y “plaga”, una vez el número de palabras incrementa, la complejidad de las combinaciones lineales a su vez incrementa. El concepto de simplex resume esta idea del prior de Dirichlet

para N-1 dimensiones (en este caso palabras), esto es gracias a que podemos asegurar una generalización de la noción de triángulo o tetraedro para dimensiones arbitrarias.

El modelo LDA en su forma base es una combinación de probabilidades en donde cada una sigue un grupo de distribuciones y un conjunto específico de parámetros que dependen de su jerarquía.

Gráfico 4.3: Ecuación de la probabilidad total modelo LDA



Fuente: <https://towardsdatascience.com/latent-dirichlet-allocation-intuition-math-implementation-and-visualisation-63ccb616e094>

De izquierda a derecha, se tiene la probabilidad total del modelo, la distribución de Dirichlet de los tópicos sobre los términos (note el β en el paréntesis), la distribución de los documentos sobre los tópicos (note el α en el paréntesis), la probabilidad de que un tópico aparezca en un documento dado y la probabilidad de que una palabra aparezca dado un tópico determinado.

Una vez establecido el modelo se comienzan a optimizar los hiper parámetros del mismo, la idea es encontrar una combinación de β , α y el número de tópicos que maximice las métricas de desempeño del modelo, para este estudio, se usarán las métricas coherence score y perplexity score que son usuales en la literatura para evaluar el desempeño de los modelos LDA y así obtener una combinación óptima orientada a una clasificación eficiente. Una vez realizada la optimización, se procede a interpretar los resultados del modelo y a establecer una marcación de tópicos por documento.

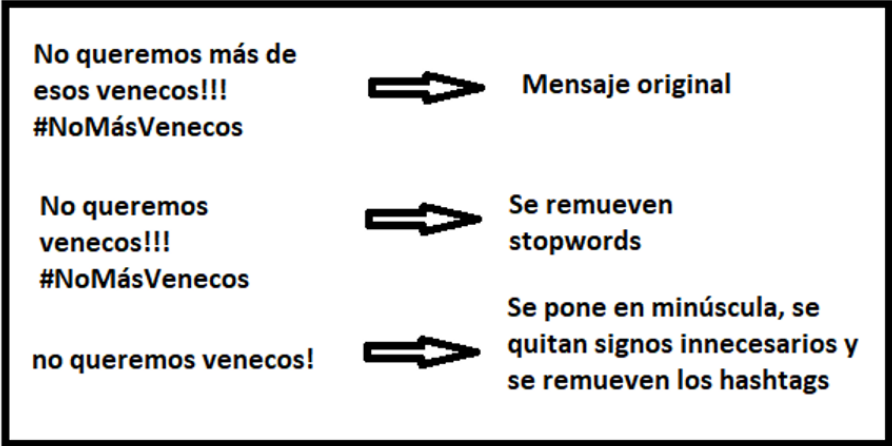
Procesamiento de Datos

Una vez obtenida la base de datos con cada una de las publicaciones de los usuarios, el objetivo es analizar el contenido de estas. Una forma de realizar el análisis es a partir de técnicas de inteligencia artificial y modelos de clasificación que permiten entender la forma en que se relacionan los mensajes. Para poder utilizar estos modelos, es necesario procesar el contenido de las publicaciones con el fin de hacerlas comparables y eliminar el ruido. Este proceso se llama vectorización de las publicaciones, y tiene como objetivo obtener una representación única y coherente de cada mensaje.

El primer paso de la vectorización busca eliminar palabras que no interesan para definir relaciones

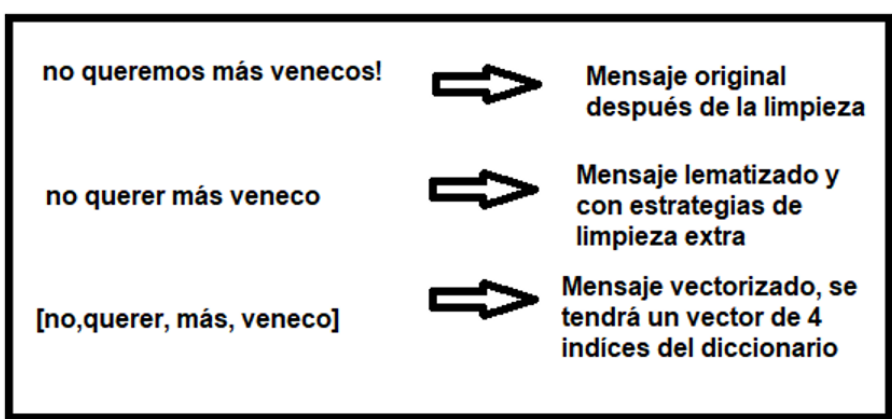
entre publicaciones. Consiste de distintas técnicas de limpieza como la eliminación de OOV (palabras fuera del vocabulario o Out-Of-Vocabulary words), stopwords (conectores lexicográficos) y las OUW (palabras demasiado usadas o Over-Used-Words). El segundo paso consiste en eliminar elementos dentro de las publicaciones que sean diferentes a contenido de léxico, ya que esto puede ensuciar la clasificación o las relaciones que encontremos entre los documentos. Siguiendo esta idea, se eliminan links (HTTP, HTTPS), hashtags, signos de puntuación innecesarios, dobles espacios, letras repetidas y letras mayúsculas. A continuación, se presentan ejemplos de cómo se verían los mensajes luego de esta limpieza y de algunas palabras que recaen en estas categorías de eliminación.

Gráfico 4.4: Ejemplo de limpieza de un mensaje de xenofobia



Una vez realizada esta primera limpieza, se propone construir una especie de diccionario con las palabras que contiene cada texto. Para esto, debemos introducir el concepto de Bag-Of-Words (BOW o bolsa de palabras), el cual sugiere que cada palabra puede ser representada con un índice que la posiciona de forma única en la bolsa (diccionario). Previo a esto, se requiere convertir palabras en forma flexionada a su lema correspondiente, este proceso lingüístico se conoce como lematización y es ampliamente usado en el procesamiento del lenguaje natural para interpretar y representar textos. De igual manera, la técnica conocida como “Stemming” permite reducir las palabras a su raíz correspondiente, la cual nos da la oportunidad de encontrar mejores relaciones entre textos, reduciendo palabras como “bibliotecario” y “biblioteca” a una raíz conocida equivalente “biblioteca”.

Gráfico 4.5: Ejemplo de vectorización de mensajes



Una vez construido el diccionario, se procede a construir los documentos que contienen las palabras modificadas. Un documento es un conjunto de vectores que comprenden la estructura numérica de las palabras contenidas en el mismo. Un uso importante de estos documentos es que permiten evidenciar n-gramas (subsecuencias de palabras) de las publicaciones por cada uno de los autores y así obtener diferentes representaciones de un mensaje. En la siguiente sección veremos cómo vamos a usar estos documentos para clasificar los mensajes basados en sus relaciones, para este informe, lo llamaremos modelamiento de tópicos.