Universidad del Desarrollo
Facultad de Ingeniería

ANÁLISIS DE TEXTOS UTILIZANDO TÉCNICAS DE NLP

Análisis de las respuestas de los ciudadanos que participaron en la iniciativa “El Chile que Queremos”

POR: CLAUDIA ANGELICA HERRERA ROJAS

Proyecto de grado presentado a la Facultad de Ingeniería de la Universidad del Desarrollo para optar al grado académico de Magíster en Data Science

PROFESOR GUÍA:
DRA. MARÍA PAZ RAVEAU MORALES

[MARZO 2024]
[SANTIAGO]

A mi amada familia, el pilar fundamental que impulsa cada uno de mis proyectos. En cada paso que doy, encuentro la fuerza y el apoyo incondicional que solo ustedes pueden brindarme. Son mi inspiración, mi guía y mi mayor motivación.

Tabla de contenido

INDICE DE TABLAS Y FIGURAS

TABLAS

FIGURAS

Resumen

El "Estallido social" de octubre de 2019 fue un reflejo de la insatisfacción de los ciudadanos frente al modelo actual. En respuesta, iniciativas como "El Chile que Queremos" (ECQQ) surgieron para iniciar un proceso de escucha social a través de diálogos ciudadanos.

El presente trabajo se enfoca en analizar las respuestas de ECQQ para comprender las percepciones de los ciudadanos sobre las necesidades del país y las emociones experimentadas después de octubre de 2019 . Se aplicó Latent Dirichlet Allocation (LDA) para identificar áreas de interés, evidenciando la consistencia de estos resultados con otros estudios realizados en el país sobre el estallido social. Las emociones predominantes fueron el miedo y la tristeza, especialmente entre las personas mayores. Sin embargo, la esperanza también se destacó, particularmente entre los jóvenes.

A través del cálculo del coeficiente V de Cramér se evidenció una asociación débil entre temas dominantes y categorías de emociones, debido a la heterogeneidad presente en estas categorías.

1. Introducción

Chile es uno de los tres países latinoamericanos más desiguales en cuanto a ingresos. Según el Informe bienal de la OCDE 2020, en 2017, el 20% de la población chilena más privilegiada ganaba 10,31 veces más que el 20% menos favorecido.

El 18 de octubre del 2019 una serie de manifestaciones masivas iniciaron lo que se conoció como "estallido social" en Chile. Los principales factores que desencadenaron esta revuelta corresponden a las lógicas estructurales del modelo socioeconómico neoliberal y las profundas desigualdades que éste produce (Mayol, 2019; Cortés, 2019; Güell, 2019; Araujo, 2015). Dicho modelo se expresa en altos índices de desconfianza en las instituciones, el malestar colectivo y la sensación de abuso y maltrato por parte de la élite política y económica chilena (Gutiérrez-Muñoz, 2020; Mayol, 2019).

Posterior a las manifestaciones de octubre 2019 surgieron diversos cabildos y algunas iniciativas de diálogos tales como "Tenemos que hablar de Chile" (TQH) y "El Chile que Queremos" (ECQQ) en dónde los ciudadanos manifestaron sus emociones y opiniones frente a las principales necesidades del país.

El informe "Demandas prioritarias y propuestas para un Chile diferente" del año 2021 menciona que las discusiones en cada cabildo levantaron un diagnóstico de los problemas que aquejaban a colectividades y territorios, como también propuestas de soluciones para subvertirlas.

En el presente estudio se presenta un análisis de las respuestas de los ciudadanos que participaron en el proyecto “El Chile que Queremos” con el objetivo de conocer sus emociones y principales expectativas y percepciones con respecto a las principales necesidades que aquejan al país.

ECQQ se llevó a cabo a través de un sitio web que permitió realizar tres tipos de consultas:

(i) Consulta individual (13.947 participantes)

(ii) Diálogos de niños, niñas y adolescentes (12.789 participantes en 864 diálogos)

(iii) Diálogos ciudadanos autoconvocados (86.747 participantes en 12.587 diálogos).

Los diálogos ciudadanos autoconvocados, y diálogos de niños, niñas y adolescentes se realizaron de manera presencial y con cobertura en 343 comunas.

Este estudio se centra en el análisis de respuestas de ECQQ, utilizando herramientas de clasificación no supervisada como Latent Dirichlet Allocation (LDA)

2. Marco Teórico

Definición de PLN

El procesamiento del lenguaje natural (PLN) hace referencia a la rama de la informática y más específicamente, a la rama de la inteligencia artificial o IA encargada de dar a los ordenadores la capacidad de comprender textos y palabras habladas de la misma manera que los seres humanos.

El procesamiento del lenguaje natural combina la lingüística computacional (modelado basado en reglas del lenguaje humano) con modelos estadísticos, de *Machine learning* y *Deep learning*. Juntas, estas tecnologías permiten a los ordenadores procesar el lenguaje humano en forma de datos de texto o voz y "comprender" su significado completo, junto con la intención y el sentimiento del orador o escritor.
(Fuente: IBM)

Importancia y aplicación de PLN

El procesamiento de lenguaje natural (PLN) es fundamental para analizar los datos de texto y voz de manera eficiente y en profundidad. Puede resolver las diferencias en dialectos, jerga e irregularidades gramaticales típicas en las conversaciones cotidianas.
Las empresas lo utilizan para varias tareas automatizadas, como:

- Procesar, analizar y archivar documentos grandes.
- Analizar los comentarios de los clientes o las grabaciones de centros de atención telefónica.
- Ejecutar chatbots para ofrecer un servicio al cliente automatizado.
- Responder preguntas de quién, qué, cuándo y dónde.
- Clasificar y extraer texto.

También se puede integrar el PLN en aplicaciones orientadas al cliente para comunicarse de manera más eficaz con ellos. Por ejemplo, un chatbots analiza y ordena las consultas de los clientes, responde automáticamente a las preguntas comunes y redirige las consultas complejas al servicio de atención al cliente. Esta automatización ayuda a reducir los costos, evita que los agentes dediquen tiempo a las consultas redundantes y mejora la satisfacción del cliente.

(Fuente: AWS)

Bolsa de palabras

El modelo de bolsa de palabras convierte el texto en vectores de longitud fija contando las veces que aparece cada palabra. Lo anterior se ilustrará en el siguiente ejemplo, considerando las siguientes frases en inglés.

- Text processing is necessary.
- Text processing is necessary and important.

- Text processing is easy.

Nos referiremos a cada una de las frases anteriores como documentos. Si quitamos las palabras únicas de todas estas frases, el vocabulario estará formado por estas 7 palabras {‘Text’, ’processing’, ’is’, ’necessary’, ’and’, ’important, ’easy’}. Para realizar el bag-of-words, simplemente tendremos que contar el número de veces que aparece cada palabra en cada uno de los documentos.

Document	Text	preprocessing	is	necessary	and	important	easy
1	1	1	1	1	0	0	0
2	1	1	1	1	1	1	0
3	1	1	1	0	0	0	1

Figura 1 Ejemplo de una Bolsa de palabras

Así, tenemos los siguientes vectores para cada uno de los documentos de longitud fija -7:

- Documento 1: [1,1,1,1,0,0,0]
- Documento 2: [1,1,1,1,1,0]
- Documento 3: [1,1,1,0,0,1]

Limitaciones de la bolsa de palabras:

- Si utilizamos la bolsa de palabras para generar vectores para documentos grandes, los vectores serían de gran tamaño y también tendrían demasiados valores nulos, lo que llevaría a la creación de vectores dispersos.
- La bolsa de palabras no aporta ninguna información sobre el significado del texto. Por ejemplo, si consideramos estas dos frases “El procesamiento de textos es fácil

pero tedioso" y "El procesamiento de textos es tedioso pero fácil", un modelo de bolsa de palabras crearía los mismos vectores para ambas, aunque tengan significados diferentes.

(Fuente: Kaggle)

TF-IDF (Term Frequency-Inverse Document Frequency):

TF-IDF es una medida ponderada que se utiliza para evaluar la importancia de una palabra en un conjunto de documentos, teniendo en cuenta la frecuencia de la palabra en los documentos y la rareza de la palabra en todo el conjunto de documentos (Salton & McGill, 1983). Esta representación puede mejorar la eficacia de los algoritmos de aprendizaje automático y análisis de texto en comparación con la representación de Bolsa de palabras.

La representación TF-IDF puede mejorar la eficacia de los algoritmos de aprendizaje automático y análisis de texto en comparación con la representación de Bolsa de palabras, ya que ajusta las ponderaciones de las palabras en función de su relevancia en el contexto del conjunto de documentos (Manning, Raghavan, & Schütze, 2008).

Stemming y lematización en el procesamiento del lenguaje natural

En el ámbito del procesamiento del lenguaje natural (PLN) y el análisis de textos, la normalización de textos desempeña un papel fundamental. Dos de las técnicas de

normalización más utilizadas en el ámbito de la ciencia de los datos y la inteligencia artificial son el “Stemming” y la “Lematización” (Fuente: SEO North).

En PLN, el Stemming recorta las palabras a su forma raíz eliminando los afijos, mientras que la Lematización reduce las palabras a su forma base del diccionario, teniendo en cuenta su contexto y significado.

Word	Stemming	Lemmatization
information	inform	information
informative	inform	informative
computers	comput	computer
feet	feet	foot

Figura 2 Ejemplo de Stemming y Lematización

(Fuente: SEO North)

El algoritmo de Stemming más popular, sobre todo en inglés, es el Porter Stemmer. Desarrollado por Martin Porter, elimina los sufijos (y en algunos casos los prefijos) de las palabras. También destaca el algoritmo Snowball, más agresivo y compatible con varios idiomas. (Fuente: SEO North)

La Lematización es un proceso más sofisticado que Stemming. Consiste en reducir una palabra a su forma base o de diccionario, conocida como lema. A diferencia del Stemming, la lematización tiene en cuenta el significado de la palabra, su parte de la oración y el análisis morfológico para lograr esta reducción. (Fuente: SEO North)

Enfoques para el procesamiento de lenguaje natural

PLN supervisado

Los métodos de PLN supervisados entrenan el software con un conjunto de entradas y salidas etiquetadas o conocidas. Primero, el programa procesa grandes volúmenes de datos conocidos y aprende a producir el resultado correcto a partir de cualquier entrada desconocida. Por ejemplo, las empresas entrenan a las herramientas de PLN para categorizar los documentos según etiquetas específicas. (Fuente: AWS)

PLN no supervisado

El PLN no supervisado utiliza un modelo de lenguaje estadístico para predecir el patrón que se produce cuando se alimenta mediante entradas no etiquetadas. Por ejemplo, la función de autocompletar en los mensajes de texto sugiere palabras relevantes que tienen sentido para la oración al monitorear la respuesta del usuario. (Fuente: AWS)

Modelo LDA

LDA, conocido como Asignación Latente de Dirichlet, es un modelo probabilístico generativo para colecciones de datos discretos tales como corpus de texto. LDA se presenta como un modelo bayesiano jerárquico de tres niveles, en el que cada elemento

de una colección se modela como una mezcla finita sobre un conjunto subyacente de temas. Cada tema, a su vez, se modela como una mezcla infinita sobre un conjunto subyacente de probabilidades de tema. En el contexto del modelado de texto, las probabilidades del tema proporcionan una representación explícita de un documento (Blei, D. M., Ng, A. Y., & Jordan, M. I., 2003).

Es un método utilizado en Topic Modeling para identificar y extraer tópicos latentes en corpus de texto. Esta técnica es ampliamente utilizada en el análisis de datos textuales para descubrir patrones subyacentes y estructuras temáticas en grandes conjuntos de documentos.

Características del modelo LDA

El modelo LDA tiene, entre otras, las siguientes características:

- Es un modelo probabilístico, lo que quiere decir que se basa en las probabilidades de cercanía, entre otras.
- Es un modelo de aprendizaje no supervisado, es decir, no tenemos información a priori de los posibles topics que hay o, al menos, no están etiquetados. Aquí no tenemos que darle ningún tipo de etiquetado a los datos, simplemente le damos un texto y él, a partir del submodelo probabilístico, es capaz de hacer los cálculos pertinentes.
- El modelo LDA asume que:

- Documentos con topics o temáticas similares emplearán palabras similares.
- Los documentos están compuestos por un conjunto de topics que siguen una determinada distribución.
- Los topics están compuestos por un conjunto de palabras que, al igual que estos, siguen una determinada distribución

Algunas cosas que debemos tener en cuenta cuando hagamos uso del algoritmo o modelo LDA son:

- Debe fijarse el vocabulario al inicio, es decir, antes del entrenamiento. Nosotros le damos los datos al algoritmo, no hacen falta etiquetas, pero sí es necesario pasarle el vocabulario y tenemos que hacer un preprocesamiento anticipado de esos datos, para que así tenga un buen rendimiento.
- Debemos tener en cuenta que el modelo LDA es probabilístico basado en frecuencias, por tanto, si no aplicamos limpieza para que pueda calcular bien esas frecuencias, será un problema de cara a la generación de temas. En este caso, eliminar stopwords suele arrojar mejores resultados.
- El modelo LDA se basa en la representación de Bag-of-words, esto es, los vocabularios con frecuencias.
- Debemos definir previamente el número de tópicos que queremos que extraiga.

(Fuente: KeepCoding)

Test de chi-cuadrado.

La prueba de Chi-cuadrado es una excelente opción para comprender e interpretar la relación entre dos variables categóricas.

La tabulación cruzada presenta las distribuciones de dos variables categóricas simultáneamente, con las intersecciones de las categorías de las variables que aparecen en las celdas de la tabla.

El cálculo estadístico de Chi-Cuadrado y su comparación con un valor crítico de la distribución Chi-Cuadrado permite al investigador evaluar si los recuentos de celdas observados son significativamente diferentes de los recuentos de celdas esperados.

Debido a la forma en que se calcula el valor de Chi-Cuadrado, es extremadamente sensible al tamaño de la muestra: cuando el tamaño de la muestra es demasiado grande (~500), casi cualquier pequeña diferencia parecerá estadísticamente significativa.

También es sensible a la distribución dentro de las celdas. Esto puede solucionarse utilizando siempre variables categóricas con un número limitado de categorías.

Tipos de pruebas de Chi-Cuadrado

Existen diferentes tipos de pruebas de Chi-Cuadrado: Prueba de bondad de ajuste, prueba de independencia y prueba de homogeneidad.

Prueba de bondad de ajuste

La prueba de bondad de ajuste Chi-cuadrado se utiliza para comparar una muestra recogida aleatoriamente que contiene una única variable categórica con una población mayor.

Esta prueba se utiliza con mayor frecuencia para comparar una muestra aleatoria con la población de la que se ha recogido potencialmente.

Prueba de independencia

La prueba de independencia de Chi-Cuadrado busca una asociación entre dos variables categóricas dentro de la misma población.

A diferencia de la prueba de bondad de ajuste, la prueba de independencia no compara una única variable observada con una población teórica, sino dos variables dentro de un conjunto de muestras entre sí.

Prueba de homogeneidad de Chi-Cuadrado

La prueba de homogeneidad de Chi-Cuadrado se organiza y ejecuta exactamente igual que la prueba de independencia.

La principal diferencia que hay que recordar entre ambas es que la prueba de independencia busca una asociación entre dos variables categóricas dentro de la misma población, mientras que la prueba de homogeneidad determina si la distribución de una variable es la misma en cada una de varias poblaciones (asignando así la propia población como segunda variable categórica).
(Fuente: QuestionPro)

V de Cramer

El V de Cramer es una medida de la fuerza de asociación entre dos variables nominales.

- Va de 0 a 1 donde:
  - 0 indica que no hay asociación entre las dos variables.
  - 1 indica una fuerte asociación entre las dos variables.
- Se calcula como:

V de Cramer = $\sqrt{}$ (X 2 / n) / min (c-1, r-1)

- dónde:
  - X 2 : el estadístico Chi-cuadrado

- n: tamaño total de la muestra
- r: número de filas
- c: número de columnas

(Fuente: Statologos)

3. Hipótesis y Objetivos

Hipótesis del Trabajo:

"Se plantea que existe una relación entre las emociones experimentadas posterior al estallido social de octubre 2019 y las principales necesidades país identificadas en las respuestas de los ciudadanos que participaron en la iniciativa "El Chile Que Queremos". Además, se postula que mediante el método de clasificación no supervisada Latent Dirichlet Allocation (LDA), es posible identificar temas relevantes que reflejen estas necesidades país de manera agrupada"

Objetivo General:

Investigar la relación entre las principales necesidades país percibidas por los ciudadanos y las emociones experimentadas después de octubre de 2019, utilizando técnicas de análisis de clusterización no supervisada y modelado de temas.

Objetivos Específicos:

- Realizar un análisis de clusterización mediante el método de clasificación no supervisada Latent Dirichlet Allocation (LDA) para modelar temas en respuestas de las necesidades que enfrenta el país.
- Realizar un análisis de la relación entre temas de respuestas de las necesidades que enfrenta el país y las emociones experimentadas posterior al estallido social de octubre 2019.

4. Datos y Metodología

Se realizó un análisis de las respuestas de los ciudadanos que participaron en “El Chile que Queremos” con respecto a las respuestas que proporcionaron en relación con las siguientes preguntas post octubre 2019:

1. ¿Cómo me he sentido dentro de las últimas semanas?
2. ¿Cuáles son las necesidades que enfrenta el país?

Selección y preparación de los datos.

1. Los datos fueron extraídos desde el GitHub del Ministerio de Ciencia, Tecnología, Conocimiento, e Innovación: https://github.com/MinCiencia/ECQQ
2. Se escogió la base “BBDD_Dialogos.csv” (13324 Filas x 289 Columnas) ya que contenía las respuestas a las preguntas en relación con las emociones y necesidades del país. Las respuestas de las personas que participaron en los diálogos se fueron registrando en diferentes columnas , dado lo anterior para poder continuar con el análisis fue necesario realizar ajustes en la base original; registrando las respuestas a una pregunta en una sola columna. Con la implementación de estos cambios las dimensiones de la base cambiaron a 66620 Filas x 289 Columnas.

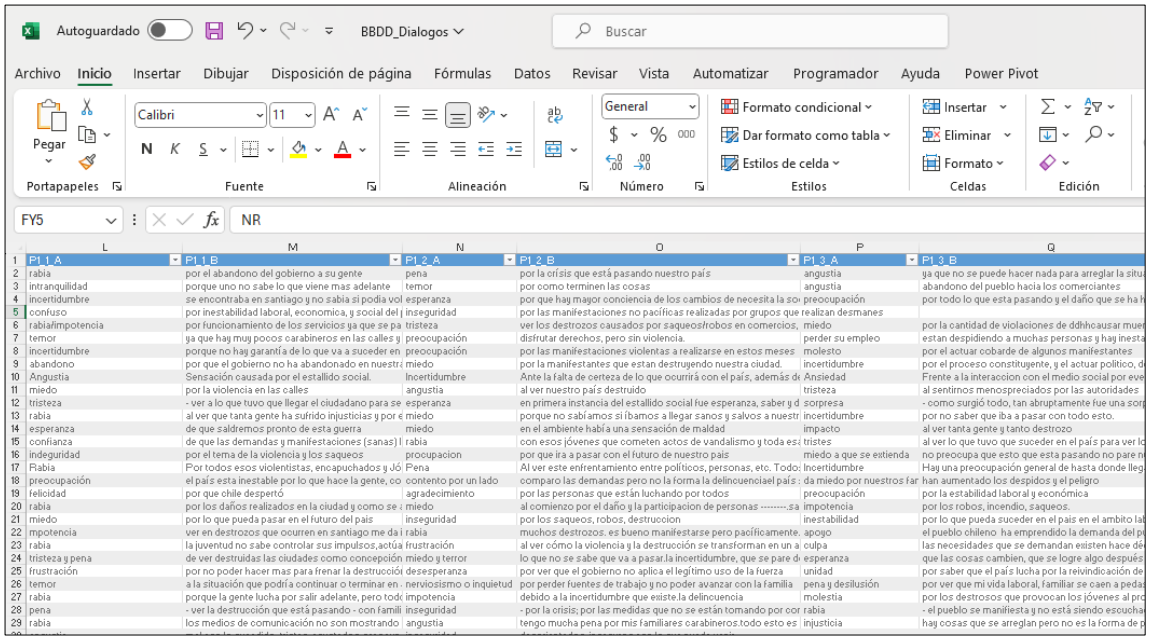

Figura 3 Base "BBDD_Dialogos.csv"

3. Asimismo, para realizar el análisis de la relación entre las necesidades que enfrenta el país y las emociones fue necesario añadir a la base original:

    - Una columna con la categoría a la cuál correspondía la emoción, a la cual se llamó "categoria_emocion" (se extrajo esta información desde la base "emotions.csv" y se incorporó a la base "BBDD_Dialogos.csv" utilizando la función de Excel "BUSCARV")
    - Una columna con el nombre "sentimiento", en dónde se asignó la etiqueta de negativo a las categorías de emociones "Aversión", "Enojo", "Miedo", "Sorpresa" y "Tristeza" y la etiqueta de positivo a la categoría "Alegría"

- Una columna con el nombre “Categoria_edad”, en dónde se asignó la etiqueta de Juventud a las personas con edades igual o menor a 29 años[1], la etiqueta de Adultez a las personas con edades comprendidas entre los 30 y 59 años y la etiqueta de Adultez mayor a las personas con edades mayores o iguales a 60 años (SENAMA, 2002).
- Una columna a la cual se llamó “Emoción_Persona” en dónde se concatenaron las variables “Emoción” y “Respuesta Emoción”
- Una columna a la cual se llamó “Necesidades_pais” en dónde se concatenaron las variables “Necesidades” y “Respuesta Necesidades”

4. La base final “BBDD.csv” quedó conformada con 66620 Filas x 34 Columnas y a partir de ella se creó un nuevo DataFrame (66620 Filas x 9 Columnas) filtrando solo las variables que guardan relación con el análisis.

Tabla 1 Descripción del Dataframe final para análisis

NOMBRE	DESCRIPCIÓN
EMOCION	Emoción de la persona con respecto a cómo se ha sentido post octubre 2019
RESPUESTA_EMOCION	Respuesta de la emoción

[1] Esta definición se encuentra definida por el Instituto Nacional de la Juventud (INJUV) como aquellas personas que se encuentra entre los 15 y 29 años.

EMOCION_PERSONA	Concatenación de la variable "EMOCION" y "RESPUESTA_EMOCION"
CATEGORIA_EMOCION	Agrupa las emociones en las categorías Miedo, Tristeza, Enojo, Alegría, Aversión y Sorpresa.
SENTIMIENTO	Etiqueta las emociones como negativas o positivas
NECESIDADES_PAIS	Concatenación de la variable "NECESIDADES" y "RESPUESTA_NECESIDADES"
LP_EDAD	Edad de los participantes
CATEGORIA_EDAD	Agrupa las edades en las categorías Juventud, Adultez y Adultez Mayor
LP_SEXO	Sexo de los participantes

Fuente: Elaboración propia.

Bibliotecas y Módulos utilizados

- Módulo os: Proporciona funciones para interactuar con el sistema operativo, como la manipulación de archivos y directorios. (Fuente: Tutz.tv)
- Matplotlib: Biblioteca completa para la creación de visualizaciones estáticas, animadas e interactivas en Python. (Fuente: matplotlib.net)
- Módulo pyplot: Es un módulo Matplotlib que propone varias funciones sencillas para añadir elementos tales como líneas, imágenes o textos a los ejes de un gráfico. (Fuente: DataScientest)

- Seaborn: Seaborn es una biblioteca para crear gráficos estadísticos en Python. Está basada en Matplotlib, y se integra con las estructuras de Pandas. (Fuente: DataScientest)
- Pandas: Biblioteca de Python utilizada para manipulación y análisis de datos. (Fuente: Master Data Scientist)
- NumPy: Biblioteca de Python utilizada para realizar operaciones matemáticas y estadísticas en Python. (Fuente: DataScientest)
- Módulo math: Este módulo proporciona acceso a las funciones matemáticas definidas en el estándar de C. (Fuente: Python Software Foundation)
- NLTK (Natural Language Toolkit): Es un conjunto de bibliotecas y programas de software. Este conjunto de herramientas reúne los algoritmos más comunes en el procesamiento del lenguaje natural , como tokenización, etiquetado de partes del discurso, derivación, análisis de sentimientos, segmentación de temas y reconocimiento de entidades nombradas. (Fuente: DataScientest)
- Módulo itertools: El módulo itertools de Python proporciona una colección de funciones que crean objetos iterables. Estos objetos pueden utilizarse para realizar operaciones de iteración complejas de forma eficiente. (Fuente: CodigosPython)
- Módulo re: Proporciona funciones para trabajar con expresiones regulares en Python. (Fuente: Python Software Foundation)
- Módulo String: El Módulo String ofrece operaciones y constantes para trabajar con cadenas de caracteres. (Fuente: Python Software Foundation)

- Stanza: Biblioteca de procesamiento del lenguaje natural para análisis de texto en varios idiomas. El objetivo de Stanza no es reemplazar las herramientas de modelado que elija, sino ofrecer implementaciones para patrones comunes útiles de experimentos en aprendizaje automático. (Fuente: Editorialia)
- Gensim: Es una biblioteca de código abierto para el modelado de temas no supervisados. (Fuente: DataScientest)
- pyLDAvis: Biblioteca Python para visualización interactiva de modelos de temas. Esta es una adaptación del paquete R de Carson Sievert y Kenny Shirley. (Fuente: The Python Package Index PyPI)
- Wordcloud: Librería para crear nubes de palabras en Python. Comúnmente utilizada para visualizar las palabras más frecuentes en un conjunto de datos de texto. (Fuente: DataScientest)
- SciPy: Biblioteca de código abierto para Python que ofrece herramientas y algoritmos matemáticos, incluyendo optimización, álgebra lineal, integración, interpolación, funciones especiales, FFT, procesamiento de señales e imagen, resolución de EDOs y otras tareas relacionadas con la ciencia e ingeniería. (Fuente: Unipython)

Metodología

La metodología utilizada para realizar el diagnóstico y etiquetado de tópicos utilizando LDA (Latent Dirichlet Allocation) fue la siguiente:

1. Limpieza y preparación de los datos:
    - Realizar la limpieza de datos eliminando stopwords y caracteres especiales.
    - Aplicar Tokenización para dividir el texto en palabras (tokens) para facilitar el análisis.
    - Aplicar lematización para normalizar textos y reducir la cardinalidad del vocabulario asociado, lo que facilita su procesamiento y análisis.
    - Obtener un corpus de texto relevante para el análisis.

2. Construcción del modelo LDA:
    - Utilizar bibliotecas como Gensim o Scikit-Learn en Python para construir el modelo LDA.

3. Diagnóstico del modelo:
    - Evaluar la coherencia semántica de los tópicos generados para determinar la calidad del modelo.

- Evaluar la perplejidad del modelo LDA para determinar la calidad del modelo.

4. Visualización de tópicos:
    - Utilizar herramientas como pyLDAvis para visualizar la distribución de tópicos y facilitar su interpretación.

5. Etiquetado de tópicos:
    - Revisar las palabras de mayor probabilidad asociadas a cada tópico generado por el modelo.
    - Asignar etiquetas descriptivas basadas en el significado de las palabras clave.
6. Identificación de los tópicos dominantes.
7. Análisis de relación entre Temas y Emociones.
8. Evaluación del modelo.

5. Desarrollo

Para llevar a cabo el desarrollo de este trabajo se realizó un modelado de temas aplicando el método clasificación no supervisada LDA (Latent Dirichlet Allocation). El objetivo fue obtener clústeres de diferentes temas en base a las palabras que aparecían en las respuestas de las necesidades que enfrenta el país, y luego realizar un análisis de la relación entre las necesidades que enfrenta el país y las emociones post octubre 2019.

Cómo se mencionó anteriormente para llevar a cabo el análisis se utilizó la base "BBDD.csv" y a partir de ella se creó un nuevo DataFrame (66620 Filas x 10 Columnas) filtrando solo las variables que guardan relación con el análisis.

Limpieza y preparación de los datos.

Se realizó un conjunto de transformaciones en la columna 'necesidades_pais' del DataFrame con el objetivo de realizar la limpieza de los datos.

Las transformaciones consistieron en:

- Convertir todas las letras en el texto a minúsculas.
- Eliminar todos los dígitos numéricos del texto.
- Eliminar toda la puntuación del texto.
- Reemplazar los saltos de línea (\n) con espacios en blanco.
- Eliminar los espacios en blanco al principio y al final de cada texto

Luego de la limpieza de los datos las dimensiones finales del Dataframe quedaron en 50318 Filas x 9 Columnas.

Generación del diccionario y corpus, principales entradas del modelo LDA

El corpus se construyó aplicando lematización de textos utilizando Stanza, tokenización utilizando word_tokenize de NLTK y eliminación de las stopwords en la variable "necesidades_pais" del dataframe. Finalmente se creó el diccionario basado en el texto tokenizado y lematizado.

Construcción del Modelo LDA y Diagnóstico de Tópicos.

Se creó el modelo LDA utilizando Gensim y se realizó un Diagnóstico de tópicos, calculando la coherencia semántica del modelo LDA y la perplejidad. Los resultados se presentan en las Figuras 4 y 5.

Se observa que la coherencia alcanza su valor máximo en el rango comprendido entre 10 y 15 tópicos. Dado lo anterior se procedió a calcular la coherencia y la perplejidad para los tópicos comprendidos en este rango. Los resultados se presentan en la tabla 2

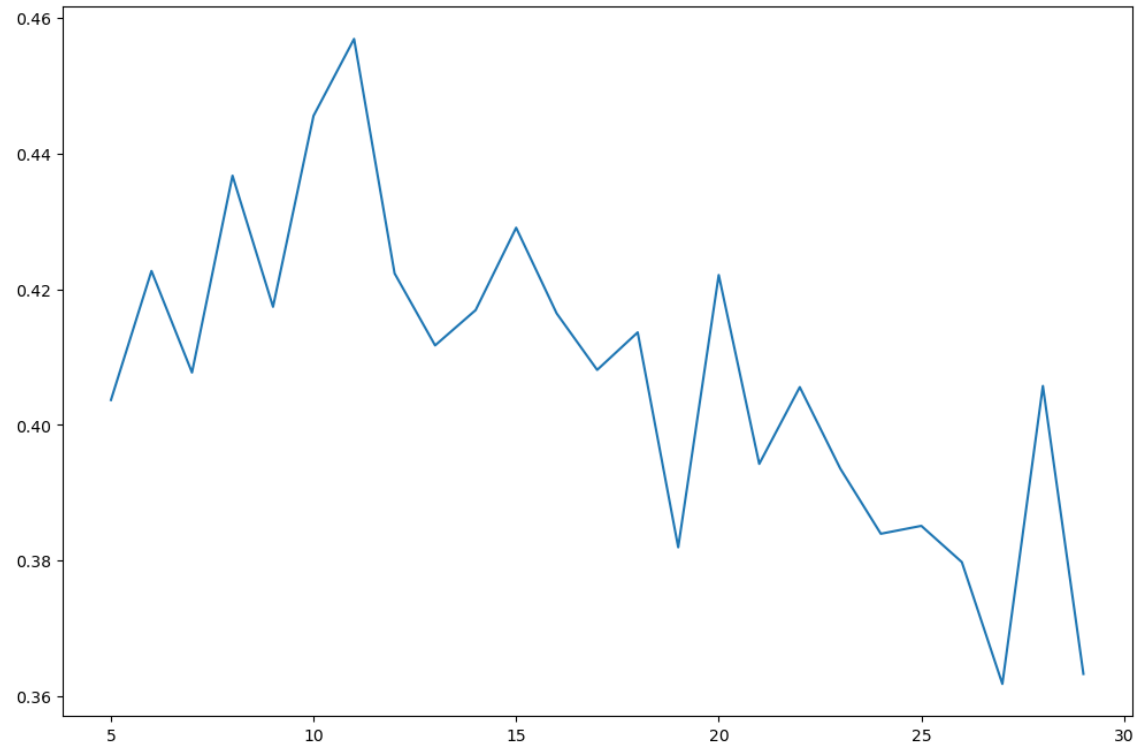

Figura 4 Gráfico de Coherencia de tópicos

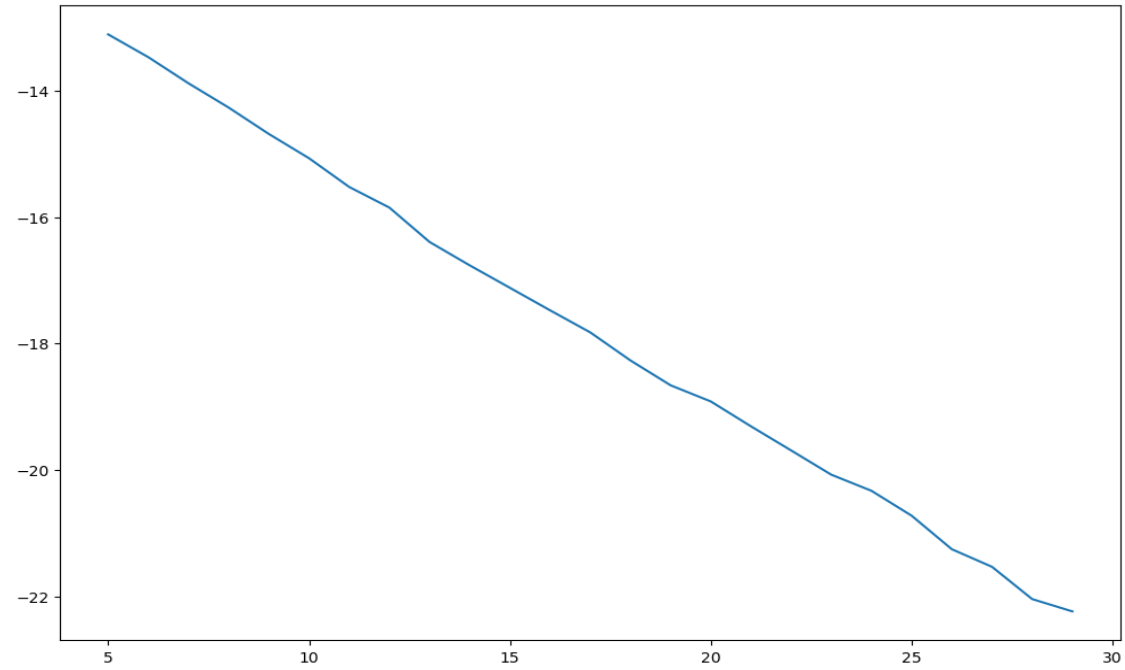

Figura 5 Gráfico de perplejidad

Tabla 2 Coherencia y Perplejidad

Número de Tópicos	Coherencia	Perplejidad
10	0.4544680321686836	-9.218708531439786
11	0.4339181219591785	-9.554720405540268
12	0.3987753616947329	-9.965868334037705
13	0.4377247925594714	-10.382386974157185
14	0.4049816786927236	-10.693116753636716
15	0.41613261895724746	-10.902109704371519

Fuente: Elaboración propia.

Elección del modelo

Observamos que el puntaje de coherencia más alto es de 0.4545 para el modelo de 10 tópicos. Este resultado sugiere una moderada coherencia en los tópicos identificados. Observamos que la perplejidad es -9.2187, y, en general, valores bajos son deseables, indicando que el modelo tiene una buena capacidad predictiva.

Por lo tanto, el modelo escogido fue el de 10 tópicos.

Visualización con pyLDAvis

Se puede apreciar en la Figura 6 que la mayoría de los tópicos son independientes entre sí (separación entre los distintos tópicos es grande) excepto los que pyLDAvis identifica como tópico 10, 7, 5 y 4.

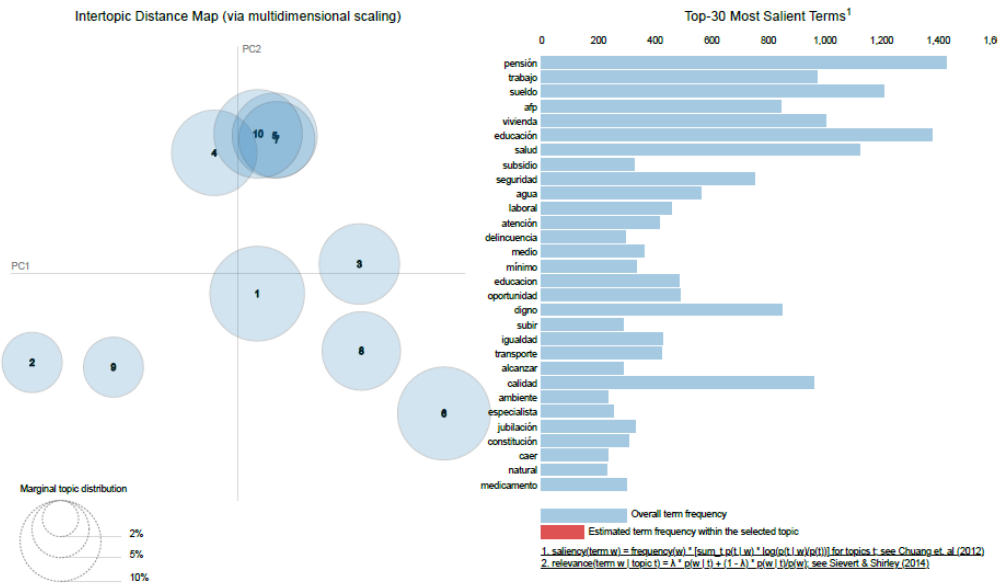

Figura 6 Visualización con pyLDAvis

Etiquetado de Tópicos

En la Tabla 3 se imprimieron las 15 Palabras de mayor probabilidad para cada tópico (Se obtuvo a través de la matriz de tópicos por palabras). Esto significa que, por ejemplo, las

palabras de mayor probabilidad para el tópico 1 fueron: salud, atención, mejorar, medicamento, especialista, mayor, calidad, espera, mejor, público, hospital, lista, falta, trato y malo. Por lo tanto, está claro que este tópico se centra en el Servicio de Salud Pública y Atención Médica.

Tabla 3 Palabras de mayor probabilidad para cada tópico

Tópico	***Palabras de mayor probabilidad para cada tópico***
Tópico 0	salud, atención, mejorar, medicamento, especialista, mayor, calidad, espera, mejor, público, hospital, lista, falta, trato, malo
Tópico 1	extranjero, zona, calle, libre, mejoramiento, funcionar, emprendimiento, tolerancia, escaso, indigno, seguro, honorario, seguridad, sequía, barrio
Tópico 2	trabajo, laboral, oportunidad, transporte, mujer, empleo, mayor, sueldo, fuente, haber, rural, estabilidad, hombre, condición, mejor
Tópico 3	seguridad, delincuencia, constitución, justicia, respeto, ley, ciudadano, cambio, nuevo, respetar, autoridad, política, haber, mano, protección
Tópico 4	educación, calidad, educacion, caer, gratuidad, profesor, colegio, gratuito, mejorar, niño, eliminar, mejor, universidad, cultura, firme
Tópico 5	sueldo, pensión, vida, bajo, digno, mínimo, subir, alcanzar, costo, vivir, minimo, jubilación, mejorar, edad, luz
Tópico 6	educación, igualdad, derecho, equidad, desigualdad, gratis, oportunidad, social, calidad, deber, acceso, pobre, tener, carretera, mismo

Tópico 7	pensión, afp, sistema, digno, adulto, mayor, carabinero, mejorar, eliminar, jubilación, fondo, cambiar, modificar, justo, acorde
Tópico 8	subsidio, vivienda, igual, estudio, burocracia, terreno, entrega, sacar, burla, campo, institución, aplicar, x, gestión, sustentable
Tópico 9	vivienda, agua, medio, recurso, ambiente, natural, clase, educación, acceso, calidad, derecho, beneficio, tener, casa, fiscalización

Fuente: Elaboración propia.

Para complementar el resultado obtenido en la Tabla 3, a continuación, podemos visualizar las palabras más importantes de cada tópico como nubes de palabras:

Figura 7 WordCloud Tópico 0

Figura 8 WordCloud Tópico 1

Figura 9 WordCloud Tópico 2

Figura 10 WordCloud Tópico 3

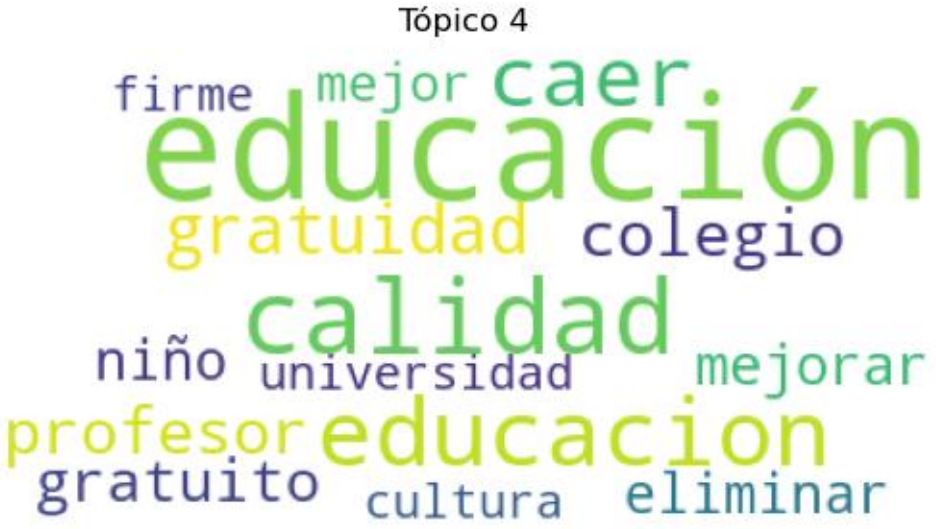

Figura 11 WordCloud Tópico 4

Figura 12 WordCloud Tópico 5

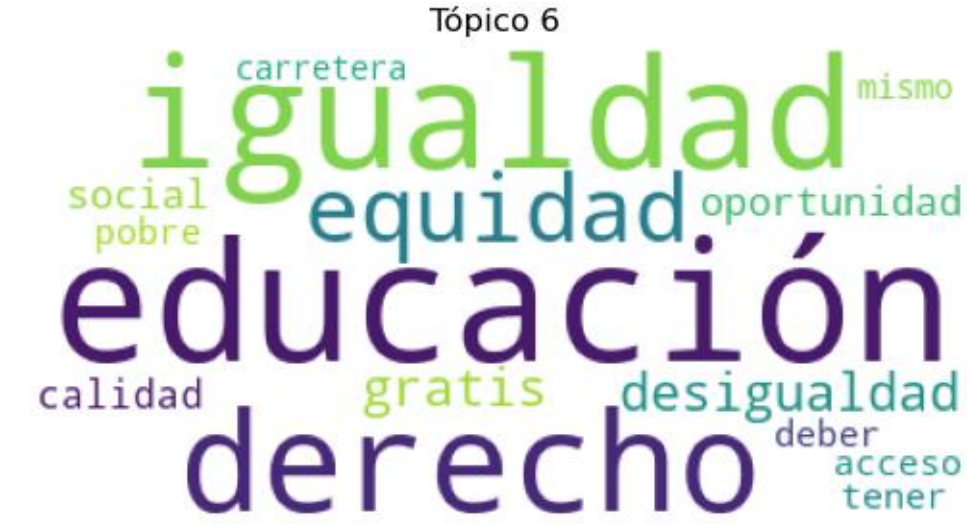

Figura 13 WordCloud Tópico 6

Figura 14 WordCloud Tópico 7

Figura 15 WordCloud Tópico 8

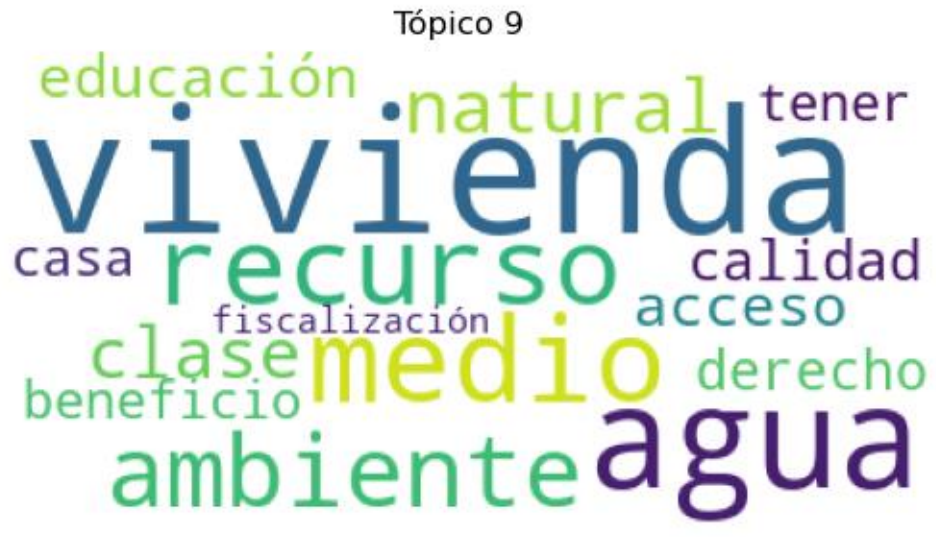

Figura 16 WordCloud Tópico 9

Por lo tanto, a partir de estos resultados se etiquetaron los tópicos con la nomenclatura indicada en la tabla a continuación.

Tabla 4 Etiquetado de Tópicos

Tópico	*Etiqueta*
Tópico 0	Servicio de Salud Pública y Atención Médica
Tópico 1	Problemas Sociales
Tópico 2	Empleo y Trabajo
Tópico 3	Seguridad Ciudadana
Tópico 4	Educación
Tópico 5	Economía y Finanzas Personales

Tópico 6	Igualdad y Equidad
Tópico 7	Sistema de Pensiones
Tópico 8	Vivienda y Desarrollo Urbano
Tópico 9	Política Pública

Fuente: Elaboración propia.

Tópico Dominante

El tópico dominante es aquel que tiene la mayor probabilidad en el documento. Se obtuvo el tópico dominante para cada documento utilizando la matriz de probabilidad de temas dominantes y al extender el análisis del tópico más dominante al corpus total se obtuvieron los siguientes resultados:

Tabla 5 Porcentaje de cada Tópico en el corpus total

Tópico	*Porcentaje en el corpus total*
Tópico 0	16.47%
Tópico 1	1.80%
Tópico 2	9.73%
Tópico 3	11.86%
Tópico 4	12.01%
Tópico 5	16.20%
Tópico 6	7.88%

Tópico 7	9.32%
Tópico 8	1.65%
Tópico 9	13.07%

Fuente: Elaboración propia.

Por lo tanto, se puede observar que los tópicos 0, 5 y 9 presentan los valores más altos y de esta manera son los más dominantes del corpus total.

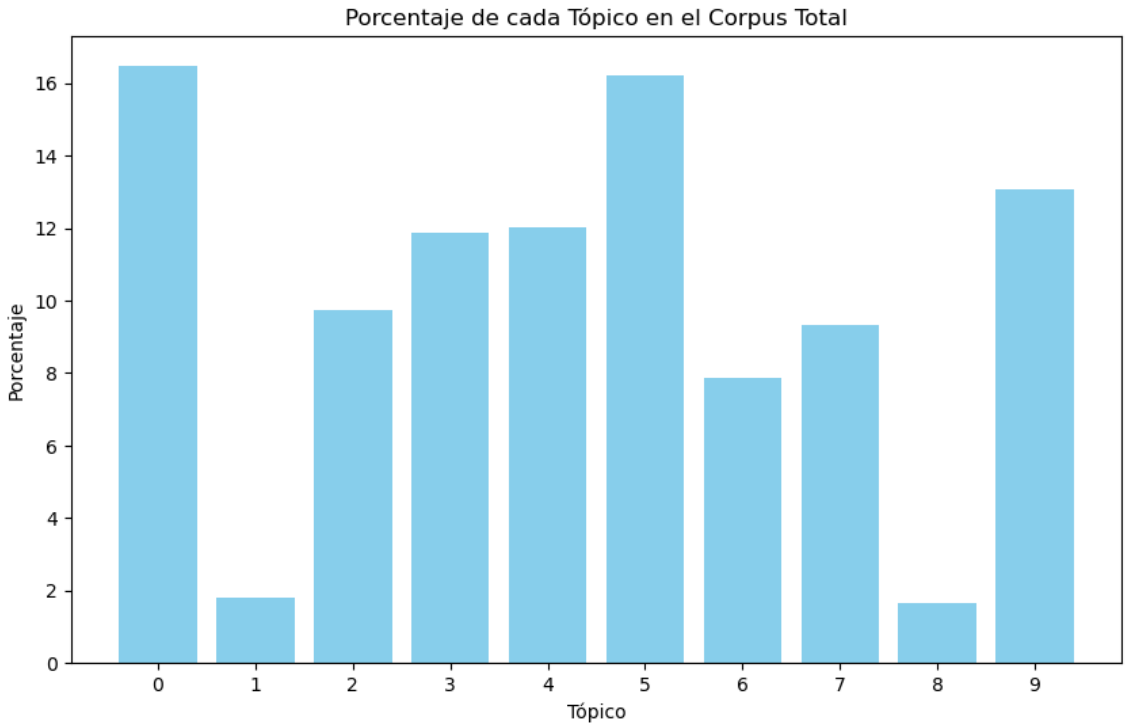

Figura 17 Porcentaje de cada tópico en el corpus total

Análisis de relación entre Temas, Emociones y Variables Demográficas

En el Anexo de este informe encuentran los resultados de todos los análisis en las relaciones entre los temas dominantes, las emociones y variables demográficas como la edad y el sexo.

A continuación, se examinan los aspectos y conclusiones principales producto del análisis.

Resultados Generales.

- Los tópicos 0 “Servicio de Salud Pública y Atención Médica” (16.47%), 5 “Economía y Finanzas Personales” (16.20%) y 9 “Política Pública” (13.07%) son los más dominantes del corpus total.
- Existe una mayor proporción de mujeres (58%) en comparación con los hombres (32%) en el corpus total. El porcentaje restante (10%) corresponde a los documentos en dónde el sexo no estaba identificado en la data.
- El miedo representa el 38% del corpus total, lo que indica una alta prevalencia de esta emoción en el conjunto de datos. Con un 17%, la tristeza también tiene una presencia significativa en el corpus, lo sigue el Enojo con un 15%, la Alegría con un 13%, la Aversión con un 10% y finalmente con solo un 2%, la sorpresa es la emoción menos representada en el corpus. El porcentaje restante (5%) corresponde a los documentos en dónde la categoría emoción no estaba identificada en la data.

- El porcentaje de cada sentimiento en el corpus total analizado reveló una tendencia mayoritaria hacia sentimientos negativos (81%) en el conjunto de datos examinado.
- La categoría de edad etiquetada como “Adultez” representa el 49% del corpus total, seguida de la Adultez Mayor con un 28% y luego la juventud con un 16%. El porcentaje restante (7%) corresponde a los documentos en dónde la categoría edad no estaba identificada en la data.

Relación entre Temas Dominantes en las respuestas de las principales necesidades del país y las emociones de los ciudadanos Post-octubre 2019.

En el análisis presentado en la sección 8.5 del Anexo se puede observar cómo se distribuyen las emociones en diferentes categorías de temas dominantes. Evidenciando que, aunque ciertos temas son dominantes en el corpus, pueden variar dependiendo de la categoría a la cual pertenece la emoción, destacando la presencia del tópico 4 "Educación" en la categoría de emociones de "Miedo" como un hallazgo interesante, ya que se posiciona como el tercer tópico más dominante en esta categoria desplazando de esta manera al tópico 9 “Política Pública”. Lo anterior se puede observar en la matriz de contingencia presentada en la Tabla 6, en dónde se expresa la contribución porcentual de cada tópico en las categorías de emociones (Los porcentajes se calcularon normalizando hacia el lado, por filas)

Por otro lado, se obtiene un coeficiente V de Cramér de 0.0294 lo que indica una asociación débil entre las variables “dominant_topic” y “categoria_emocion”

Tabla 6 Matriz de contingencia variable “categoria_emocion” y “dominant_topic” en porcentaje

dominant_topic	***0***	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***
categoria_emocion										
Alegría	15.09	1.91	9.66	12.35	11.16	16.04	8.67	8.79	1.88	14.46
Aversión	14.43	1.81	10.15	14.21	10.75	15.28	8.03	9.55	1.77	14.02
Enojo	16.52	1.73	9.99	10.66	11.91	17.39	7.42	10.03	1.48	12.87
Miedo	18.56	1.70	9.09	11.48	12.66	15.86	7.90	9.28	1.56	11.93
Sorpresa	16.47	1.74	10.17	12.50	10.95	15.41	9.59	9.21	1.84	12.11
Tristeza	15.44	1.92	9.68	11.70	12.48	16.71	7.88	9.45	1.69	13.05

Fuente: Elaboración propia.

Por otro lado, en la figura 18 se observa que la categoría de emoción “Miedo” concentra aproximadamente en promedio el 40% de los documentos en los tópicos.

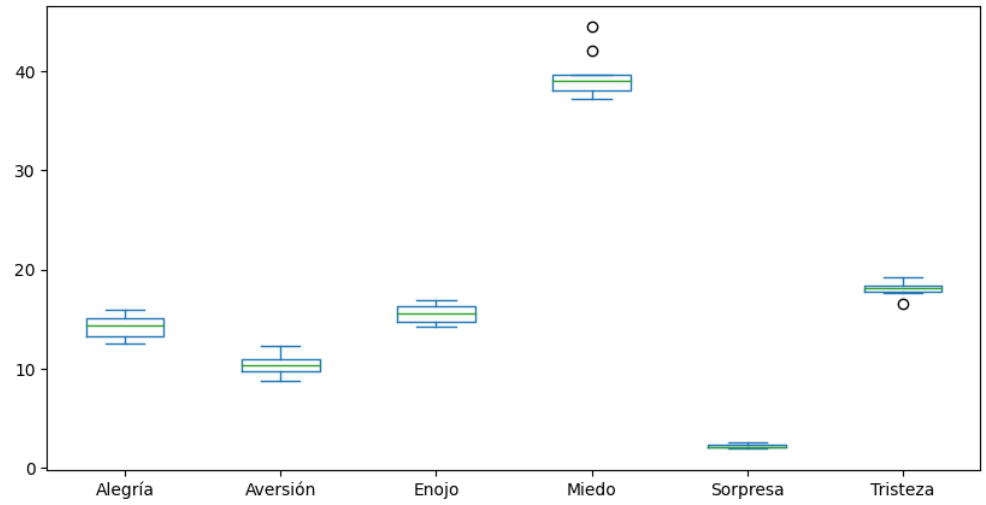

Figura 18 Boxplot variable "categoria_emocion" en análisis con la "variable dominant_topic"

A continuación, se analiza la categoría "Alegría" por ser la única que apunta a un sentimiento positivo y la categoría "Miedo" dada su dominancia en el corpus total.

Categoría Alegría

Al imprimir las 10 palabras más frecuentes de la categoría emoción "Alegría" podemos observar que "Esperanza" es la palabra más usada en las respuestas brindadas por los ciudadanos en esta categoría lo cual es consistente con los resultados de la Encuesta nacional realizada por la Universidad de Santiago en conjunto con la Asociación Chilena de Municipalidades (AChM) y la empresa de investigación de mercados Sargon el año

2019 titulada “Legitimidad, Miedo y Esperanza en el Chile post estallido” [2]en dónde se concluyó que Tristeza y Esperanza son las emociones más intensas de la ciudadanía con el estallido social. Por otro lado, en esta encuesta también se concluyó que la esperanza se concentra preferentemente en personas entre los 18 a 29 años lo cual es consistente con los resultados de la tabla 9 “Matriz de contingencia variable "Categoria_edad" y "categoria_emocion" en porcentaje”

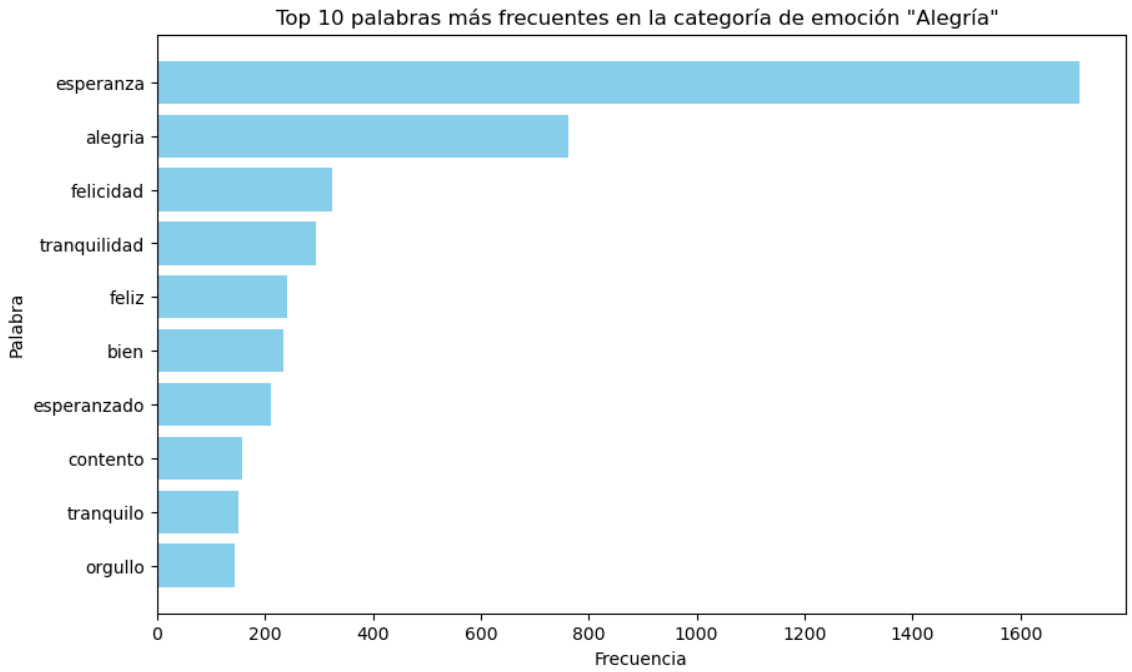

Figura 19 Top 10 palabras más frecuentes en la categoría de emoción "Alegría"

Los resultados de la tabla 6 confirman que los temas más relevantes para la categoría Alegría fueron:

[2] https://www.usach.cl/news/tristeza-y-esperanza-son-las-emociones-mas-intensas-la-ciudadania-estallido-social

- Tópicos 0 “Servicio de Salud Pública y Atención Médica”
- Tópico 5 “Economía y Finanzas Personales”
- Tópico 9 “Política Pública”

Lo cual es consistente con la información proporcionada en la columna de opinión publicada en Ciperchile titulada “Escuchando a los chilenos en medio del estallido: Liberación emocional, reflexividad y el regreso de la palabra pueblo” [3]en dónde a partir de un estudio con grupos focales quedó de manifiesto la fuerte expectativa de un cambio expresada de numerosas y múltiples maneras por los participantes, mencionando repetidas veces la palabra “esperanza”. Asimismo, los participantes expresaron aspiraciones específicas sobre sueldos e ingresos, educación y universidad, transporte público o jubilaciones.

Categoria Miedo

Al Imprimir las 10 palabras más frecuentes de la categoría emoción “Miedo” podemos observar que “Miedo” es la palabra más usada en las respuestas brindadas por los ciudadanos en esta categoría. En los resultados de la Encuesta nacional realizada por la Universidad de Santiago comentados anteriormente también se destaca el miedo como una de las emociones dominantes[4]. Por otro lado, la Columna de opinión publicada en Ciperchile también concluyó que el miedo es más agudo entre personas mayores lo cual

[3] https://www.ciperchile.cl/2020/03/02/escuchando-a-los-chilenos-en-medio-del-estallido-liberacion-emocional-reflexividad-y-el-regreso-de-la-palabra-pueblo/
[4] https://www.latercera.com/tendencias/noticia/las-emociones-del-estallido-social/952835/

es consistente con los resultados de la tabla 9 “Matriz de contingencia variable "Categoria_edad" y "categoria_emocion" en porcentaje”

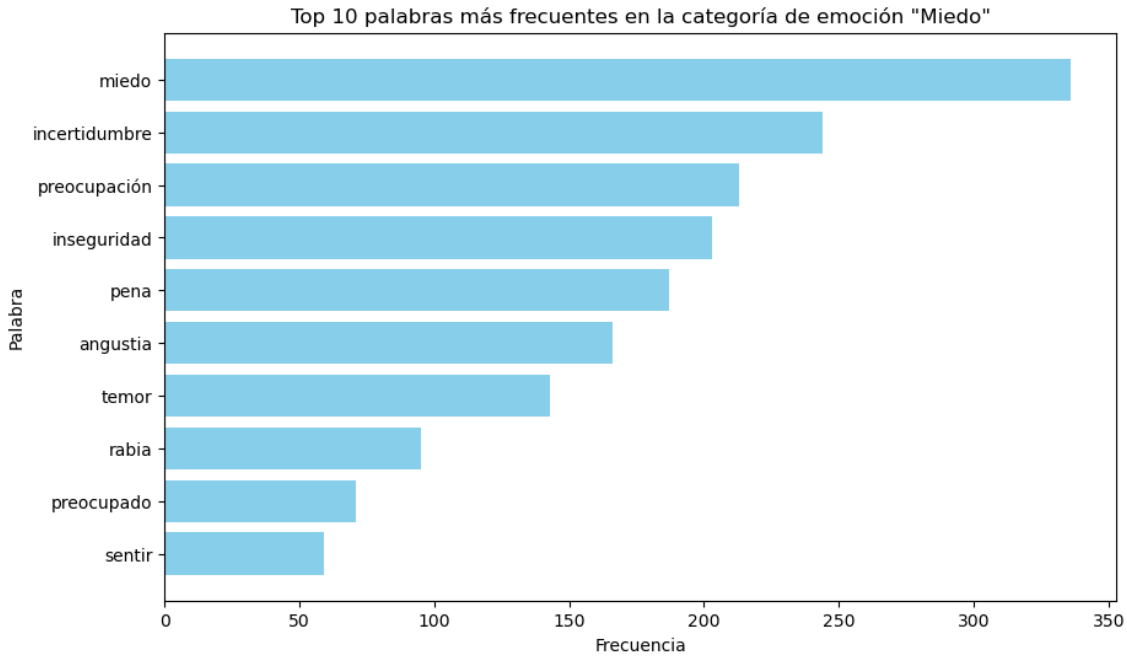

Figura 20 Top 10 palabras más frecuentes en la categoría de emoción "Miedo"

Relación entre Temas Dominantes y Edad:

En el análisis presentado en la sección 8.7 del Anexo se puede observar cómo se distribuyen las diferentes categorías de edad en los temas dominantes. Evidenciando que, aunque ciertos temas son dominantes en el corpus, pueden variar dependiendo la categoria de edad a la cual pertenece el documento, destacando la presencia del tópico 3 " Seguridad Ciudadana" en la categoría de edad “Adultez Mayor” como un hallazgo interesante, ya que se posiciona como el tercer tópico más dominante en esta categoria desplazando de esta manera al tópico 9 “Política Pública”. Lo anterior se puede observar en la matriz de

contingencia presentada en la Tabla 7, en dónde se expresa la contribución porcentual de cada tópico en las categorías de edades (Los porcentajes se calcularon normalizando hacia el lado, por filas)

Por otro lado, se obtiene un coeficiente V de Cramér de 0.0426 lo que indica una asociación débil entre las variables “dominant_topic” y “Categoria_edad”

Tabla 7 Matriz de contingencia variable "Categoria_edad" y "dominant_topic" en porcentaje

dominant_topic	***0***	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***
Categoria_edad										
Adultez	15.92	1.96	9.79	11.60	11.88	15.50	8.48	9.48	1.63	13.76
Adultez mayor	18.08	1.56	10.05	12.51	11.77	17.01	6.46	9.36	1.73	11.46
Juventud	15.47	1.71	8.94	11.08	12.89	17.13	8.46	9.16	1.65	13.53

Fuente: Elaboración propia.

Relación entre Sexo y Emociones:

En el análisis presentado en la sección 8.9 del Anexo se puede observar cómo se distribuyen las diferentes emociones entre los sexos. Evidenciando que, las emociones, como el miedo, tienen una distribución porcentualmente mayor en el sexo femenino.

Por otro lado, se obtiene un coeficiente V de Cramér de 0.0287 lo que indica una asociación débil entre las variables.

Tabla 8 Matriz de contingencia variable “LP_SEXO" y "categoria_emocion" en porcentaje

categoria_emocion	*Alegría*	*Aversión*	*Enojo*	*Miedo*	*Sorpresa*	*Tristeza*
LP_SEXO						
F	13.780727	9.716482	15.267203	41.237039	1.994054	18.004496
M	14.517694	10.247784	15.831015	38.752669	2.497250	18.153587

Relación entre Edad y Emociones:

En el análisis presentado en la sección 8.10 del Anexo se puede observar el análisis realizado entre estas dos variables. Evidenciando que, el porcentaje de documentos en la categoría emoción “miedo” es porcentualmente mayor en los adultos mayores, mientras que en los jóvenes es menor. Por otro lado, se evidencia que el porcentaje de documentos en la categoría emoción “Alegría” es porcentualmente mayor en los jóvenes (17%), mientras que en los Adultos mayores (12%) es menor.

Por otro lado, se obtiene un coeficiente V de Cramér de 0.0553 lo que indica una asociación débil entre las variables.

Tabla 9 Matriz de contingencia variable "Categoria_edad" y "categoria_emocion" en porcentaje

categoria_emocion	*Alegría*	*Aversión*	*Enojo*	*Miedo*	*Sorpresa*	*Tristeza*
Categoria_edad						
Adultez	14.259752	10.286970	15.561472	40.015215	2.235747	17.640844
Adultez mayor	11.910966	8.948113	15.201370	43.891908	1.674980	18.372664
Juventud	16.633871	11.412830	17.447199	34.251607	2.754821	17.499672

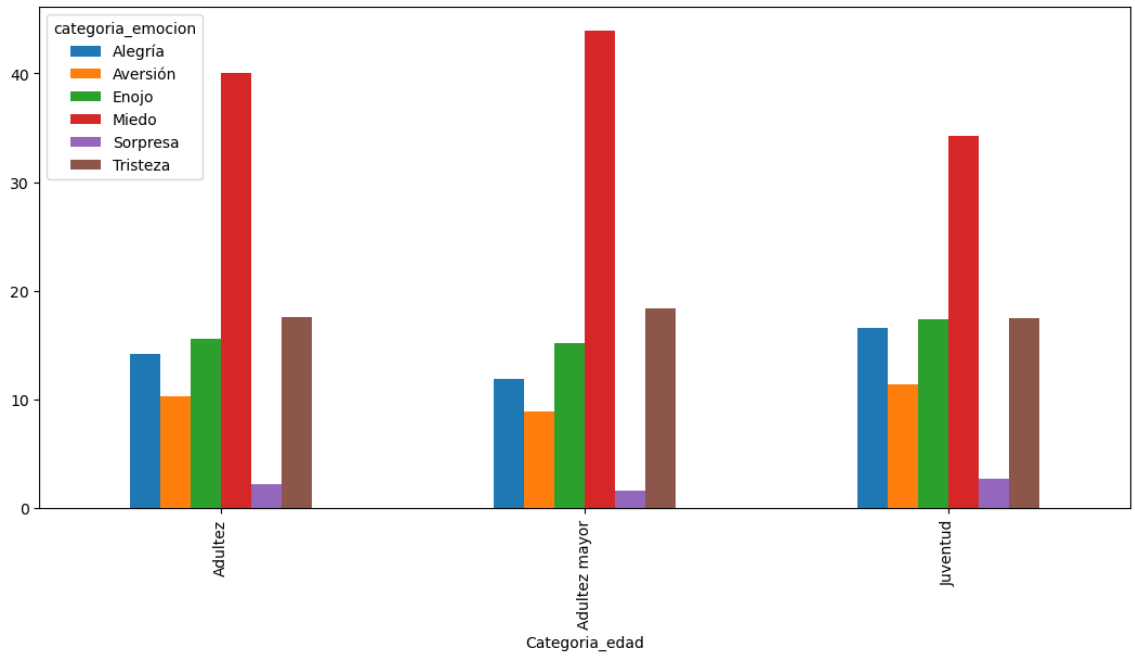

Figura 21 Gráfico de barras Análisis variables "Categoria_edad" y "categoria_emocion"

Evaluación del modelo

Para evaluar el rendimiento del modelo, se imprimirá las respuesta brindada por una persona con respecto a sus emociones y las necesidades del país en un documento aleatorio y posteriormente se estimarán sus tópicos más dominantes.

Documento Aleatorio Número 88

Los resultados de la figura 22 presentan a los tópicos 2 “Empleo y Trabajo”, 5 “Economía y Finanzas Personales” y 4 “Educación” como los tópicos más relevantes en el documento aleatorio número 88.

Se imprimen las principales palabras de cada tópico con el objetivo de determinar si los tópicos identificados por el modelo fueron consistentes con el documento aleatorio.

Tópico 2 “Empleo y Trabajo”: trabajo, laboral, oportunidad, transporte, mujer, empleo, mayor, sueldo, fuente, haber, rural, estabilidad, hombre, condición, mejor
Tópico 4 “Educación”: educación, calidad, educación, caer, gratuidad, profesor, colegio, gratuito, mejorar, niño, eliminar, mejor, universidad, cultura, firme
Tópico 5 “Economía y Finanzas Personales”: sueldo, pensión, vida, bajo, digno, mínimo, subir, alcanzar, costo, vivir, mínimo, jubilación, mejorar, edad, luz.

Al analizar estos resultados con la tabla 10 se concluye.

- Los tópicos 2 (44%) y 5 (34%) tienen un alto grado de coherencia con la respuesta de la principal necesidad del país, puesto que se plantea una mejor distribución de riqueza e igualdad social lo que puede traducirse en mejores sueldos, pensiones y mejores oportunidades laborales palabras que pueden observarse forman parte de los tópicos mencionados.
- Por otro lado, vemos que el tópico 4 se presenta en un 0.3%, por lo cual no presenta una mayor contribución al documento.
- Se observa que en la emoción se manifiesta esperanza y optimismo ante la posibilidad de lograr una mayor equidad social, por lo tanto, vemos que la emoción y la respuesta a las necesidades del país en dicho documento aleatorio guardan una estrecha relación.

Tabla 10 Evaluación del Modelo Esperanza - Igualdad Social

Documento Aleatorio:	88
Emoción de la persona:	Esperanza a partir del 18 de octubre , hay un optimismo por ver una nueva oportunidad de cambios que garanticen la equidad social
Necesidades del país:	Igualdad social lograr mayor empatía por parte del gobierno y autoridades para realizar cambios y leyes que garanticen una mejor distribución de la riqueza

```
Número documento aleatorio: 88

Tópicos más importantes del documento aleatorio:
Tópico 2: 0.44
Tópico 5: 0.34
Tópico 4: 0.03
```

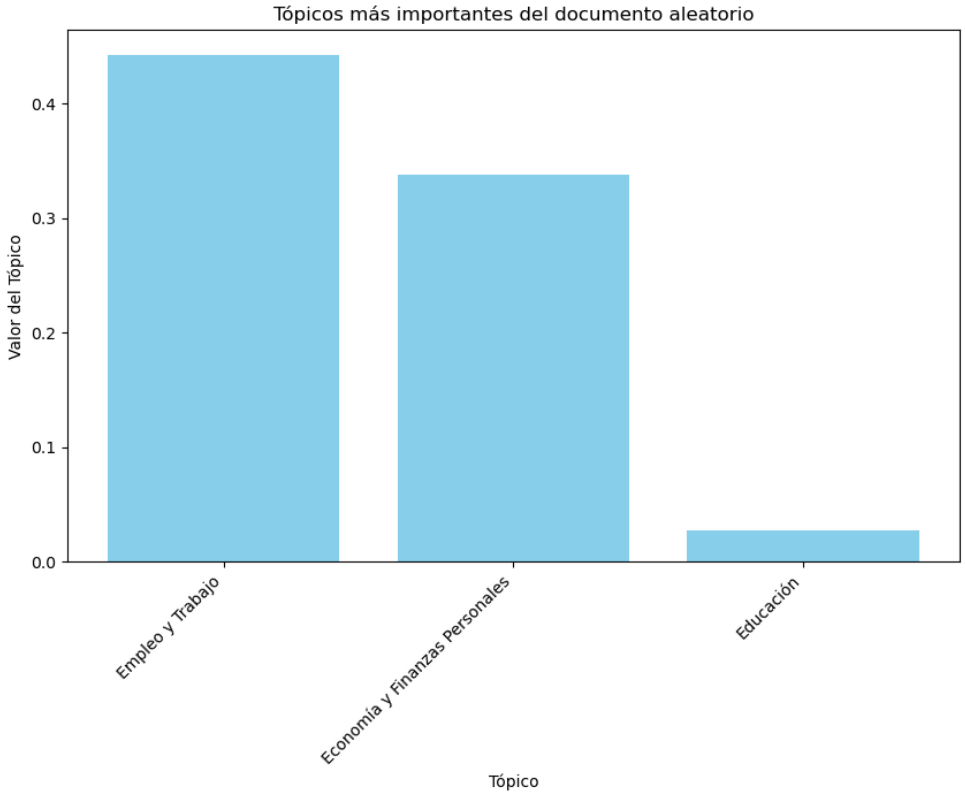

Figura 22 Evaluación del Modelo Esperanza - Igualdad Social

6. Conclusiones

La realización de este trabajo se ha centrado en identificar temas relevantes que reflejen las principales necesidades del país en las respuestas proporcionadas por los ciudadanos que participaron en la iniciativa “El Chile que Queremos” post octubre de 2019 y luego observar si es que existe una relación entre las principales necesidades país y las emociones experimentadas durante las últimas semanas posteriores al estallido social.

Para ello se utilizó información extraída desde el GitHub del Ministerio de Ciencia, Tecnología, Conocimiento, e Innovación.

Para llevar a cabo el desarrollo de este trabajo se realizó un modelado de temas aplicando el método de clasificación no supervisada LDA (Latent Dirichlet Allocation)

De los resultados obtenidos, se observó que el que el puntaje de coherencia más alto fue de 0.4545 para el modelo de 10 tópicos, lo cual indicó una moderada coherencia en los tópicos identificados con respecto a las necesidades del país.

Las percepciones ciudadanas se centraron en áreas tales como: “Servicio de Salud Pública y Atención Médica”, “Problemas Sociales”, “Empleo y Trabajo”, “Seguridad Ciudadana”, “Educación”, “Economía y Finanzas Personales”, “Igualdad y Equidad”, “Sistema de

Pensiones", "Vivienda y Desarrollo Urbano" y "Política Pública". Esto se evidenció a través de la obtención de las palabras de mayor probabilidad para cada tópico

Las temáticas anteriormente identificadas fueron consistentes con la información proporcionada en la columna de opinión publicada en Ciperchile titulada "Escuchando a los chilenos en medio del estallido: Liberación emocional, reflexividad y el regreso de la palabra pueblo"

El miedo representó el 38% del corpus total lo cual indicó una alta prevalencia de esta emoción en el conjunto de datos. Asimismo, también se evidenció la categoría miedo fue más aguda entre personas mayores.

Se analizó la categoría "Alegría" por ser la única que apuntaba a un sentimiento positivo en dónde se pudo observar que la esperanza fue una de las palabras más frecuentes en esta categoría. Asimismo, también se evidencio que la categoría Alegría fue más aguda en los Jóvenes.

Los resultados anteriores fueron igualmente consistentes con los resultados de la Encuesta nacional realizada por la Universidad de Santiago en conjunto con la Asociación Chilena de Municipalidades (AChM) y la empresa de investigación de mercados Sargon el año 2019 titulada "Legitimidad, Miedo y Esperanza en el Chile post estallido"

A través del cálculo del coeficiente V de Cramér se evidenció una asociación débil entre temas dominantes y categorías de emociones, debido a la heterogeneidad presente en estas categorías.

El estudio tuvo ciertas restricciones. La principal se presentó al momento de seleccionar las variables de análisis, puesto que la data no estaba balanceada, por lo tanto, este punto impidió extender el análisis hacia otras variables como por ejemplo el nivel educacional de las personas que participaron en la iniciativa ECQQ.

Sería interesante que se realizaran trabajos futuros en dónde se profundizara acerca de las percepciones que personas de diferentes edades manifiestan con respecto a cada una de las temáticas identificadas, puesto que a partir del análisis se identificó que entre los jóvenes es más frecuente identificarse con una emoción positiva con respecto a la población adulta y adulta mayor. Mientras tanto que por ejemplo al analizar los resultados de la categoría miedo vemos que este resultado es totalmente opuesto.

7. Bibliografía

1. El Mostrador. (2020, 9 de marzo). Según informe de la OCDE: Chile es uno de los tres países latinoamericanos más desiguales en cuanto a ingresos. El Mostrador. https://www.elmostrador.cl/noticias/2020/03/09/segun-informe-de-la-ocde-chile-es-uno-de-los-tres-paises-latinoamericanos-mas-desiguales-en-cuanto-a-ingresos/
2. Riffo-Pavón, I., Basulto, Ó., & Segovia, P. (2021). El Estallido Social chileno de 2019: un estudio a partir de las representaciones e imaginarios sociales en la prensa. Revista mexicana de ciencias políticas y sociales, 66(243), 345-368. https://www.scielo.org.mx/scielo.php?pid=S0185-19182021000300345&script=sci_arttext
3. Demandas prioritarias y propuestas para un Chile diferente: Sistematización de 1233 cabildos ciudadanos (2021). Disponible en https://repositorio.uchile.cl/handle/2250/178886
4. Raveau, M.P., Couyoumdjian, J.P., Fuentes-Bravo, C. y Candia, C. 2023. Consideraciones sobre la democracia deliberativa y lecciones del caso chileno. Estudios Públicos. 171 (sep. 2023), 9-40. https://www.estudiospublicos.cl/index.php/cep/article/view/2212
5. IBM. Natural Language Processing. https://www.ibm.com/es-es/topics/natural-language-processing
6. Amazon Web Services. What is PLN?

https://aws.amazon.com/es/what-is/PLN/#:~:text=tareas%20de%20PLN%3F-,%C2%BFQu%C3%A9%20es%20la%20PLN%3F,y%20comprender%20el%20lenguaje%20humano

7. Kaggle. Bag of Words vs TF-IDF. https://www.kaggle.com/code/fernandobordi/fb-bolsa-palabras-vs-tfidf
8. Alaminos-Fernández, A. F. (2023). Introducción a la minería de texto y análisis de sentimiento con R. https://rua.ua.es/dspace/bitstream/10045/133098/1/Introduccion_a_la_mineria_de_texto_y_analisis_de_sentimiento_con_R.pdf
9. Stemming and Lemmatization - SEO North https://seonorth.ca/es/PLN/stemming-and-lemmatization/
10. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022. https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf?ref=http://githubhelp.com
11. KeepCoding. ¿Qué es el modelo LDA? https://keepcoding.io/blog/que-es-el-modelo-lda/
12. QuestionPro. Prueba de chi-cuadrado: ¿Qué es y cómo se realiza? https://www.questionpro.com/blog/es/prueba-de-chi-cuadrado-de-pearson/
13. Statologos. (2021, 7 de mayo). Cómo calcular la V de Cramer en R. https://statologos.com/cramers-v-en-r/

14. Huawei. Aprendizaje supervisado y no supervisado. https://forum.huawei.com/enterprise/es/aprendizaje-supervisado-y-no-supervisado/thread/667228966026625025-667212895009779712
15. Tutz.tv. Módulo os en Python. https://tutz.tv/python/os
16. Matplotlib. Matplotlib: visualización con Python. https://es.matplotlib.net/
17. DataScientest. Matplotlib: todo lo que tienes que saber sobre la librería. https://datascientest.com/es/todo-sobre-matplotlib
18. DataScientest. Seaborn: todo sobre la herramienta de Data Visualization Python. https://datascientest.com/es/seaborn-la-herramienta-de-data-visualization-python
19. Universidad de Alcalá. Master Data Scientist. Pandas: herramienta básica para el Data Science en Python. https://www.master-data-scientist.com/pandas-herramienta-data-science/
20. DataScientest. NumPy: La biblioteca de Python más utilizada en Data Science. https://datascientest.com/es/numpy-la-biblioteca-python
21. Python Software Foundation. math - Funciones matemáticas. https://docs.python.org/es/3/library/math.html
22. Mastering NLTK: Your Ultimate Guide to Python's NLP Toolkit (2023, October 16). DataScientest. https://datascientest.com/en/mastering-nltk-your-ultimate-guide-to-pythons-nlp-toolkit.
23. CodigosPython. (2023, octubre 4). Explorando el Módulo "itertools" en Python. https://codigospython.com/itertools-explorando-el-modulo-itertools-en-python/

24. Python Software Foundation. re - Operaciones con expresiones regulares. https://docs.python.org/es/3/library/re.html
25. Python Software Foundation. Common string operations https://docs.python.org/es/3/library/string.html
26. La Biblia de la IA - The Bible of AI™ Journal Stanza – una biblioteca de Python NLP para muchos idiomas humanos. https://editorialia.com/2020/03/26/stanza-una-biblioteca-de-python-nlp-para-muchos-idiomas-humanos/.
27. Datascientest. (2024, January 7). Gensim: The Python library for topic modelling. https://datascientest.com/en/gensim-the-python-library-for-topic-modelling
28. pyLDAvis. PyPI. https://pypi.org/project/pyLDAvis/
29. DataScientest. ¿Cómo generar un Wordcloud con Python? https://datascientest.com/es/como-generar-un-wordcloud-con-python
30. Unipython. SciPy: Funciones principales. https://unipython.com/scipy-funciones-principales/
31. Sandoval, M. (2007). Caracterización de la juventud chilena actual. https://biblioteca.clacso.edu.ar/Chile/ceju/20120913094504/sandov.pdf.
32. USACH. (2020, 9 enero). Tristeza y esperanza son las emociones más intensas de la ciudadanía con el estallido social. Universidad de Santiago de Chile. https://www.usach.cl/news/tristeza-y-esperanza-son-las-emociones-mas-intensas-la-ciudadania-estallido-social

33. La Tercera. (2019, 27 de diciembre). Las emociones del estallido social. https://www.latercera.com/tendencias/noticia/las-emociones-del-estallido-social/952835/
34. (Mac-Clure, O., Barozet, E., Conejeros, J., & Jordana, C. (2020). Escuchando a los chilenos en medio del estallido: Liberación emocional, reflexividad y el regreso de la palabra “pueblo”. CIPER Académico, 2, 2020.) https://www.ciperchile.cl/2020/03/02/escuchando-a-los-chilenos-en-medio-del-estallido-liberacion-emocional-reflexividad-y-el-regreso-de-la-palabra-pueblo

8. Anexos

8.1 Porcentaje de cada sexo en el corpus total

La tabla 11 proporciona el porcentaje de cada sexo en el corpus total analizado. Estos porcentajes indican la distribución relativa de los sexos dentro del corpus utilizado para el análisis. En este caso, hay una mayor proporción de mujeres en comparación con los hombres en el corpus total.

Tabla 11 Porcentaje de cada sexo en el corpus total

LP_SEXO	Porcentaje en el corpus total
F	58%
M	32%

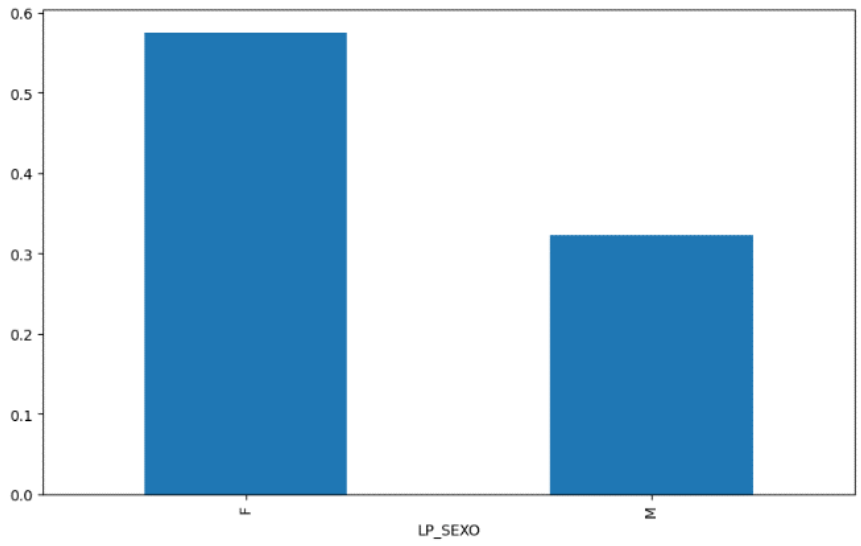

Figura 23 Porcentaje de cada sexo en el corpus total

8.2 Porcentaje de cada categoría de emoción en el corpus total

La tabla 12 muestra el porcentaje de cada categoría de emoción en el corpus total analizado. Podemos observar que miedo representa el 38% del corpus total, lo que indica una alta prevalencia de esta emoción en el conjunto de datos. Con un 17%, la tristeza también tiene una presencia significativa en el corpus. Por otro lado, con solo un 2%, la sorpresa es la emoción menos representada en el corpus.

Tabla 12 Porcentaje de cada categoría de emoción en el corpus total

categoria_emocion	Porcentaje en el corpus total
Alegría	13%
Aversión	10%
Enojo	15%
Miedo	38%
Sorpresa	2%
Tristeza	17%

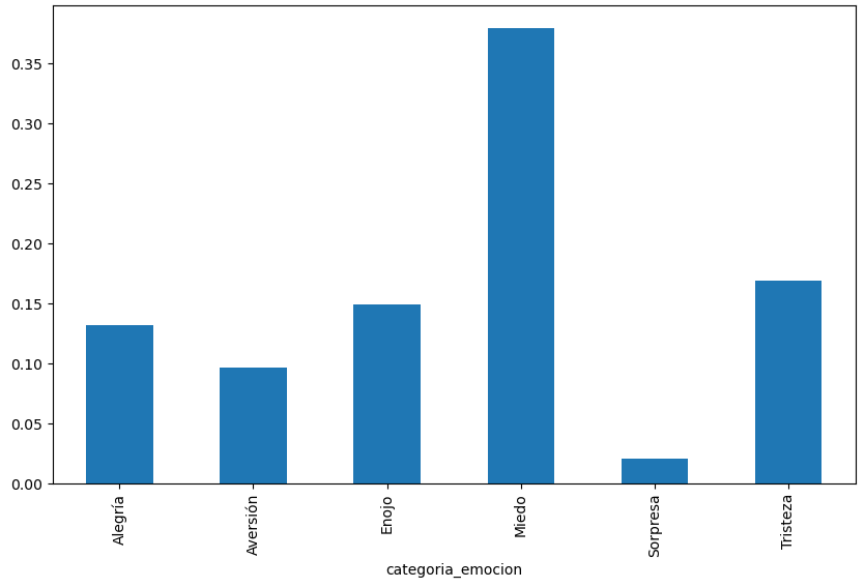

Figura 24 Porcentaje de cada categoría de emoción en el corpus total

8.3 Porcentaje de cada sentimiento en el corpus total

La tabla 13 muestra el porcentaje de cada sentimiento en el corpus total analizado. Estos resultados revelan una tendencia mayoritaria hacia sentimientos negativos en el conjunto de datos examinado.

Tabla 13 Porcentaje de cada sentimiento en el corpus total

sentimiento	Porcentaje en el corpus total
negativo	81%
positivo	13%

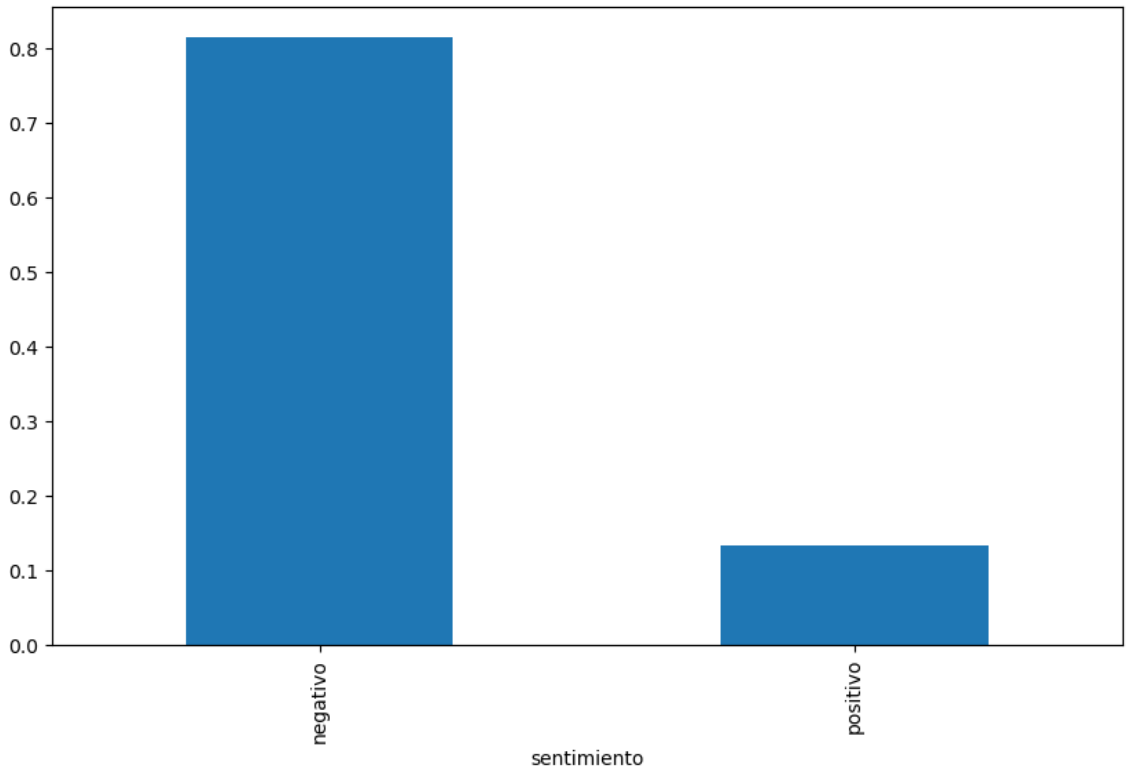

Figura 25 Porcentaje de cada sentimiento en el corpus total

8.4 Porcentaje de cada categoría de edad en el corpus total

La tabla 14 presenta el análisis de la distribución porcentual de diferentes grupos de edad en un corpus de datos. Se observa que la “Adultez” Representa el 49% del corpus total, indicando una presencia significativa de individuos en la categoría de adultez.

Estos resultados revelan la composición del corpus según diferentes grupos de edad, destacando la preponderancia de la adultez y la adultez mayor.

Tabla 14 Porcentaje de cada categoría de edad en el corpus total

Categoria_edad	*Porcentaje en el corpus total*
Adultez	49%
Adultez mayor	28%
Juventud	16%

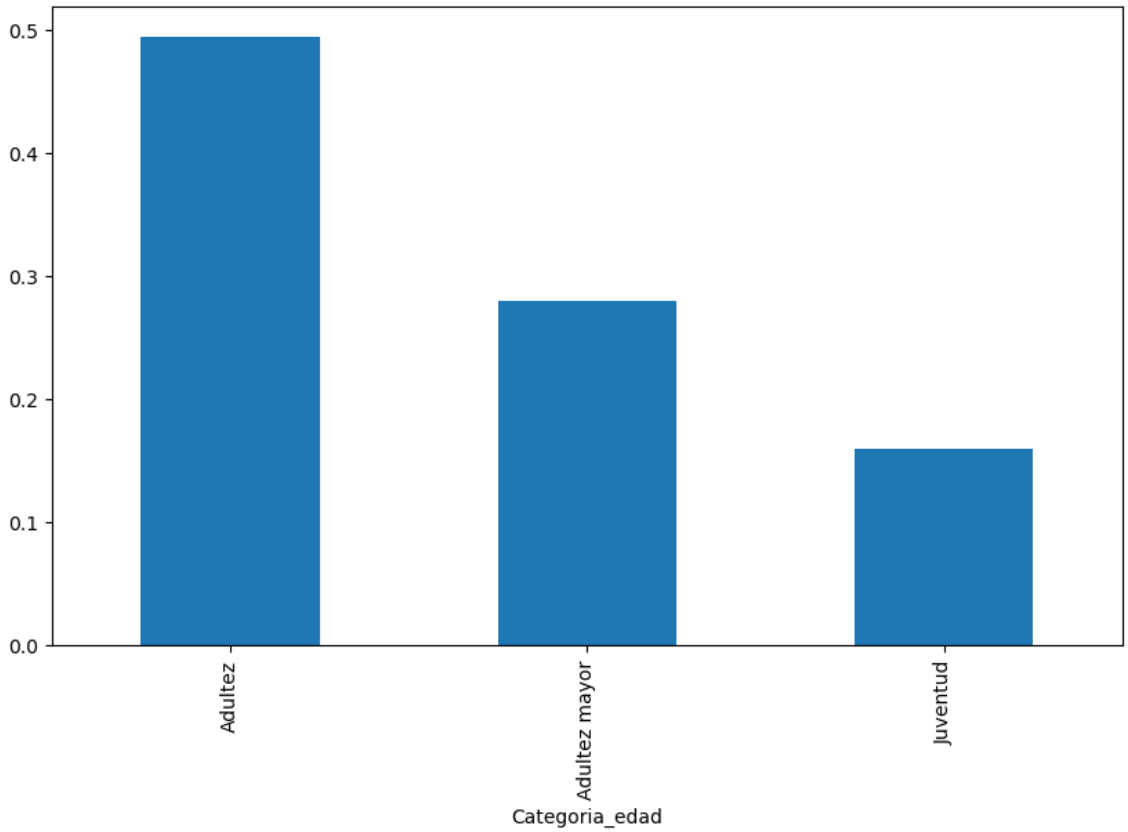

Figura 26 Porcentaje de cada categoría de edad en el corpus total

8.5 Análisis Variable “dominant_topic” y “categoria_emocion”

El análisis de la tablas 15 y 16 muestran la distribución de temas dominantes en diferentes categorías de emociones y la tabla 17 el análisis a la inversa. Se Observa que la categoría de emoción “Miedo” concentra aproximadamente en promedio el 40% de los documentos en los tópicos. Por otro lado, se obtiene un coeficiente V de Cramér de 0.0294 lo que indica una asociación débil entre las variables.

En la figura 28 observamos que los tópicos 0, 5 y 9 son los más dominantes, lo cual es consistente con los resultados de los tópicos más relevantes en el corpus, sin embargo, al observar la tabla 17 vemos que al analizar los tópicos más relevantes para la emoción de miedo esta tendencia cambia, ya que el tópico 4 “Educación” se posiciona como el tercer tópico más dominante en esta categoria.

Tabla 15 Matriz de contingencia variable “dominant_topic” y “categoria_emocion”

categoria_emocion	*Alegría*	*Aversión*	*Enojo*	*Miedo*	*Sorpresa*	*Tristeza*
dominant_topic						
0	1005	701	1239	3543	170	1316
1	127	88	130	324	18	164
2	643	493	749	1735	105	825
3	822	690	799	2191	129	997
4	743	522	893	2418	113	1064
5	1068	742	1304	3028	159	1424
6	577	390	556	1508	99	672
7	585	464	752	1771	95	805
8	125	86	111	297	19	144
9	963	681	965	2277	125	1112

Tabla 16 Matriz de contingencia variable “dominant_topic” y “categoria_emocion” en porcentaje

categoria_emocion	*Alegría*	*Aversión*	*Enojo*	*Miedo*	*Sorpresa*	*Tristeza*
dominant_topic						
0	12.603461	8.791071	15.537998	44.431904	2.131929	16.503637

1	14.923619	10.340776	15.276146	38.072855	2.115159	19.271445
2	14.131868	10.835165	16.461538	38.131868	2.307692	18.131868
3	14.605544	12.260128	14.196873	38.930348	2.292111	17.714996
4	12.915001	9.073527	15.522336	42.030245	1.964193	18.494698
5	13.825243	9.605178	16.880259	39.197411	2.058252	18.433657
6	15.176223	10.257759	14.623882	39.663335	2.603893	17.674908
7	13.081395	10.375671	16.815742	39.601968	2.124329	18.000894
8	15.984655	10.997442	14.194373	37.979540	2.429668	18.414322
9	15.727585	11.121999	15.760248	37.187653	2.041483	18.161032

Tabla 17 Matriz de contingencia variable “categoria_emocion” y “dominant_topic” en porcentaje

dominant_topic	***0***	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***
categoria_emocion										
Alegría	15.094623	1.907480	9.657555	12.346050	11.159507	16.040853	8.666266	8.786422	1.877441	14.463803
Aversión	14.432777	1.811818	10.150299	14.206300	10.747375	15.276920	8.029648	9.553222	1.770640	14.021001
Enojo	16.524407	1.733796	9.989330	10.656175	11.909843	17.391304	7.415311	10.029341	1.480395	12.870099
Miedo	18.557511	1.697046	9.087576	11.476011	12.664991	15.860046	7.898596	9.276137	1.555625	11.926461
Sorpresa	16.472868	1.744186	10.174419	12.500000	10.949612	15.406977	9.593023	9.205426	1.841085	12.112403
Tristeza	15.440573	1.924205	9.679690	11.697759	12.483867	16.707732	7.884548	9.445031	1.689546	13.047049

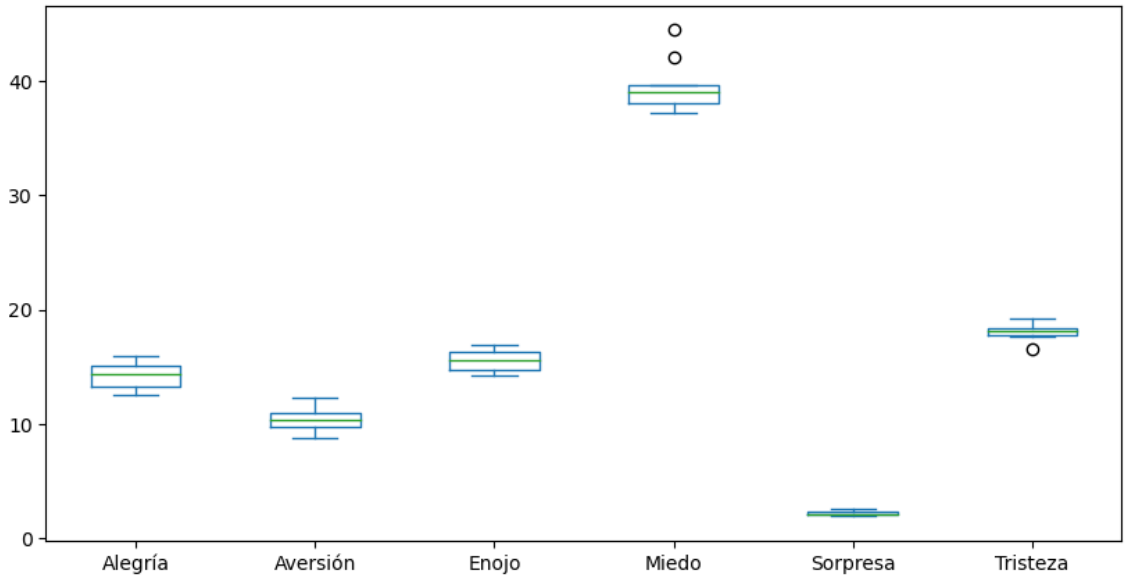

Figura 27 Boxplot variable "categoria_emocion" en análisis con la "variable dominant_topic"

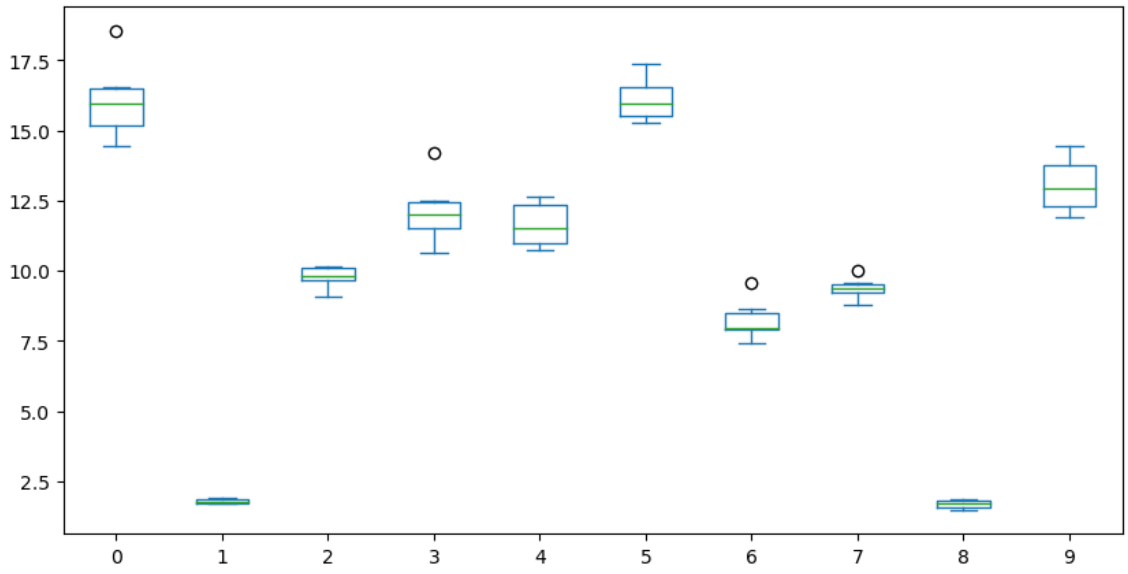

Figura 28 Boxplot variable "dominant_topic" en análisis con la "categoria_emocion"

8.6 Análisis Variable “dominant_topic” y “Sentimiento”

El análisis de la tablas 18 y 19 muestran la distribución de temas dominantes con respecto a los sentimientos y la tabla 20 el análisis a la inversa. Se observa que los sentimientos “negativos” concentran aproximadamente en promedio el 86% de los documentos distribuidos por tópico, lo cual es consistente con los resultados del porcentaje de cada sentimiento en el corpus total. Por otro lado, se obtiene un coeficiente V de Cramér de 0.0312 lo que indica una asociación débil entre las variables.

Tabla 18 Matriz de contingencia variable "dominant_topic" y "sentimiento"

sentimiento	***negativo***	***positivo***
dominant_topic		
0	6969	1005
1	724	127
2	3907	643
3	4806	822
4	5010	743
5	6657	1068
6	3225	577
7	3887	585
8	657	125
9	5160	963

Tabla 19 Matriz de contingencia variable "dominant_topic" y "sentimiento" en porcentaje

sentimiento	negativo	positivo
dominant_topic		
0	87.396539	12.603461
1	85.076381	14.923619
2	85.868132	14.131868
3	85.394456	14.605544
4	87.084999	12.915001
5	86.174757	13.825243
6	84.823777	15.176223
7	86.918605	13.081395
8	84.015345	15.984655
9	84.272415	15.727585

Tabla 20 Matriz de contingencia variable "sentimiento" y "dominant_topic" en porcentaje

dominant_topic	0	1	2	3	4	5	6	7	8	9
sentimiento										
negativo	16.996732	1.765768	9.528803	11.721379	12.218916	16.235793	7.865470	9.480025	1.602361	12.584752
positivo	15.094623	1.907480	9.657555	12.346050	11.159507	16.040853	8.666266	8.786422	1.877441	14.463803

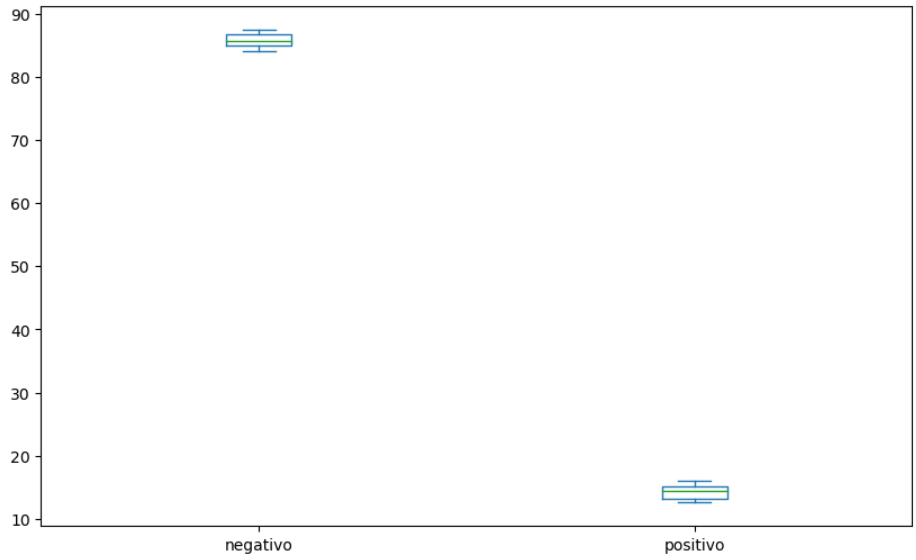

Figura 29 Boxplot variable "sentimiento" en análisis con la variable "dominant_topic"

En la figura a continuación observamos que los tópicos 0, 5 y 9 son los más dominantes, lo cual es consistente con los resultados de los tópicos más relevantes en el corpus.

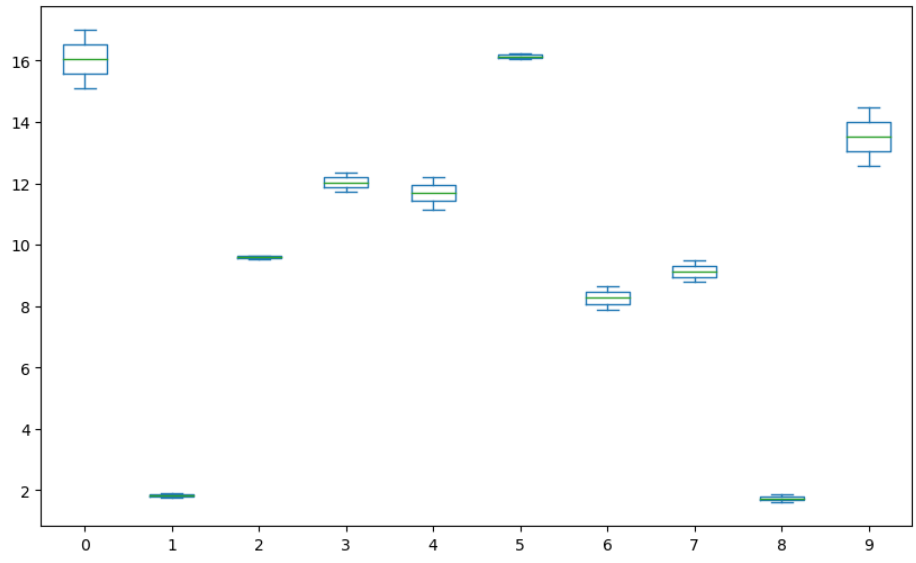

Figura 30 Boxplot variable "dominant_topic" en análisis con la variable sentimiento

8.7 Análisis Variable “dominant_topic” y “Categoria_edad”

El análisis de la tablas 21 y 22 muestran la distribución temas dominantes en diferentes categorías de edades y la tabla 23 el análisis a la inversa. Se observa que la categoria edad “Adultez” concentra aproximadamente en promedio el 54% de los documentos distribuidos por tópico, lo cual es consistente con los resultados del porcentaje de cada categoria edad en el corpus total. Por otro lado, se obtiene un coeficiente V de Cramér de 0.0426 lo que indica una asociación débil entre las variables.

En la figura 32 observamos que los tópicos 0, 5 y 9 son los más dominantes, lo cual es consistente con los resultados de los tópicos más relevantes en el corpus, sin embargo, al observar la tabla 23 vemos que al analizar los tópicos más relevantes para la Adultez mayor esta tendencia cambia, ya que el tópico 3 “Seguridad Ciudadana” se posiciona como el tercer tópico más dominante en esta categoria.

Tabla 21 Matriz de contingencia variable "dominant_topic" y "Categoria_edad"

Categoria_edad	***Adultez***	***Adultez mayor***	***Juventud***
dominant_topic			
0	3957	2543	1240
1	486	220	137
2	2433	1413	717
3	2883	1760	888
4	2953	1655	1033

5	3851	2392	1373
6	2106	908	678
7	2355	1317	734
8	405	244	132
9	3419	1612	1085

Tabla 22 Matriz de contingencia variable "dominant_topic" y "Categoria_edad" en porcentaje

Categoria_edad	***Adultez***	***Adultez mayor***	***Juventud***
dominant_topic			
0	51.124031	32.855297	16.020672
1	57.651246	26.097272	16.251483
2	53.320184	30.966469	15.713346
3	52.124390	31.820647	16.054963
4	52.348874	29.338770	18.312356
5	50.564601	31.407563	18.027836
6	57.042254	24.593716	18.364030
7	53.449841	29.891058	16.659101
8	51.856594	31.241997	16.901408
9	55.902551	26.357096	17.740353

Tabla 23 Matriz de contingencia variable "Categoria_edad" y "dominant_topic" en porcentaje

dominant_topic	0	1	2	3	4	5	6	7	8	9
Categoria_edad										
Adultez	15.924823	1.955892	9.791533	11.602543	11.884256	15.498229	8.475531	9.477624	1.629910	13.759659
Adultez mayor	18.081627	1.564278	10.046928	12.514221	11.767634	17.007964	6.456200	9.364334	1.734926	11.461889
Juventud	15.467132	1.708869	8.943495	11.076463	12.885119	17.126107	8.457029	9.155544	1.646501	13.533741

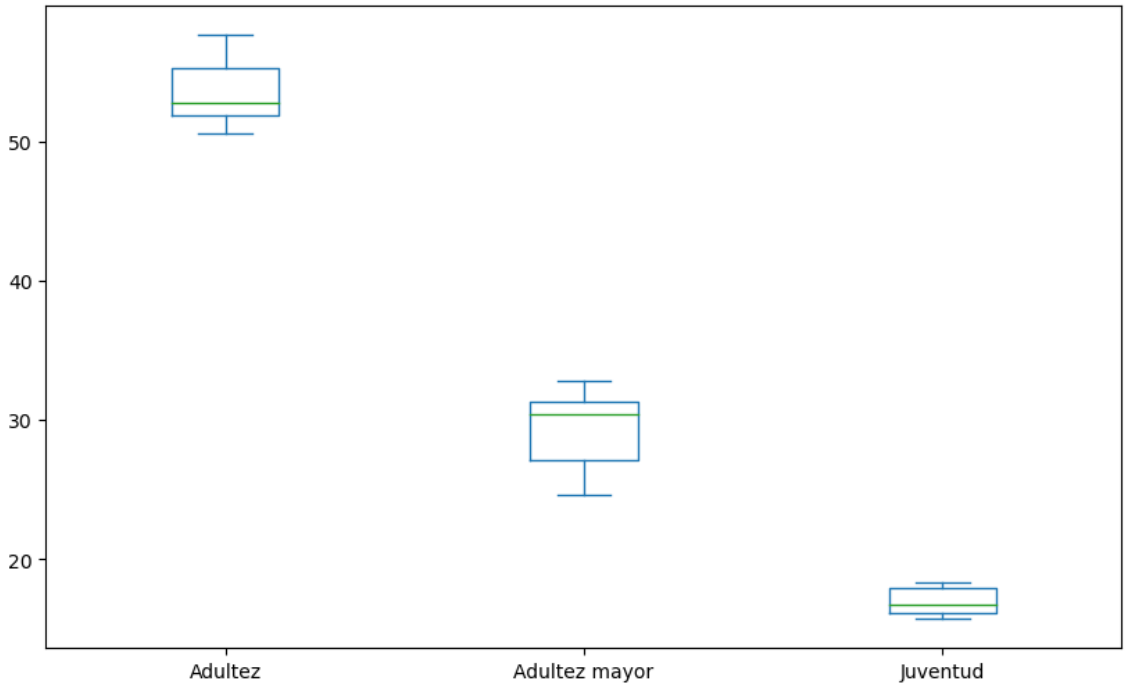

Figura 31 Boxplot variable "Categoria_edad" en análisis con la variable "dominant_topic"

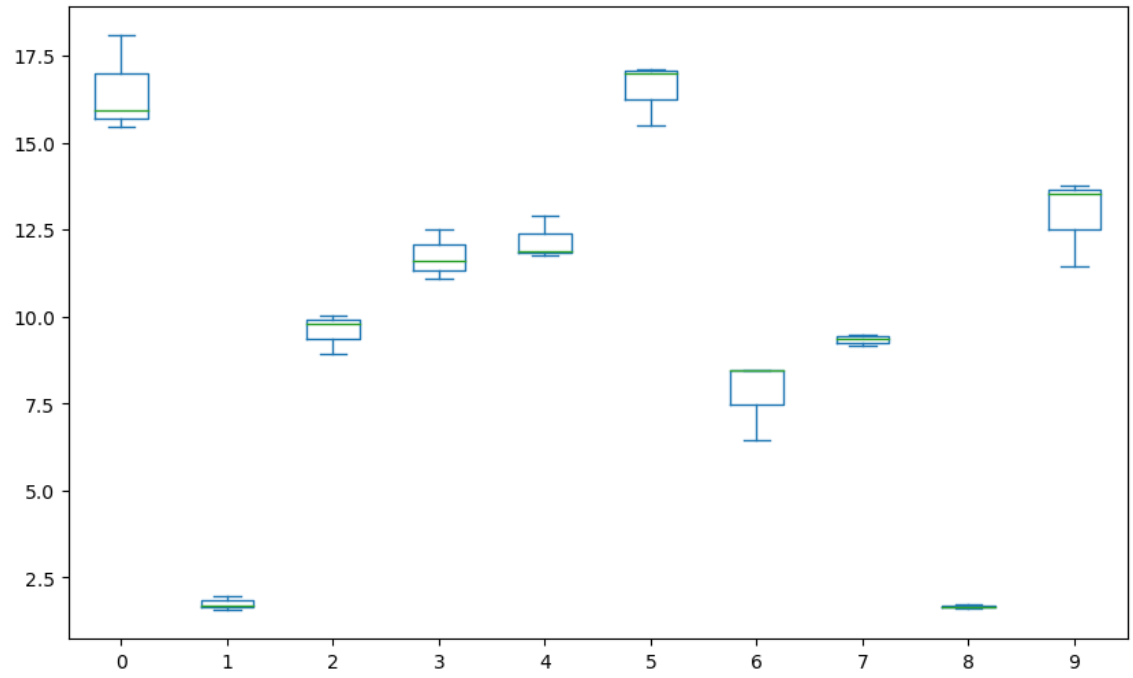

Figura 32 Boxplot variable "dominant_topic" en análisis con la variable categoria de edad

8.8 Análisis Variable "dominant_topic" y "LP_SEXO"

El análisis de la tablas 24 y 25 muestran la distribución de temas dominantes por sexo y la tabla 26 el análisis a la inversa. Se observa que el sexo "Femenino" concentra aproximadamente en promedio el 64% de los documentos distribuidos por tópico, lo cual es consistente con los resultados del porcentaje de cada sexo en el corpus total. Por otro lado, se obtiene un coeficiente V de Cramér de 0.0444 lo que indica una asociación débil entre las variables.

Tabla 24 Matriz de contingencia variable "dominant_topic" y "LP_SEXO"

LP_SEXO	F	M
dominant_topic		
0	4950	2554
1	499	318
2	2845	1527
3	3167	2128
4	3596	1846
5	4743	2512
6	2192	1387
7	2700	1504
8	485	251
9	3759	2190

Tabla 25 Matriz de contingencia variable "dominant_topic" y "LP_SEXO" en porcentaje

LP_SEXO	F	M
dominant_topic		
0	65.964819	34.035181
1	61.077111	38.922889
2	65.073193	34.926807
3	59.811143	40.188857
4	66.078648	33.921352
5	65.375603	34.624397
6	61.246158	38.753842
7	64.224548	35.775452
8	65.896739	34.103261

9	63.187090	36.812910

Tabla 26 Matriz de contingencia "LP_SEXO" y "dominant_topic" en porcentaje

dominant_topic	***0***	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***
LP_SEXO										
F	17.106718	1.724495	9.832043	10.944844	12.427426	16.391346	7.575339	9.330937	1.676113	12.990738
M	15.748905	1.960905	9.416045	13.122032	11.383116	15.489918	8.552753	9.274218	1.547759	13.504347

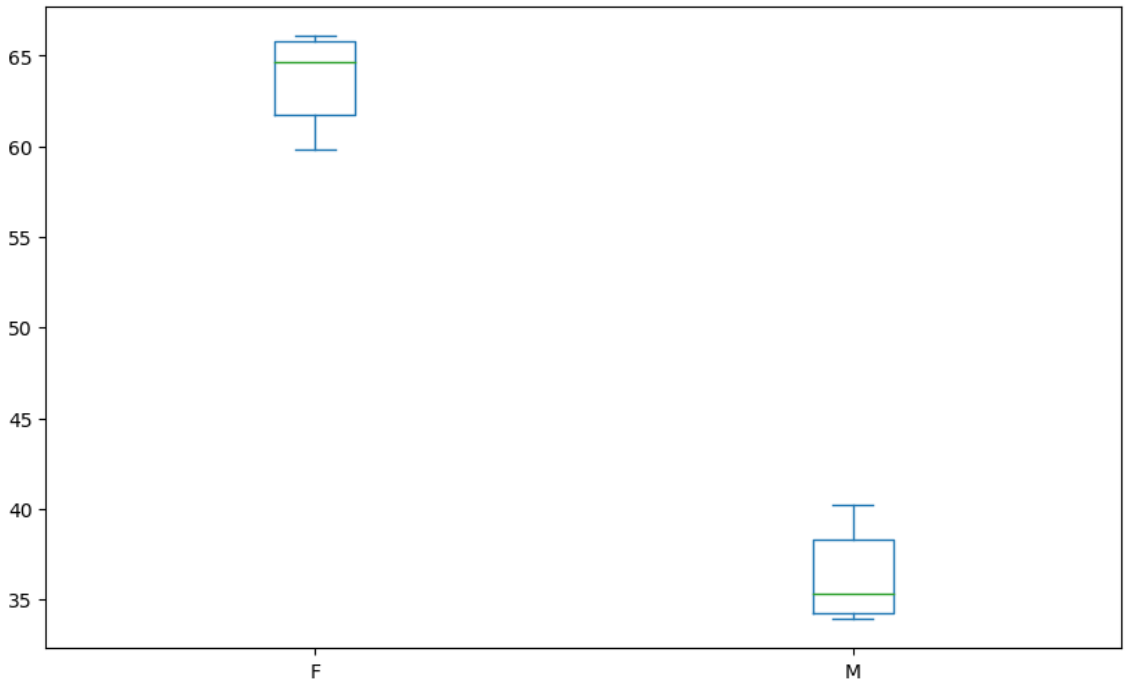

Figura 33 Boxplot variable "LP_SEXO" en análisis con la variable "dominant_topic"

En la figura a continuación observamos que los tópicos 0, 5 y 0 son los más dominantes, lo cual es consistente con los resultados de los tópicos más relevantes en el corpus.

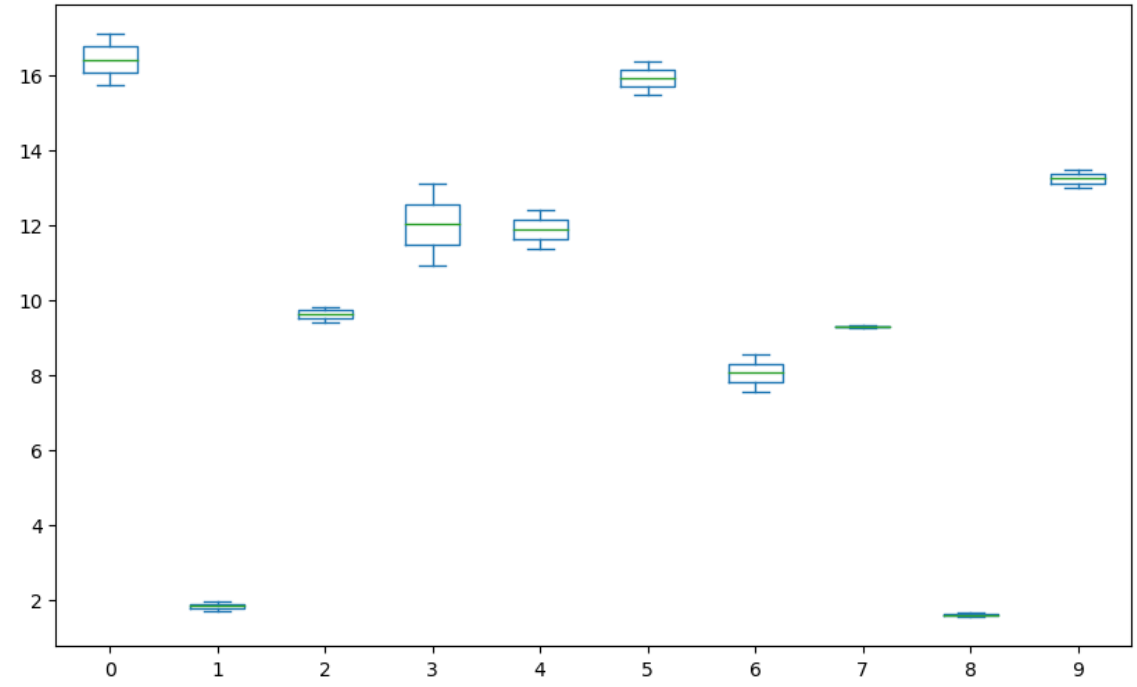

Figura 34 Boxplot variable "dominant_topic" en análisis con la variable "LP_SEXO"

8.9 Análisis Variable "LP_SEXO" y "categoria_emocion"

El análisis de la tablas 27 y 28 muestran la distribución de cada sexo en diferentes categorías de emociones y la tabla 29 el análisis a la inversa. Se observa que el sexo "Femenino" concentra aproximadamente en promedio el 63% de los documentos distribuidos por categoria de emoción, lo cual es consistente con los resultados del porcentaje de cada sexo en el corpus total. Por otro lado, se obtiene un coeficiente V de Cramér de 0.0287 lo que indica una asociación débil entre las variables.

En el gráfico 35 observamos que el porcentaje de documentos en la categoría emoción miedo es porcentualmente mayor en el sexo femenino.

Tabla 27 Matriz de contingencia variable “LP_SEXO" y "categoria_emocion"

categoria_emocion	Alegría	Aversión	Enojo	Miedo	Sorpresa	Tristeza
LP_SEXO						
F	3801	2680	4211	11374	550	4966
M	2244	1584	2447	5990	386	2806

Tabla 28 Matriz de contingencia variable “LP_SEXO" y "categoria_emocion" en porcentaje

categoria_emocion	Alegría	Aversión	Enojo	Miedo	Sorpresa	Tristeza
LP_SEXO						
F	13.780727	9.716482	15.267203	41.237039	1.994054	18.004496
M	14.517694	10.247784	15.831015	38.752669	2.497250	18.153587

Tabla 29 Matriz de contingencia variable "categoria_emocion" y “LP_SEXO" en porcentaje

LP_SEXO	F	M
categoria_emocion		
Alegría	62.878412	37.121588

Aversión	62.851782	37.148218
Enojo	63.247221	36.752779
Miedo	65.503340	34.496660
Sorpresa	58.760684	41.239316
Tristeza	63.896037	36.103963

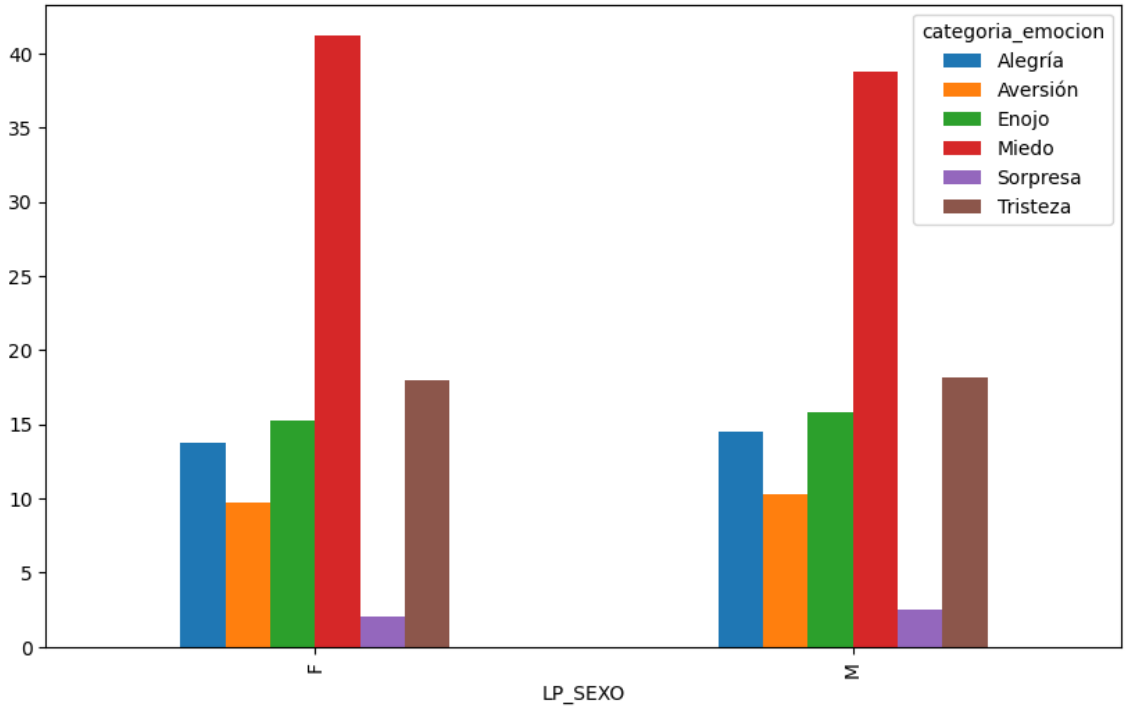

Figura 35 Gráfico de barras variable "LP_SEXO" en análisis con la variable "Categoria_emocion"

8.10 Análisis Variable "Categoria_edad" y "categoria_emocion"

El análisis de la tablas 30 y 31 muestran la distribución de cada categoria de edad en diferentes categorías de emociones y la tabla 32 el análisis a la inversa. Se observa que la

categoria edad “Adultez” concentra aproximadamente en promedio el 53% de los documentos distribuidos por categoria de emoción, lo cual es consistente con los resultados del porcentaje de cada categoria edad en el corpus total. Se Observa que la categoría de emoción “Miedo” concentra aproximadamente en promedio el 39% de los documentos distribuidos por categoria de edad, lo cual es consistente con los resultados del porcentaje de cada categoria de emoción en el corpus total.

En la Figura 36 observamos que el porcentaje de documentos en la categoría emoción miedo es porcentualmente mayor en los adultos mayores, mientras que en los jóvenes es menor.

Por otro lado, se obtiene un coeficiente V de Cramér de 0.0553 lo que indica una asociación débil entre las variables.

Tabla 30 Matriz de contingencia variable "Categoria_edad" y "categoria_emocion"

categoria_emocion	*Alegría*	*Aversión*	*Enojo*	*Miedo*	*Sorpresa*	*Tristeza*
Categoria_edad						
Adultez	3374	2434	3682	9468	529	4174
Adultez mayor	1600	1202	2042	5896	225	2468
Juventud	1268	870	1330	2611	210	1334

Tabla 31 Matriz de contingencia variable "Categoria_edad" y "categoria_emocion" en porcentaje

categoria_emocion	*Alegría*	*Aversión*	*Enojo*	*Miedo*	*Sorpresa*	*Tristeza*
Categoria_edad						
Adultez	14.259752	10.286970	15.561472	40.015215	2.235747	17.640844
Adultez mayor	11.910966	8.948113	15.201370	43.891908	1.674980	18.372664
Juventud	16.633871	11.412830	17.447199	34.251607	2.754821	17.499672

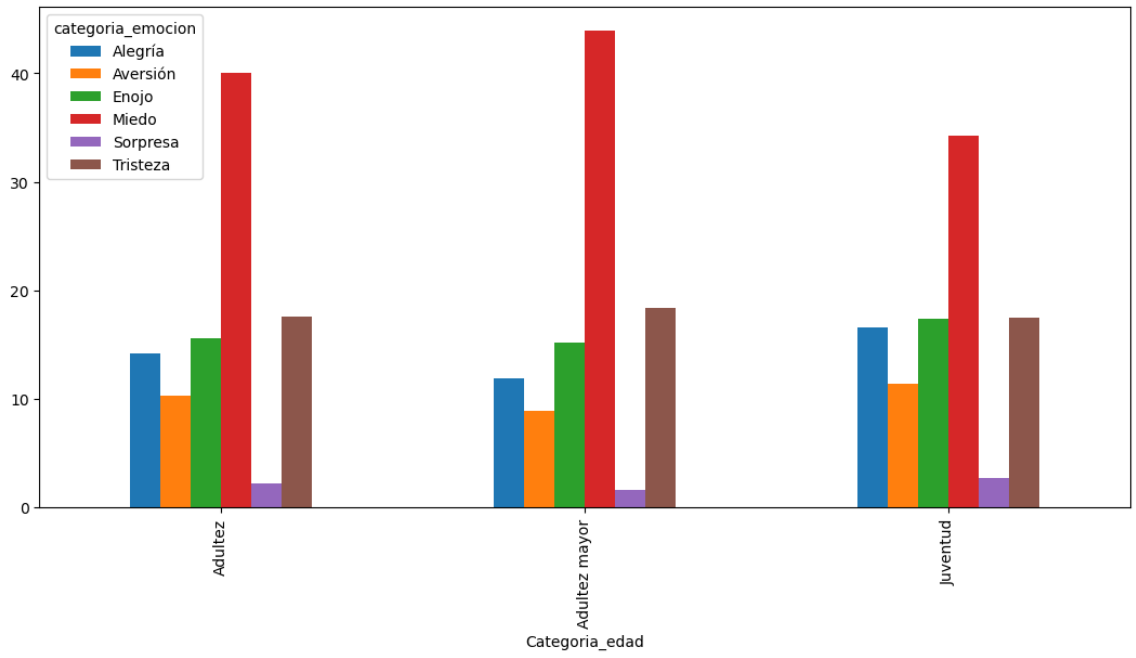

Figura 36 Gráfico de barras Análisis variables "Categoria_edad" y "categoria_emocion"

Tabla 32 Matriz de contingencia variable "categoria_emocion" y “Categoria_edad"

Categoria_edad	Adultez	Adultez mayor	Juventud
categoria_emocion			
Alegría	54.053188	25.632810	20.314002
Aversión	54.016866	26.675544	19.307590
Enojo	52.197335	28.948115	18.854551
Miedo	52.673157	32.801113	14.525730
Sorpresa	54.875519	23.340249	21.784232
Tristeza	52.331996	30.942828	16.725176

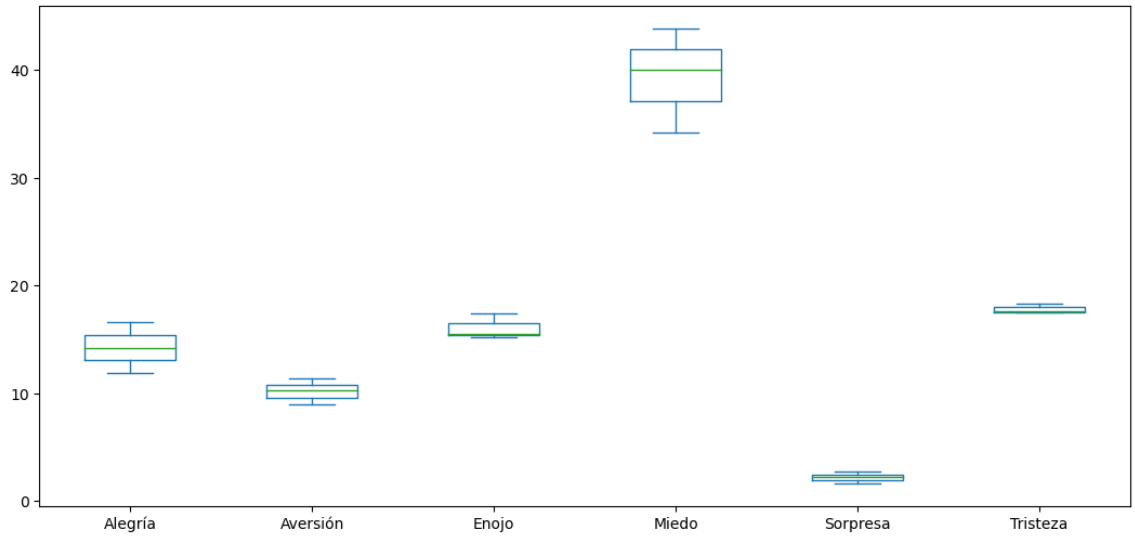

Figura 37 Boxplot variable “categoria_emocion” en análisis con la variable “Categoria_edad”

8.11 Análisis Variable “Categoria_edad” y “sentimiento”

El análisis de la tablas 33 y 34 muestran la distribución de cada categoria de edad con respecto a los sentimientos y la tabla 35 el análisis a la inversa. Se observa que los sentimientos “negativos” concentran aproximadamente en promedio el 86% de los documentos distribuidos por cada categoria de edad, lo cual es consistente con los resultados del porcentaje de cada sentimiento en el corpus total. Se observa que la categoria edad “Adultez” concentra aproximadamente en promedio el 53% de los documentos distribuidos por cada sentimiento, lo cual es consistente con los resultados del porcentaje de cada categoria edad en el corpus total. Por otro lado, se obtiene un coeficiente V de Cramér de 0.0459 lo que indica una asociación débil entre las variables.

Tabla 33 Matriz de contingencia “Categoria_edad" y "sentimiento"

sentimiento	*negativo*	*positivo*
Categoria_edad		
Adultez	20287	3374
Adultez mayor	11833	1600
Juventud	6355	1268

Tabla 34 Matriz de contingencia “Categoria_edad" y "sentimiento” en porcentaje

sentimiento	*negativo*	*positivo*
Categoria_edad		
Adultez	85.740248	14.259752
Adultez mayor	88.089034	11.910966
Juventud	83.366129	16.633871

Tabla 35 Matriz de contingencia "sentimiento" y “Categoria_edad” en porcentaje

Categoria_edad	*Adultez*	*Adultez mayor*	*Juventud*
sentimiento			
negativo	52.727745	30.755036	16.517219
positivo	54.053188	25.632810	20.314002

8.12 Análisis Variable “Categoria_edad” y “LP_SEXO”

El análisis de la tablas 36 y 37 muestran la distribución del sexo con respecto a cada categoria de edad la tabla 38 el análisis a la inversa. Se observa que la categoria edad “Adultez” concentra aproximadamente en promedio el 53% de los documentos distribuidos por cada sentimiento, lo cual es consistente con los resultados del porcentaje de cada categoria edad en el corpus total. Se observa que el sexo “Femenino” concentra aproximadamente en promedio el 63% de los documentos distribuidos por cada categoria de edad, lo cual es consistente con los resultados del porcentaje de cada sexo en el corpus

total. Por otro lado, se obtiene un coeficiente V de Cramér de 0.0560 lo que indica una asociación débil entre las variables.

Tabla 36 Matriz de contingencia "LP_SEXO" y “Categoria_edad”

Categoria_edad	*Adultez*	*Adultez mayor*	*Juventud*
LP_SEXO			
F	15477	8637	4409
M	8167	4611	3164

Tabla 37 Matriz de contingencia "LP_SEXO" y “Categoria_edad” en porcentaje

Categoria_edad	*Adultez*	*Adultez mayor*	*Juventud*
LP_SEXO			
F	54.261473	30.280826	15.457701
M	51.229457	28.923598	19.846945

Tabla 38 Matriz de contingencia “Categoria_edad” y "LP_SEXO” en porcentaje

LP_SEXO	*F*	*M*
Categoria_edad		
Adultez	65.458467	34.541533
Adultez mayor	65.194746	34.805254
Juventud	58.219992	41.780008