



ANÁLISIS DE DISCURSO BASADO EN MODELOS GRANDES DE  
LENGUAJE

Discourse analysis based on Large Language Models.

EDGAR LEANDRO JIMENEZ JAIMES

Tesis de maestría

Asesor

Edwin Nelson Montoya Munera

UNIVERSIDAD EAFIT

ESCUELA DE CIENCIAS.

MAESTRÍA EN CIENCIA DE LOS DATOS Y ANALÍTICA.

MEDELLÍN.

2024

## CONTENIDO

|         |                                                                    |    |
|---------|--------------------------------------------------------------------|----|
| 1       | INTRODUCCIÓN.....                                                  | 9  |
| 1.1     | PLANTEAMIENTO DEL PROBLEMA .....                                   | 9  |
| 1.2     | JUSTIFICACIÓN .....                                                | 10 |
| 1.3     | OBJETIVOS .....                                                    | 11 |
| 1.4     | METODOLOGÍA DEL PROYECTO .....                                     | 11 |
| 1.5     | ESTRUCTURA DEL PROYECTO .....                                      | 13 |
| 2       | MARCO TEÓRICO Y ESTADO DEL ARTE.....                               | 14 |
| 2.1     | MARCO TEÓRICO.....                                                 | 14 |
| 2.1.1   | Procesamiento de lenguaje natural .....                            | 14 |
| 2.1.1.1 | Aplicaciones del NLP.....                                          | 14 |
| 2.1.2   | Procesamiento de lenguaje natural en el análisis de discurso ..... | 15 |
| 2.2     | ESTADO DEL ARTE.....                                               | 17 |
| 2.2.1   | Avances en LLM.....                                                | 17 |
| 2.2.1.1 | Embeddings.....                                                    | 17 |
| 2.2.1.2 | Modelos generativos .....                                          | 19 |
| 2.2.1.3 | Prompt engineering .....                                           | 20 |
| 2.2.1.4 | RAG - Retrieval-Augmented Generation.....                          | 22 |
| 2.2.1.5 | Aplicaciones en el análisis de discurso. ....                      | 24 |
| 3       | DESARROLLO DEL TRABAJO.....                                        | 25 |
| 3.1     | ENTENDIMIENTO Y COMPRENSIÓN DEL NEGOCIO .....                      | 25 |
| 3.2     | ENTENDIMIENTO Y COMPRENSIÓN DEL DATO .....                         | 26 |

|                                                                           |    |
|---------------------------------------------------------------------------|----|
| 3.3. PREPARACIÓN DE DATOS .....                                           | 28 |
| 3.4. MODELADO DE DATOS.....                                               | 28 |
| 3.4.1. Modelado de tópicos utilizando Prompt Engineering y LLMs .....     | 28 |
| 3.4.2. Modelado de Sentimiento utilizando Prompt Engineering y LLMs ..... | 31 |
| 3.4.3. Agrupamiento y visualización .....                                 | 32 |
| 3.4.4. Asistente conversacional .....                                     | 36 |
| 3.4.5. Modelado de clasificación de texto.....                            | 39 |
| 3.5. EVALUACIÓN .....                                                     | 45 |
| 4 RESULTADOS .....                                                        | 46 |
| 4.1. MODELADO DE TÓPICOS UTILIZANDO PROMPT ENGINEERING Y LLMs .....       | 46 |
| 4.2. MODELADO DE SENTIMIENTO UTILIZANDO PROMPT ENGINEERING Y LLMs .....   | 50 |
| 4.3. AGRUPAMIENTO Y VISUALIZACIÓN .....                                   | 55 |
| 4.4. ASISTENTE CONVERSACIONAL .....                                       | 62 |
| 4.5. MODELADO DE CLASIFICACIÓN DE TEXTO.....                              | 64 |
| 5 CONCLUSIONES Y TRABAJOS FUTUROS.....                                    | 69 |
| 5.1 CONCLUSIONES .....                                                    | 69 |
| 5.2. TRABAJOS FUTUROS .....                                               | 71 |
| 6 REFERENCIAS.....                                                        | 72 |
| ANEXOS.....                                                               | 78 |

## LISTA DE FIGURAS

|                                                                                                  |    |
|--------------------------------------------------------------------------------------------------|----|
| <b>Figura 1.</b> Etapas metodología CRISP-DM. ....                                               | 12 |
| <b>Figura 2.</b> Relaciones lingüísticas modeladas por desplazamiento de vector lineal.<br>..... | 18 |
| <b>Figura 3.</b> Sistema RAG con LLM para una aplicación de QA. ....                             | 23 |
| <b>Figura 4.</b> Función de clasificación de tópicos usando API de OpenAI. ....                  | 29 |
| <b>Figura 5.</b> Contexto para asignar tópicos utilizando prompt Engineering. ....               | 30 |
| <b>Figura 6.</b> Proceso general para asignar tópicos a respuestas. ....                         | 30 |
| <b>Figura 7.</b> Función para calcular sentimiento usando API OpenAI. ....                       | 31 |
| <b>Figura 8.</b> Proceso general para asignar sentimiento a respuestas. ....                     | 32 |
| <b>Figura 9.</b> Función para construir macro etiquetas a partir de API de OpenAI. ....          | 34 |
| <b>Figura 10.</b> Contexto utilizado para función de macro etiquetas. ....                       | 35 |
| <b>Figura 11.</b> Proceso general para agrupamiento y visualización. ....                        | 36 |
| <b>Figura 12.</b> Proceso de creación y almacenamiento de embeddings. ....                       | 37 |
| <b>Figura 13.</b> Proceso general de funcionamiento del asistente. ....                          | 37 |
| <b>Figura 14.</b> Función para crear contexto completo. ....                                     | 38 |
| <b>Figura 15.</b> Función para hacer llamado al modelo con contexto completo. ....               | 39 |
| <b>Figura 16.</b> Función para construir embeddings a partir de API OpenAI. ....                 | 40 |
| <b>Figura 17.</b> Proceso general de clasificación con aproximación 1. ....                      | 41 |
| <b>Figura 18.</b> Ejemplo de construcción de texto a partir de respuestas. ....                  | 41 |
| <b>Figura 19.</b> Función para mejorar texto concatenado usando LLMs de OpenAI. ....             | 42 |
| <b>Figura 20.</b> Ejemplo de uso de salida de función para transformar respuesta. ....           | 43 |
| <b>Figura 21.</b> Función para obtener embeddings. ....                                          | 43 |
| <b>Figura 22.</b> Proceso general de clasificación con aproximación 2. ....                      | 44 |
| <b>Figura 23.</b> Varianza explicada en función del número de componentes principales.<br>.....  | 44 |
| <b>Figura 24.</b> Proceso general de clasificación con aproximación 3. ....                      | 45 |
| <b>Figura 25.</b> Análisis de tópicos y temas principales. ....                                  | 46 |
| <b>Figura 26.</b> Análisis de tema por sentimientos top 10 temas generales. ....                 | 47 |

|                                                                                                      |    |
|------------------------------------------------------------------------------------------------------|----|
| <b>Figura 27.</b> Análisis por género. ....                                                          | 48 |
| <b>Figura 28.</b> Análisis por zonas. ....                                                           | 49 |
| <b>Figura 29.</b> Análisis por otras categorías. ....                                                | 50 |
| <b>Figura 30.</b> Distribución del análisis de sentimientos.....                                     | 51 |
| <b>Figura 31.</b> Valor medio sentimiento por género y categoría de pregunta 1.....                  | 52 |
| <b>Figura 32.</b> Valor medio sentimiento por grupo etario y categoría de pregunta 1...              | 53 |
| <b>Figura 33.</b> Valor medio sentimiento por Región y categoría de pregunta 1. ....                 | 53 |
| <b>Figura 34.</b> Valor medio sentimiento por grupo étnico y categoría de pregunta 1. .              | 54 |
| <b>Figura 35.</b> Valor medio sentimiento por zona y categoría de pregunta 1. ....                   | 55 |
| <b>Figura 36.</b> Resultado método del codo para clustering de categoría “cambiar” . .               | 56 |
| <b>Figura 37.</b> Grafica 2D de clustering y etiquetas a respuesta de ¿Qué deberíamos cambiar?.....  | 56 |
| <b>Figura 38.</b> Grafica 3D de clustering y etiquetas a respuesta de ¿Qué deberíamos cambiar?.....  | 57 |
| <b>Figura 39.</b> Resultado método del codo para clustering de categoría “mejorar” . ...             | 58 |
| <b>Figura 40.</b> Grafica 2D de clustering y etiquetas a respuesta de ¿Qué deberíamos mejorar?.....  | 59 |
| <b>Figura 41.</b> Grafica 3D de clustering y etiquetas a respuesta de ¿Qué deberíamos mejorar?.....  | 59 |
| <b>Figura 42.</b> Resultado método del codo para clustering de categoría “mantener” . .              | 60 |
| <b>Figura 43.</b> Grafica 2D de clustering y etiquetas a respuesta de ¿Qué deberíamos mantener?..... | 61 |
| <b>Figura 44.</b> Grafica 3D de clustering y etiquetas a respuesta de ¿Qué deberíamos mantener?..... | 61 |
| <b>Figura 45.</b> Respuesta del asistente conversacional. ....                                       | 63 |
| <b>Figura 46.</b> Respuesta del asistente conversacional ante otra pregunta. ....                    | 64 |

## LISTA DE TABLAS

|                                                                                         |    |
|-----------------------------------------------------------------------------------------|----|
| <b>Tabla 1:</b> Resumen modelos y técnicas .....                                        | 25 |
| <b>Tabla 2</b> Indicadores verbales y descripción.....                                  | 26 |
| <b>Tabla 3</b> Muestra de datos empleados (ciclo 1).....                                | 27 |
| <b>Tabla 4</b> Conjunto de datos seleccionado para modelo de clasificación de texto ... | 40 |
| <b>Tabla 5</b> Resultado Regresión Logística con diferentes aproximaciones .....        | 65 |
| <b>Tabla 6</b> Resultado Random Forest con diferentes aproximaciones.....               | 66 |
| <b>Tabla 7</b> Resultado Naives Bayes con diferentes aproximaciones.....                | 67 |

## RESUMEN

En esta tesis se explora la implementación de técnicas de procesamiento de lenguaje natural y modelos grandes de lenguaje (LLMs por sus siglas en inglés) para apoyar tareas de análisis del discurso en el contexto del programa "Tenemos que Hablar Colombia". Se han abordado técnicas como el modelado de tópicos, análisis de sentimientos, clustering, visualización, y la creación de un asistente conversacional basado en Retrieval Augmented Generation (RAG, por sus siglas en inglés) utilizando aproximaciones avanzadas de modelado de texto vector embeddings e ingeniería de prompts. También se presenta un modelo de clasificación de texto enfocado en predecir la etiqueta de la variable de indicador verbal, asignada manualmente por el entrevistador, aunque este modelo no se aplica directamente al análisis de discurso.

Este trabajo se suma a los estudios del programa "Tenemos que Hablar Colombia", donde otros autores han contribuido mediante análisis de lingüística computacional y técnicas de aprendizaje automático. Utilizando técnicas avanzadas de NLP, se ha buscado mejorar la interpretación de los datos de texto y su aplicación en el análisis de discurso. Los resultados han mostrado mejoras en la precisión de la clasificación y análisis de datos a través de las técnicas exploradas, proporcionando una mejor comprensión de las percepciones ciudadanas.

Palabras clave: Análisis de discurso, análisis de sentimientos, asistente conversacional, clustering, modelos grandes de lenguaje, procesamiento de lenguaje natural, PCA, prompt engineering, retrieval-augmented generation, vector embeddings, visualización de datos.

## ABSTRACT

This thesis explores the implementation of natural language processing techniques and large language models (LLMs) to support discourse analysis tasks in the context of the "Tenemos que hablar Colombia" program. Techniques such as topic modeling, sentiment analysis, clustering, visualization, and the creation of a conversational assistant based on Retrieval Augmented Generation (RAG) have been addressed using advanced text modeling, vector embeddings, and prompt engineering approaches.

A text classification model focused on predicting the label of the verbal indicator variable, assigned manually by the interviewer, is also presented, although this model is not directly applied to discourse analysis.

This work adds to the studies of the "Tenemos que hablar Colombia" program, where other authors have contributed through computational linguistics analysis and machine learning techniques. Using advanced NLP techniques, we have sought to improve the interpretation of text data and its application in discourse analysis. The results have shown improvements in the accuracy of data classification and analysis through the techniques explored, providing a better understanding of citizen perceptions.

**Keywords:** discourse analysis, sentiment analysis, conversational assistant, clustering, large language models, natural language processing, PCA, prompt engineering, retrieval-augmented generation, vector embeddings, data visualization.

# 1 INTRODUCCIÓN

## 1.1 PLANTEAMIENTO DEL PROBLEMA

Los métodos más frecuentes utilizados para conocer la percepción y opinión de las personas frente a un evento son las encuestas y entrevistas estructuradas; por un lado, las encuestas, implican la recopilación sistemática de información mediante cuestionarios estandarizados los cuales suelen estructurarse con escalas y/o entradas de datos en texto libre (Babbie, 2013). Las entrevistas, por su parte, recopilan información cualitativa, tanto verbal como no verbal (gestos, tono de voz), y buscan analizar en profundidad los temas que surgen durante la interacción de acuerdo con los objetivos y fenómenos descritos (Kvale, 2007). Ambos métodos pueden recopilar datos cualitativos y cuantitativos, los cuales suelen ser sometidos a procesamientos estadísticos para identificar patrones, tendencias y significados dentro de las respuestas (Kvale, 2007; Babbie, 2013; Creswell & Creswell, 2017). Pese a que se han desarrollado técnicas y algoritmos que facilitan el análisis de los grandes volúmenes de datos de texto, aún se evidencian desafíos y limitaciones en algunos de ellos (Manning et al., 2008; Miner et al., 2012; Jurafsky & Martin, 2018).

Técnicas tradicionales para el procesamiento de lenguaje natural (NLP por sus siglas en inglés, *Natural Language Processing*), con representaciones como BoW (*Bag-of-Words*), TF-IDF (*Term Frequency - Inverse Document Frequency*) y los modelos de n-gramas presentan limitaciones para captar la complejidad y el contexto de las respuestas textuales. Estas técnicas tienden a ignorar el contexto y el orden de las palabras, lo que resulta en una representación superficial del lenguaje y la pérdida de matices importantes (Kamkarhaghighi & Makrehchi, 2017; Asudani et al., 2023; Johnson, 2024). Además, desafíos como la integración y normalización de datos de múltiples fuentes, la ambigüedad semántica, detección de sentimientos, emociones y la escalabilidad para manejar grandes volúmenes de datos de texto complejizan aún más el análisis de los datos (Ghazizadeh, 2020). Sin embargo, los recientes avances en las técnicas de NLP han logrado una integración efectiva entre la lingüística computacional, el aprendizaje automático y la inteligencia artificial, resultando en métodos más sofisticados para examinar y entender las estructuras lingüísticas complejas en textos largos (O'Neill et al., 2021).

De acuerdo a lo anterior, optar por una selección de modelos de aprendizaje profundo y técnicas de procesamiento avanzadas como embeddings para obtener mejores resultados es vital, es por ello que, el presente trabajo busca explorar y aplicar estas técnicas en un caso de estudio particular como lo es el proyecto “Tenemos que Hablar Colombia” (TQHC; <https://tenemosquehablarcolombia.co/>), el cual proporciona una importante información en texto, debido a la estructura y forma de su recolección (entrevistas estructuradas). El caso particular de estas entrevistas es que se obtienen en el levantamiento de opinión pública, son base de la generación de diferentes trabajos y proyectos en el procesamiento de lenguaje natural para apoyar las funciones de análisis de discurso (Fajardo Becerra, 2023).

Es por ello que, este trabajo se enfocará en explorar e implementar métodos de técnicas avanzadas de NLP como el uso de modelos de transformers, embeddings, modelos grandes de lenguaje (ej, GPT, Llama3) y de reducción de dimensionalidad (como PCA) para analizar y visualizar la estructura de los datos textuales. De igual manera, se aplicará la combinación de dichas técnicas para clasificar etiquetas de datos, y obtener insights sobre las opiniones y percepciones de los participantes del proyecto “TQHC”. Con la implementación de estas técnicas, se espera mejorar la interpretación del análisis de discurso, permitiendo una comprensión más amplia de las conversaciones y temas relevantes. Adicionalmente, este estudio examinará la capacidad de manejar datos no estructurados como el texto y la necesidad de capturar el contexto y los matices en el lenguaje natural. Se espera que este trabajo contribuya a mejorar la interpretación del análisis de discurso, proporcionando técnicas y aplicaciones prácticas para el procesamiento y análisis de textos de opinión, superando las limitaciones de los métodos tradicionales de NLP para comprender la riqueza y complejidad de este tipo de datos, y así generar nuevo conocimiento a partir de los resultados.

## **1.2 JUSTIFICACIÓN**

Las metodologías clásicas de NLP presentan varias limitaciones que han sido documentadas en la literatura académica. En primer lugar, técnicas de representación como el Bag of Words (BoW) y TF-IDF ignoran el orden de las palabras y por ende no capturan la estructura gramatical ni tampoco la semántica contextual de los textos analizados, lo que puede llevar a una pérdida de información semántica (Naseem et al, 2021). Así mismo, aunque TF-IDF puede superar ciertas limitaciones de BoW, al dar mayor peso a las palabras menos frecuentes en los documentos, puede no captar adecuadamente el contexto y las relaciones entre palabras, lo que limita su precisión y efectividad en tareas más complejas de procesamiento del lenguaje natural (Gruetzemacher, 2022).

Con la llegada de modelos de lenguaje como BERT (Bidirectional Encoder Representation from Transformers), se ha experimentado un avance importante y exponencial, permitiendo un análisis más acertado del lenguaje. Estos modelos, utilizan arquitecturas avanzadas para representar las relaciones contextuales, superando las limitaciones de las técnicas tradicionales (Devlin et al., 2018). En ese mismo sentido, los modelos de lenguaje grandes pueden manejar grandes volúmenes de datos de texto de manera eficiente, permitiendo una mejor comprensión y gestión de dichos datos (Brown, 2020). En general los LLMs han facilitado la extracción de significados contextuales complejos, y han mejorado la precisión en la identificación de relaciones de coherencia, superando en repetidas ocasiones limitaciones de las técnicas clásicas de NLP (O’Neill, 2021)

Este proyecto presenta aproximaciones modernas en la forma en cómo se procesan, modelan, analizan y visualizan los datos de texto frente a los métodos tradicionales de NLP para apoyar las diferentes tareas en el análisis de discurso.

Utilizar este tipo de aproximaciones permitirá una mejor comprensión y gestión de este tipo de datos, optimizando recursos y mejorando la calidad del análisis, permitiendo entender patrones en datos de texto, lo que es paso importante para el entendimiento de la percepción ciudadana, así como para la formulación de políticas públicas y toma de decisiones informadas.

La necesidad de este trabajo surgió al identificar las limitaciones en los enfoques clásicos de NLP para manejar datos de texto, utilizados en programas como “Tenemos que hablar Colombia”, fuente principal de datos para el presente trabajo. Se espera estudiar y aplicar avances recientes en IA y NLP, para transformar y acelerar el análisis de discurso, mejorando así la calidad de las decisiones basadas en datos en programas de opinión pública.

### **1.3 OBJETIVOS**

#### **1.2.1. Objetivo General**

Explorar y aplicar técnicas avanzadas de modelos grandes de lenguaje (LLMs) para apoyar diferentes tareas de análisis de discurso.

#### **1.2.2. Objetivos específicos**

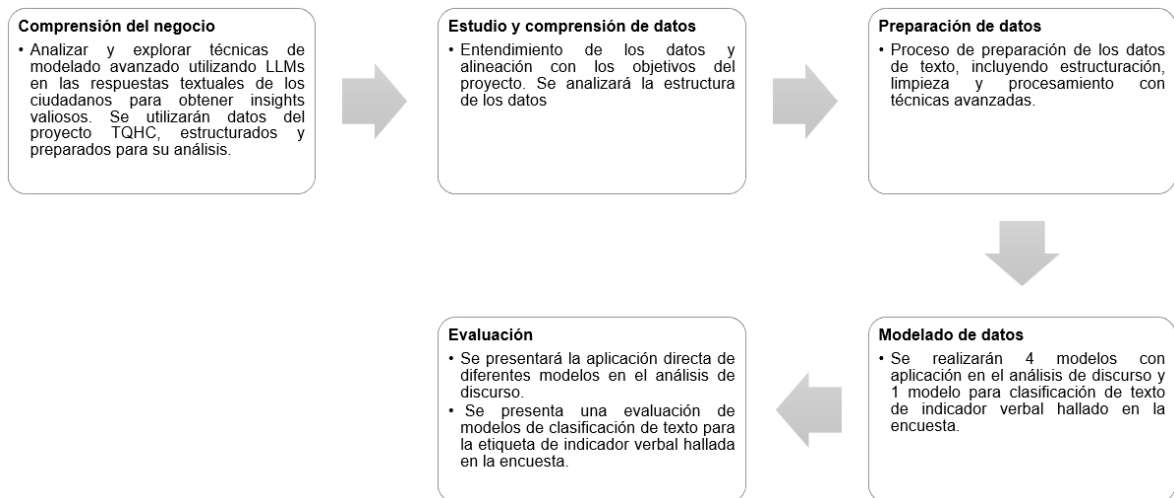
- Implementar un modelo de análisis de sentimientos utilizando prompt engineering y modelos generativos y analizar su aplicación en el análisis de discurso.
- Implementar un modelo de tópicos mediante prompt engineering y modelos generativos y analizar su aplicación en el análisis de discurso.
- Implementar un modelo de clustering para agrupar temas y desarrollar gráficos en 2D y 3D para visualizar relaciones y clusters, utilizando técnicas de reducción de dimensionalidad y analizar su aplicación en el análisis de discurso
- Implementar un asistente conversacional utilizando técnicas de RAG para analizar su aplicación en el análisis de discurso
- Implementar y evaluar diferentes modelos de clasificación de texto utilizando aproximación de embeddings para apoyar procesos de etiquetados en los datos.

### **1.4 METODOLOGÍA DEL PROYECTO**

El proceso de extraer conocimiento de los datos ha requerido grandes esfuerzos en los últimos años debido al crecimiento exponencial de la generación de datos y la forma en que interactuamos con la tecnología. En este contexto, se han desarrollado metodologías que buscan alinear la ejecución de los proyectos analíticos con

objetivos de negocio en el ámbito empresarial o de investigación en el ámbito académico. Una de estas metodologías es CRISP-DM (Cross Industry Standard Process for Data Mining; Chapman et al, 2020), que, a través de seis etapas, aborda de manera integral el despliegue de un sistema analítico, comenzando por el entendimiento y comprensión del negocio. Este proyecto propone seguir esta metodología. A continuación, en la Figura 1 muestra las etapas de CRISP-DM y su aplicación en este proyecto.

**Figura 1.** Etapas metodología CRISP-DM.



Fuente:

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide. *SPSS inc*, 9(13), 1-73.

Tal como se expone en la Figura 1, este trabajo sigue la metodología estructurada que abarca desde la comprensión del negocio y los datos hasta la evaluación de técnicas avanzadas de modelado, con un enfoque particular en los modelos de lenguaje grande. Inicialmente, se profundiza en los objetivos del proyecto y en la naturaleza de los datos disponibles por el programa, asegurando una alineación con los objetivos planteados. Luego, se lleva a cabo la preparación de los datos de texto necesaria para cada modelo, para luego ser procesados por los LLMs, aplicando en algunos casos, procesos de limpieza y estructuración. Con los datos procesados, se examinaron diferentes técnicas de modelación de lenguaje para analizar las respuestas y opiniones, evaluando su desempeño y explicación en el análisis de discurso. Finalmente, los resultados obtenidos se presentan mediante visualizaciones, y se explora el uso de un asistente conversacional para enriquecer el análisis y la interpretación de los datos de la encuesta, proporcionando una visión aplicada de los hallazgos desde el enfoque de los LLMs. Se exceptúa la fase de despliegue dado que el alcance de este proyecto se limita a explorar nuevas técnicas avanzadas y revisar su aplicación en el análisis de discurso.

## **1.5 ESTRUCTURA DEL PROYECTO**

La presente tesis está organizada en seis secciones: La primera es la introducción, que argumenta el planteamiento del problema de estudio seguidamente por su justificación, objetivos (generales y específicos), metodología y estructura. La segunda sección, presenta el marco teórico y el estado del arte, que incluye los desarrollos conceptuales y soportes teóricos requeridos para el estudio. La tercera sección provee el desarrollo del trabajo, la cual se desarrolla en 5 fases: entendimiento y comprensión del negocio, entendimiento y comprensión de los datos, preparación datos, modelado de datos, y evaluación de resultados. La cuarta sección, expone detalladamente los resultados del estudio. La quinta sección ofrece la discusión global, la cual consolida e interpreta las evidencias y hallazgos e incluye las conclusiones de la tesis y trabajo futuro. Finalmente se presentan las referencias y los anexos.

## 2 MARCO TEÓRICO Y ESTADO DEL ARTE

### 2.1 MARCO TEÓRICO

#### 2.1.1 Procesamiento de lenguaje natural

El procesamiento del lenguaje natural es un campo que combina la informática, la lingüística y la inteligencia artificial (IA) para crear y desarrollar algoritmos y sistemas capaces de interpretar el lenguaje de los humanos de una manera comprensible para las máquinas (Brookshear, 1993). En informática, NLP se basa principalmente en teorías de autómatas y lenguajes formales para diseñar modelos y algoritmos eficientes, por su parte la lingüística contribuye a la comprensión de la estructura del lenguaje en sí mismo, abarcando la sintaxis, la semántica y la pragmática (Vicente et al., 2015). La inteligencia artificial, en cambio, a través del aprendizaje automático (también llamado Machine Learning - ML por sus siglas en inglés), abarca métodos matemáticos y estadísticos avanzados como las redes neuronales artificiales, que aprenden patrones particularmente en grandes volúmenes de datos (Bengio et al., 2013). Las técnicas estadísticas empleadas en NLP abarcan los enfoques de modelos supervisados y no supervisados para diferentes tareas como clasificación de textos y modelado de tópicos. Estos modelos estadísticos se entrenan utilizando grandes corpus de texto, lo que les permite realizar tareas específicas como traducción automática, análisis de sentimientos y generación de texto. Manning (1999) indica que los enfoques basados en reglas implican la creación de sistemas que utilizan reglas lingüísticas para analizar el texto, estas se derivan del conocimiento lingüístico y se implementan en gramáticas y diccionarios formales. De igual forma, Manning y Schütze (1999) afirman que, estos enfoques de métodos basados en reglas pueden ofrecer una buena precisión en tareas específicas, su definición y mantenimiento manual requieren un alto esfuerzo y pueden no ser tan escalables como las técnicas estadísticas o de ML.

##### 2.1.1.1 Aplicaciones del NLP

NLP ha sido implementado en diferentes dominios, uno de los usos más conocidos y utilizados es en la *traducción automática*, donde cada vez se observan avances para lograr una mayor precisión. Estos modelos, utilizan grandes corpus y arquitecturas avanzadas (ej., transformers) para capturar de una mejor manera la tonalidad de los idiomas, mejorando la calidad en las traducciones en el tiempo (Sajjad et al., 2013; Vaswani et al., 2017). Otra importante aplicación del NLP son los *asistentes virtuales*, en ellos se utiliza el lenguaje natural para interpretar y responder preguntas de los usuarios que interactúan con él, tradicionalmente en lenguaje natural. Este tipo de sistemas emplean técnicas como el reconocimiento de emociones y la extracción de entidades (NER, por sus siglas en inglés), las cuales utilizan la extracción de información para identificar y obtener datos relevantes. En este mismo orden de ideas, se encuentra la *generación automática de textos*, su aplicación facilita la creación de contenidos apropiados al contexto de

los datos de entrada; su utilidad se centra en la creación de contenido, generación de resúmenes y de texto personalizado a partir de grandes textos (Rush et al., 2015; Dong et al., 2019; Brown et al., 2020).

En el ámbito de la clasificación, se aplica para *categorizar textos* en diversas clases predefinidas. Los algoritmos de clasificación de texto, como los modelos de bolsa de palabras (Bag-of-Words) y los clasificadores basados en redes neuronales, son fundamentales para tareas como la categorización de correos electrónicos en spam y no spam, la clasificación de artículos de noticias por temas y la organización de documentos legales por categorías específicas (Taha, 2024). Para el análisis de sentimiento, el NLP permite determinar la polaridad emocional de un texto, es decir, si el contenido expresa una opinión positiva, negativa o neutral. Esta técnica se basa en la identificación de palabras y frases con carga emocional, y en el análisis de su contexto dentro del texto. Los modelos de análisis de sentimiento son utilizados en el monitoreo de redes sociales, donde permiten a las empresas evaluar la percepción pública de sus productos y servicios (Medhat et al, 2014).

Otra de las aplicaciones del NLP, se puede identificar en el *modelado de tópicos*, la cual agrupa palabras en temas basados en sus co-ocurrencias en un corpus de textos. Algoritmos como Latent Dirichlet Allocation (LDA) y Non-negative Matrix Factorization (NMF) son empleados para descubrir temas en grandes colecciones de documentos. Este enfoque es útil en la minería de textos y en la organización de información en bibliotecas digitales, permitiendo a los investigadores identificar y estudiar tendencias sin necesidad de leer cada documento individualmente.

### **2.1.2 Procesamiento de lenguaje natural en el análisis de discurso**

El análisis de discurso mediante NLP se centra en comprender la estructura y el significado de textos extensos y complejos. Joty et al. (2019) afirma que, con el NLP, es posible identificar y analizar estructuras lingüísticas complejas, facilitando tareas como la resolución de correferencias, el reconocimiento de entidades nombradas y la detección de relaciones entre segmentos de texto. Estas capacidades son importantes para el análisis de discurso, ya que permiten inspeccionar la cohesión, la coherencia y las relaciones semánticas dentro de un texto. Adicionalmente, el uso de técnicas de NLP en el análisis de discurso permite automatizar la identificación de temas, el análisis de sentimientos y la estructura narrativa de textos largos (ej., discursos políticos, documentos académicos y conversaciones cotidianas, Crossley et al., 2014). Las herramientas de NLP pueden procesar grandes volúmenes de texto, ofreciendo una perspectiva cuantitativa que complementa los métodos cualitativos tradicionales del análisis de discurso (Öhman & Metcalfe, 2021).

Según Mochtak (2024) en el análisis de discursos políticos, el NLP puede revelar patrones de lenguaje que indican persuasión, manipulación o cambio de tema, por ejemplo, el análisis de discursos políticos en Filipinas ha demostrado cómo el NLP puede identificar características estilísticas y retóricas específicas, proporcionando una visión más amplia de las estrategias discursivas empleadas por los políticos.

Además, el NLP basado en reglas ha sido aplicado para analizar textos en lenguas específicas, como el maratí, donde se utilizan técnicas de resolución de correferencias y reconocimiento de entidades nombradas para indagar las estructuras del discurso (Khandale & Mahender, 2020).

En el contexto educativo, la aplicación del NLP en el análisis de discurso puede ayudar a evaluar la claridad y efectividad de la comunicación escrita, permitiendo a los educadores y estudiantes identificar áreas de mejora y desarrollar habilidades comunicativas más efectivas (Crossley, 2012). Las técnicas de NLP también pueden facilitar la creación de herramientas de retroalimentación automática, mejorando así la calidad del aprendizaje y la enseñanza.

### **2.1.3 Principales tareas de análisis de discurso apoyadas por NLP y lingüística computacional**

El análisis de discurso en el ámbito del procesamiento del lenguaje natural (NLP) y la lingüística computacional incluye diversas tareas que permiten una comprensión detallada y estructurada del texto. A continuación, se describen las principales tareas de análisis de discurso, utilizando soporte teórico de Huber y Carenini (2022):

*Frecuencia de palabras:* Se emplea para identificar las palabras, frases o n-grams más utilizados en un conjunto de datos. Esta técnica se utiliza para detectar patrones léxicos recurrentes en el discurso. Se representa gráficamente mediante histogramas que muestran la frecuencia de aparición de cada término en orden ascendente o descendente.

*Nube de palabras:* Visualiza la frecuencia de términos en un formato donde el tamaño de cada palabra es proporcional a su frecuencia. Esta técnica facilita la identificación visual de los temas más frecuentes en el texto y es útil para la exploración preliminar de grandes conjuntos de datos.

*Construcción de árboles de constituyentes del discurso:* Organiza las Unidades de Discurso Elementales (EDU) en una estructura jerárquica, reflejando las relaciones semánticas y pragmáticas entre ellas. Utiliza la programación dinámica CKY y el algoritmo de Eisner para generar árboles de constituyentes y de dependencias a partir de las matrices de autoatención de los modelos de transformadores como BERT y BART.

*Agrupación:* La agrupación permite organizar respuestas en grupos basados en similitudes, facilitando la identificación de temas comunes y patrones. Huber y Carenini (2022) discuten la importancia de capturar estructuras discursivas en los modelos pre-entrenados, lo que puede mejorar la precisión de la agrupación de respuestas basadas en las relaciones semánticas entre EDUs.

*Clasificación:* Asigna una categoría específica a una respuesta dada, utilizando modelos pre-entrenados para tareas específicas. El uso de LLMs como BERT y

BART, afinados en conjuntos de datos etiquetados, permite una clasificación precisa de las respuestas basadas en la estructura discursiva capturada.

*Análisis de sentimientos:* Determina el sentimiento de un texto (positivo, negativo o neutro). Los modelos pre-entrenados capturan información discursiva que puede mejorar la precisión del análisis de sentimientos al proporcionar un contexto adicional que ayuda a determinar el sentimiento correcto en el análisis del texto.

*Detección de emociones:* La detección de emociones extiende el análisis de sentimientos al identificar una gama más amplia de emociones. Los modelos como BERT y BART, que capturan relaciones discursivas complejas, pueden ser afinados para detectar múltiples categorías emocionales en el texto.

*Identificación de entidades:* La identificación de entidades clasifica entidades mencionadas en el texto, como personas, lugares y organizaciones, ayudando a entender los temas discutidos.

*Preguntas y respuestas (QA):* La tarea de preguntas y respuestas (QA) permite realizar preguntas sobre el conjunto de datos y obtener respuestas precisas. Los modelos afinados para QA, pueden extraer información relevante del texto utilizando la estructura discursiva capturada.

*Resúmenes:* Generar resúmenes de documentos largos, extrayendo la información más relevante. La capacidad de los LLMs para capturar grandes estructuras discursivas permite crear resúmenes más precisos y coherentes, mejorando la comprensión y el análisis de grandes volúmenes de texto.

## **2.2 ESTADO DEL ARTE**

### **2.2.1. Avances en LLM**

#### **2.2.1.1. Embeddings**

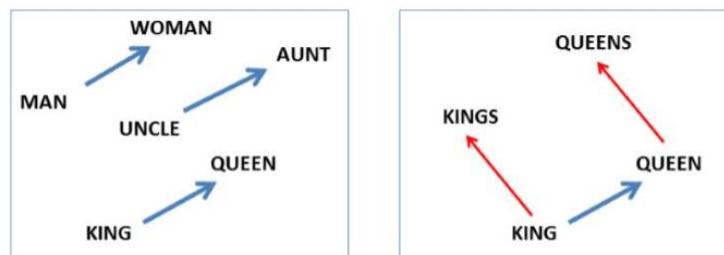
Los embeddings de palabras son representaciones vectoriales en un espacio continuo que capturan relaciones semánticas y contextuales entre las palabras. Estas representaciones permiten que palabras con significados similares tengan vectores similares al convertir palabras en vectores numéricos, donde las similitudes semánticas se reflejan en las proximidades de los vectores en el espacio. Varma (2019) considera que la conversión de palabras en vectores numéricos facilita su uso en modelos de aprendizaje automático, redes neuronales y modelos generativos para diversas tareas de procesamiento del lenguaje natural (NLP), como la clasificación de texto, el análisis de sentimientos y la traducción automática. Es importante asegurar que las representaciones vectoriales capturen adecuadamente todos los aspectos lingüísticos de una palabra para el éxito de las aplicaciones de estas técnicas.

La base teórica de los embeddings de palabras se encuentra en la teoría de la semántica distribucional, que postula que el significado de una palabra puede inferirse a partir de los contextos en los que aparece (Lenci, 2018). Esta teoría es respaldada por la hipótesis distribucional, que sugiere que palabras con significados similares ocurren en contextos similares (Boleda, 2020). Varma (2019) asegura que técnicas populares como Word2Vec, GloVe y FastText implementan esta idea al entrenar modelos en grandes corpus de texto, aprendiendo representaciones vectoriales que capturan estas relaciones contextuales.

En términos matemáticos, los embeddings transforman palabras en vectores en un espacio de alta dimensión. En este espacio, las distancias y direcciones entre vectores reflejan similitudes y relaciones semánticas, por ejemplo, en un buen espacio de embeddings, la distancia entre los vectores de "rey" y "reina" debería ser similar a la distancia entre "hombre" y "mujer" (ver Figura 2, Arseniev-Koehler, 2021). Lenci (2018) comenta que estos modelos han demostrado ser efectivos en capturar muchos aspectos del significado de las palabras, facilitando la comprensión y manipulación del lenguaje natural por parte de las máquinas.

---

**Figura 2.** Relaciones lingüísticas modeladas por desplazamiento de vector lineal.



Fuente:

Varma, K. K. (2019) Deconstructing Word Embedding Models.

---

En su análisis, Varma (2019) compara los modelos de embeddings basados en conteo y los basados en predicción, señalando que estos últimos tienden a capturar mejor las relaciones semánticas complejas debido a su capacidad para predecir contextos. Esta comparación es relevante para entender por qué los modelos como word2vec, son preferidos en muchas aplicaciones de NLP. Baroni et al. (2014) realizaron una comparación entre los modelos basados en conteo y los modelos predictivos, concluyendo que los embeddings generados mediante técnicas de predicción, como word2vec, obtienen un rendimiento superior en una variedad de tareas de semántica léxica.

### **2.2.1.2. Modelos generativos**

Los modelos generativos son una clase de algoritmos de aprendizaje automático diseñados para generar datos nuevos y similares a los datos de entrenamiento. Su tarea principal es aprender la distribución subyacente de los datos para generar muestras nuevas que sigan la misma distribución. Estos modelos se utilizan en una variedad de aplicaciones, incluyendo la generación de imágenes, texto, música y video (Vardanyan et al., 2024). Uno de los principios de funcionamiento de estos modelos es la teoría de la probabilidad, donde se intenta estimar la probabilidad de ocurrencia de diferentes eventos en un espacio determinado. Algunos de los modelos más conocidos en esta categoría incluyen las redes generativas adversariales (GANs), los modelos de autocodificadores variacionales (VAEs).

Las GANs, introducidas por Goodfellow et al. (2014), utilizan dos redes neuronales en competencia: una generadora que crea datos falsos y una discriminadora que intenta distinguir entre datos reales y falsos. El objetivo de la generadora es engañar a la discriminadora produciendo datos cada vez más realistas. Este proceso adversarial mejora continuamente la calidad de los datos generados. Estas redes tienen aplicaciones en campos como generación de imágenes, mejora de imágenes, creación de datos sintéticos, videojuegos y animación (Ororbia & Kifer, 2022; Sharma et al., 2022). Por otro lado, los VAEs, son modelos generativos que utilizan una estructura de red neuronal compuesta por un codificador (Encoder) y un decodificador (Decoder), se basan en principios bayesianos para aprender una representación latente de los datos, permitiendo la generación de nuevas muestras al muestrear desde esta representación latente (Kingma & Welling, 2019).

Por su parte, los transformadores generativos, tales como GPT-3 y GPT-4, utilizan una arquitectura basada en transformers para la generación de texto de manera coherente y contextual. Estos modelos son entrenados con grandes corpus de datos de texto y aprenden las distribuciones de probabilidad de las secuencias de palabras. Están fundamentados en el mecanismo de atención introducido por Vaswani et al. (2017), los modelos GPT consideran el contexto completo de una secuencia para predecir la siguiente palabra. Este proceso se realiza a través de capas de transformers que implementan múltiples cabezas de atención y normalización sobre los datos de entrada. Las aplicaciones de estos modelos son diversas y cada vez más crecientes, entre las que se incluye generación de texto, generación de imágenes a partir de texto, y videos permitiendo la creación de contenido textual que mantiene coherencia y contexto. Además, se aplican en la traducción automática, facilitando la traducción entre distintos idiomas con alta precisión. De igual manera, se emplean en el desarrollo de asistentes virtuales y chatbots avanzados, mejorando la interacción hombre-máquina mediante respuestas consideradas adecuadas y coherentes por los usuarios (Achiam et al., 2023; Minaee et al., 2024).

Los modelos generativos han revolucionado diversas tareas en el ámbito del procesamiento de lenguaje natural. En la generación de texto, estos modelos

permiten la creación automática de contenido coherente y relevante, desde párrafos cortos hasta artículos completos, manteniendo un flujo lógico y una estructura adecuada (Bhardwaj et al., 2020). En la traducción automática, modelos como GPT-3 y GPT-4 facilitan la traducción precisa entre múltiples idiomas, comprendiendo tanto el contexto cultural como el idiomático, lo que mejora la calidad de las traducciones. Además, los chatbots y asistentes virtuales utilizan modelos generativos para mantener conversaciones naturales y contextualmente adecuadas con los usuarios, proporcionando respuestas a preguntas, recomendaciones y realizando tareas específicas, mejorando así la experiencia del usuario en aplicaciones de servicio al cliente y soporte técnico (Sun & Ortiz, 2024).

Estos modelos no solo se aplican en el contexto del NLP, se han encontrado aplicaciones en otros campos. En la generación de imágenes, estos modelos crean imágenes realistas a partir de descripciones textuales, siendo útiles en el diseño, el entretenimiento y las simulaciones (Baldassarre et al., 2023). Así mismo, Gozalo-Brizuela y Garrido-Merchán (2023) plantean la aplicación de estos modelos en la música, donde pueden componer nuevas melodías y armonías basadas en grandes bases de datos de piezas musicales, imitando estilos de compositores específicos, también en la generación de datos sintéticos es fundamental para entrenar modelos de aprendizaje automático, cuando los datos reales son escasos o costosos de obtener, o presentan problemas de privacidad. Los datos sintéticos permiten mejorar la robustez y la generalización de los modelos al proporcionar conjuntos de datos más diversos y balanceados.

### **2.2.1.3. Prompt engineering**

El prompt engineering es una técnica utilizada en los modelos de lenguaje para diseñar y optimizar las entradas o "*prompts*" que se utilizan para interactuar con estos. El propósito de esta técnica es mejorar la calidad y la relevancia de las respuestas generadas por los modelos de lenguaje, (por ej, Llama3, GPT-4). Esta técnica se utiliza en aplicaciones como la generación de texto, la traducción automática y los asistentes virtuales, donde obtener resultados precisos y contextualmente adecuados es muy importante.

El prompt engineering se fundamenta en principios teóricos relacionados con el aprendizaje supervisado, no supervisado y generativo. En el aprendizaje supervisado, los modelos se entrenan utilizando pares de entrada y salida etiquetados, lo que permite ajustar los prompts para obtener respuestas específicas y predecibles (Korzynski et al., 2023). En el aprendizaje no supervisado, los modelos se entrenan con datos sin etiquetar, lo que requiere diseñar prompts que puedan guiar al modelo a descubrir patrones y estructuras en los datos de manera autónoma (Zhang et al., 2024). Por su parte, el aprendizaje generativo, implica entrenar modelos para generar nuevos datos a partir de distribuciones aprendidas, y el

prompt engineering se utiliza para afinar estas capacidades generativas y mejorar la creatividad y coherencia de las respuestas (He et al., 2024).

En el contexto de los modelos de lenguaje, el prompt engineering también se relaciona con técnicas de aprendizaje por refuerzo, donde los prompts se optimizan a través de interacciones repetidas con el modelo y ajustes basados en retroalimentación (Franceschelli & Musolesi, 2024). El prompt engineering se ha convertido en una habilidad para optimizar la interacción con los LLMs en sistemas de inteligencia artificial, y la alfabetización en IA influye en la calidad de estos prompts y su aplicación en diferentes contextos (Knoth et al., 2024).

Se han estudiado varias técnicas que optimizan la interacción y los resultados de los modelos a partir del prompt Eng. A continuación, se detallan algunas de ellas:

*Cadena de Pensamiento (Chain of Thought, CoT):* Esta técnica CoT desglosa problemas en pasos lógicos para guiar el razonamiento del modelo. Para Wei et al (2022) estas son algunas de sus tipos:

- Zero-Shot CoT: instruye al modelo a pensar secuencialmente sin ejemplos previos, empleando mecanismos internos de generación para descomponer problemas.
- Manual CoT: proporciona ejemplos detallados de razonamiento, mejorando la precisión del modelo mediante el uso de plantillas específicas de razonamiento.

*Árbol de Pensamiento:* Esta técnica permite al modelo examinar diversas líneas de razonamiento simultáneamente, evaluando cada una para converger en la respuesta más coherente. Yao et al. (2024) afirman que esta técnica utiliza estructuras de datos de tipo árbol para gestionar diferentes trayectorias de pensamiento, optimizando la toma de decisiones en problemas complejos.

*Autoconsistencia:* Esta técnica genera varias respuestas a una misma consulta y evalúa la consistencia entre ellas. Manakul et al. (2023) utilizan métricas como la superposición de n-gramas y la similitud semántica basada en embeddings para medir la coherencia y fiabilidad de las respuestas generadas.

*Reflexión:* Permite que el modelo evalúe y revise sus propias salidas, mejorando la precisión y la coherencia. Este proceso iterativo utiliza mecanismos de retroalimentación y ajuste para corregir errores y optimizar la calidad de las respuestas (Shinn et al., 2023).

*Prompting Experto:* Simula respuestas de especialistas combinando múltiples perspectivas de expertos, empleando técnicas de ensamblado y síntesis para

producir respuestas informadas y completas. Este método utiliza modelos de transferencia de conocimiento para integrar conocimiento especializado en la generación de respuestas (Zhang et al., 2023).

*Cadenas:* Estructura tareas complejas en una secuencia de pasos interconectados. Cada componente en la cadena realiza una función específica, utilizando enfoques modulares que permiten el procesamiento secuencial y detallado de tareas complejas mediante la división en componentes manejables (Wu et al., 2022).

*Ingeniería Automática de Prompts:* Automatiza la creación y optimización de prompts utilizando los propios modelos para generar, evaluar y refinar las entradas iterativamente. Emplea algoritmos de búsqueda y optimización para identificar y ajustar los mejores prompts, mejorando continuamente la eficiencia y calidad de las respuestas generadas (Zhou et al., 2022).

Por consiguiente, Zhao et al. (2023) señalan que, cada técnica ofrece enfoques específicos para mejorar la interacción con modelos de lenguaje generativos, ajustándose a diversas necesidades y contextos de aplicación, y maximizando el potencial de los LLMs en tareas complejas y específicas.

#### **2.2.1.4. RAG - Retrieval-Augmented Generation**

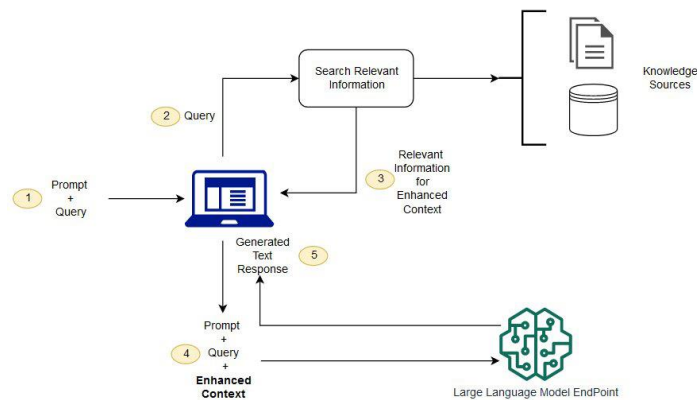
RAG se refiere a una arquitectura que combina componentes de recuperación y generación (Lewis et al., 2020). En un enfoque típico de RAG, el modelo primero recupera documentos pertinentes de una base de datos o corpus utilizando técnicas de búsqueda y recuperación de información. Posteriormente, estos documentos recuperados se emplean como contexto para un modelo generativo, como un transformador, que produce la respuesta final. Igualmente, Lewis et al. (2020) indican que la integración de ambos componentes permite una búsqueda precisa y rápida a la par de generar un texto fluido y coherente. De esta manera, el sistema proporciona respuestas más informadas y detalladas. Por un lado, en el contexto de asistentes virtuales y chatbots, los RAG mejoran la capacidad de estos sistemas para responder a preguntas complejas y variadas. Al recuperar información precisa y contextual de una amplia base de datos, el modelo generativo produce respuestas más adecuadas a las necesidades del usuario.

Desde otro enfoque, como en la investigación y consulta de documentos, los RAG pueden asistir a profesionales en campos como la medicina, el derecho y la investigación científica para encontrar y sintetizar información de manera rápida, facilitando la toma de decisiones informadas (Lewis et al., 2020). Además, en sistemas de recomendación, los RAG pueden generar recomendaciones personalizadas basadas en la recuperación de información y la generación de descripciones o resúmenes adaptados a las necesidades del usuario. En entornos educativos, los RAG pueden proporcionar respuestas precisas a preguntas de los

estudiantes, ofreciendo explicaciones y material adicional que mejora el aprendizaje y la comprensión.

El modelo RAG propuesto por Lewis et al. (2020), en su artículo “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks” demuestra la eficacia de este enfoque, mostrando mejoras en tareas de respuesta a preguntas y generación de texto basado en conocimientos. Esta arquitectura abre nuevas posibilidades para la interacción hombre-máquina y establece un marco sólido para futuras investigaciones y aplicaciones en el procesamiento del lenguaje natural. Estos modelos representan un avance metodológico en el procesamiento del lenguaje natural, integrando técnicas de recuperación de información con modelos generativos para mejorar la precisión de las respuestas generadas (ver Figura 3).

**Figura 3.** Sistema RAG con LLM para una aplicación de QA.



Fuente:

A. W. Services. (2023) Question answering using retrieval augmented generation with foundation models in amazon sagemaker jumpstart. Accessed: Date of access, e.g., December 5, 2023. [Online]. Available: <https://shorturl.at/dSV47>

En avances recientes un desarrollo de un sistema RAG, presentado por Gao et al. (2023), presenta algunas mejoras en la integración y optimización de estos sistemas. Este trabajo aborda limitaciones previas relacionadas con la eficiencia computacional y la escalabilidad, proponiendo un enfoque que optimiza la fase de recuperación de documentos mediante técnicas de indexación avanzadas y el uso de embeddings más compactos. Adicionalmente, esta investigación explora la combinación de RAG con técnicas de aprendizaje por refuerzo, optimizando la interacción entre la recuperación y la generación para maximizar la utilidad de las respuestas en aplicaciones específicas. Esto resulta útil en escenarios donde la adaptabilidad y la precisión son críticas, como en sistemas de diagnóstico médico y asesoramiento legal. Según Gao et al. (2023), la integración de técnicas de aprendizaje continuo y autorregulado, junto con mejoras en la eficiencia

computacional, posiciona a los modelos RAG como una herramienta robusta y adaptable para el procesamiento de lenguaje natural en contextos cada vez más exigentes y dinámicos

#### **2.2.1.5 Aplicaciones en el análisis de discurso.**

La integración de técnicas avanzadas como embeddings, modelos generativos, prompt engineering y RAG proporciona un enfoque potente para el análisis de discurso. Estas técnicas permiten descomponer, interpretar y generar texto con un alto grado de precisión.

En el análisis de discurso, los embeddings permiten mapear palabras y frases en espacios vectoriales, facilitando la identificación de relaciones semánticas y contextuales entre diferentes partes del texto (Benara et al., 2024). Esto es relevante para tareas como la detección de tópicos, el análisis de sentimientos y la identificación de estructuras narrativas.

Los modelos generativos, como Llama3 de Meta y GPT-4 de OpenAI, permiten la generación de texto coherente y adaptativo al contexto del discurso (Naveed et al., 2023). Esto es útil para crear simulaciones de diálogo, resúmenes y la generación de contenido que refleje el estilo y tono del discurso que está siendo analizado.

El prompt engineering por su parte mejora la interacción con los modelos generativos, permitiendo generar instrucciones específicas que guían el análisis hacia aspectos particulares del discurso, como la identificación de argumentos, la cohesión textual y la detección de inconsistencias (Barua, 2024).

Finalmente, los RAGs enriquecen el análisis de discurso al permitir que el modelo recupere y utilice información externa relevante durante el proceso de generación. Esto es particularmente útil para incorporar contextos históricos y referencias culturales que son relevantes para el análisis de discurso (Megahed et al., 2023).

### 3 DESARROLLO DEL TRABAJO

En este capítulo, se presentan cuatro modelos que emplean técnicas avanzadas de NLP. Se detalla su aplicación en el análisis del discurso, utilizando datos del programa "TQHC". Se detalla cómo se realiza la implementación y aplicación directa de estos modelos. Además, se desarrollará un modelo de clasificación de texto enfocado en predecir la etiqueta de la variable de indicador verbal, asignada manualmente por el entrevistador durante la entrevista, aunque este último modelo no se aplica directamente en el análisis de discurso, pretende automatizar el proceso de etiquetado. En conjunto, este trabajo se enfoca en la exploración y aplicación de técnicas avanzadas de NLP tanto para el análisis de discurso como para la tarea de etiquetado.

El presente trabajo se suma a los estudios implementados en el programa "TQHC" donde otros autores han contribuido mediante análisis de lingüística computacional, y otros desde el modelado de técnicas de aprendizaje automático. A través del uso de técnicas avanzadas de NLP se busca aportar al análisis de datos de texto, mejorando su interpretación y aplicación en el análisis de discurso.

En el apartado de modelación, se presentan los diferentes modelos implementados en este proyecto, junto con las técnicas utilizadas para su desarrollo. A continuación, en la Tabla 1 se detalla cada uno de ellos:

---

---

**Tabla 1:** Resumen modelos y técnicas

| Modelo Tarea                 | Técnicas Utilizadas                               |
|------------------------------|---------------------------------------------------|
| Modelo de tópicos            | Modelos generativos prompt engineering            |
| Modelo de sentimientos       | Modelos generativos + prompt engineering          |
| Agrupamiento y visualización | Embeddings, clustering, PCA, modelos generativos  |
| Asistente conversacional     | RAGs, técnicas combinadas                         |
| Modelo de clasificación      | Embeddings, modelos generativos, ML clasificación |

Fuente: Elaboración propia

---

---

#### 3.1. ENTENDIMIENTO Y COMPRESIÓN DEL NEGOCIO

Al revisar los avances logrados con los datos del programa, particularmente en el análisis lingüístico y el uso de *modelos de aprendizaje automático*, se plantea utilizar modelos grandes de lenguaje para apoyar tareas de análisis de discurso y clasificación. Estos modelos se aplicarán directamente a los datos brindados por el programa para realizar análisis de discurso y de desarrollar un modelo de clasificación que etiquete los indicadores verbales, que son etiquetas asignadas manualmente por los entrevistadores. Estos indicadores reflejan clasificaciones

basadas en la emoción, el valor y las consecuencias positivas o negativas de las respuestas (ver tabla 2).

**Tabla 2** Indicadores verbales y descripción

| Indicadores Verbales | Significado               | Descripción                                                      |
|----------------------|---------------------------|------------------------------------------------------------------|
| EM                   | Apela a emociones         | Manifiesta o expresa una emoción al responder las preguntas.     |
| RE                   | Apela a Reglas            | Sustenta las respuestas con normas sociales o constitucionales.  |
| BI                   | Apela a Fines             | El participante menciona un propósito o fin específico.          |
| CL                   | Clasificaciones           | Clasifica o etiqueta a la persona responsable en sus respuestas. |
| CN                   | Consecuencias Negativas   | Indica una emoción si no se logra el cambio o mejora esperado.   |
| NUL                  | Ninguna de las anteriores | La respuesta no encaja en ninguna categoría definida.            |
| VA                   | Valor                     | El participante menciona un valor.                               |

Fuente: Elaboración propia a partir de datos

### 3.2. ENTENDIMIENTO Y COMPRESIÓN DEL DATO

En este capítulo, se abordará el entendimiento y la comprensión de los datos utilizados, los cuales han sido preprocesados en etapas anteriores dentro de la iniciativa "TQHC". Este trabajo se suma a los esfuerzos colaborativos orientados a analizar y entender las percepciones y opiniones de los participantes y el análisis de discurso con estos resultados. Los datos utilizados provienen directamente del programa y han sido preparados a nivel de formato por el equipo de la iniciativa en etapas previas.

La estructura de la encuesta que da origen a la base de datos del proyecto "TQHC" se organiza en varios ciclos de preguntas. En el Ciclo 1, se plantean preguntas fundamentales como "¿Qué debemos cambiar en Colombia?", "¿Qué debemos mejorar en Colombia?" y "¿Qué debemos mantener en Colombia?", seguidas por una pregunta de justificación ("¿Por qué?"). El Ciclo 2 profundiza en los temas mencionados, solicitando a los participantes que identifiquen cuál consideran más importante y expliquen sus razones. Finalmente, el Ciclo 3 explora más a fondo las soluciones propuestas y la confianza en las organizaciones. Para este trabajo de tesis, nos enfocaremos en las respuestas dadas por los usuarios en el Ciclo 1.

A continuación, la siguiente tabla proporciona una pequeña muestra de las principales variables asociadas al Ciclo 1:

---

---

**Tabla 3** Muestra de datos empleados (ciclo 1)

| ID   | c1_preg  | c1_resp                           | c1_raz                            | Variables sociodemograficas | c1_iv |
|------|----------|-----------------------------------|-----------------------------------|-----------------------------|-------|
| 8725 | Cambiar  | La violencia                      | Muchos padres violan a sus hijos. | .....                       | VA    |
| 6744 | Mejorar  | Nosotros debemos mejorar la salud | Porque es mala.                   | .....                       | VA    |
| 8320 | Mantener | La biodiversidad                  | Porque es el origen de la vida.   | .....                       | NUL   |

Fuente: Datos del programa TQHC

---

---

En la tabla 3 se pueden identificar cada una de las variables del ciclo 1, adicionalmente se describen todas a continuación:

- ID: Número anonimizado que identifica al participante.
- c1\_preg: Pregunta del Ciclo 1 que el participante eligió responder, con valores posibles: 1 (Cambiar), 2 (Mejorar), 3 (Mantener), 0 (Sin respuesta).
- c1\_resp: Respuesta original del participante sobre lo que quiere cambiar, mejorar o mantener en Colombia.
- c1\_raz: Razón o motivo de la respuesta dada en c1\_resp.
- tipo\_doc\_id: Tipo de documento de identidad presentado por el participante (CC, TI, CE, PEP, Indocumentado).
- género: Género del participante (Masculino, Femenino, No binario, No sabe/No responde).
- grupo\_etario: Grupo etario del participante.
- grupo\_étnico: Grupo étnico al que pertenece el participante.
- victima\_conflicto: Si el participante ha sido víctima del conflicto armado.
- reside\_col: Si el participante reside en Colombia.
- país\_residencia\*\*: País donde reside el participante.
- ciudad: Ciudad donde reside el participante.
- departamento: Departamento donde reside el participante.
- zona: Zona (urbana o rural) donde reside el participante.
- sisben: Si el participante es beneficiario del Sisben.
- escolaridad: Nivel de escolaridad del participante.
- uso\_datos\_personales: Si el participante acepta las políticas de uso de datos personales.
- municipio: Municipio donde reside el participante.
- c1\_iv: Indicadores verbales anotados manualmente en c1\_raz, como emociones, reglas, fines, clasificaciones, consecuencias negativas o ninguna de las anteriores.

De esta manera, se ha explorado el entendimiento y la comprensión de los datos utilizados en este trabajo, resaltando su preprocesamiento en etapas anteriores dentro de la iniciativa "TQHC". Se ha detallado la estructura de la encuesta que da origen a la base de datos, enfocándose especialmente en el Ciclo 1, donde se centrarán los análisis de esta tesis. La tabla resultante contiene 54 columnas y 4766 filas, correspondientes al mismo número de participantes efectivos. De igual manera, se han enumerado las principales variables asociadas al Ciclo 1 que serán las que se utilizarán para los modelos y las variables sociodemográficas, de las cuales se utilizarán género, zona, grupo etario, región y grupo étnico para analizar los resultados de los modelos y describir mejor el análisis de discurso. Esto establece un marco para el análisis posterior de las percepciones y opiniones de los participantes.

### **3.3. PREPARACIÓN DE DATOS**

La preparación de los datos para este trabajo se realiza considerando los diferentes enfoques de modelado que se emplearán. Cada enfoque, utilizando técnicas de NLP avanzadas, requiere un tratamiento específico de los datos que se presentará a detalle en cada apartado. El punto de partida, como se mencionó anteriormente, son los datos utilizados y preprocesados por el programa "TQHC".

### **3.4. MODELADO DE DATOS**

#### **3.4.1. Modelado de tópicos utilizando Prompt Engineering y LLMs**

En este apartado, se diseña y construye un enfoque para identificar los temas más relevantes en las respuestas de los usuarios mediante el análisis de tópicos. Las respuestas utilizadas son las correspondientes al Ciclo 1. Para esta tarea, se utilizaron técnicas de prompt engineering y embeddings para procesar las respuestas, lo que permitió la creación de indicaciones específicas en el contexto para la aplicación de modelos generativos.

El modelo se implementó mediante el uso de la API OpenAI, la tarea dada fue calcular la distribución temática de las respuestas en función de indicaciones de contexto. Este enfoque permite la identificación de temas relevantes y aquellos que no son evidentes a partir del texto sin procesar, proporcionando una comprensión más amplia de la estructura y la organización de las respuestas.

El modelado de tópicos es útil en este contexto, ya que permite la identificación de patrones y relaciones sutiles. La siguiente figura presenta el proceso de construcción del modelado mediante la API de OpenAI, empleando LLMs para la identificación de temas en las respuestas de los usuarios

---

**Figura 4.** Función de clasificación de tópicos usando API de OpenAI.

```
def clasificacion_topicos(mensaje: str):
    with open('contexto_topicos.txt', 'r') as file:
        contexto = file.read()
    # Define mensaje
    prompt = f"""
    Respuesta: {mensaje} \n
    Por favor, proporciona los principales temas mencionados en la respuesta, sin el prefijo
    "Tema:" y sin saltos de línea. Los temas deben estar separados por comas.
    """
    mensajes = [
        {
            "role": "system",
            "content": contexto
        },
        {
            "role": "user",
            "content": prompt
        }
    ]
    # Crea mensaje
    response = openai.ChatCompletion.create(
        model="gpt-3.5-turbo",
        messages=mensajes,
        max_tokens=50
    )
    return response['choices'][0]['message']['content']
```

Fuente: Elaboración propia

---

El uso de la API de OpenAI para este caso implica varios pasos. Primero, se crea y carga un archivo de contexto que guía al modelo en la comprensión y procesamiento de las respuestas de los usuarios. A partir de esto, se diseña un prompt que estructura la información de manera que el modelo pueda interpretarla correctamente. Los mensajes, que combinan el contexto con el prompt, se envían a la API de OpenAI. Utilizando modelos LLMs como GPT-3.5-turbo o GPT4, la API procesa estas respuestas para generar resultados bajo las instrucciones establecidas. El archivo que contiene el contexto para la modelación de tópicos, que utiliza técnicas de *prompt engineering* se muestra a continuación:

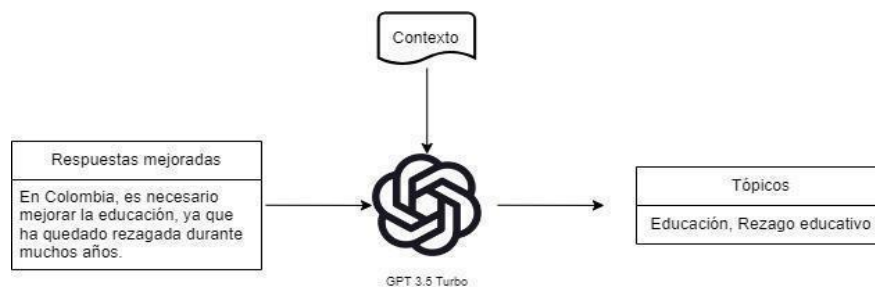
**Figura 5.** Contexto para asignar tópicos utilizando prompt Engineering.

```
'contexto_topicos.txt':  
  
En calidad de experto en análisis de texto, se te encarga evaluar las respuestas en la columna  
`ciclo1_completo` de un conjunto de datos. Estas respuestas contienen opiniones y sugerencias sobre  
diversos temas. Tu tarea consiste en identificar y extraer los principales temas abordados en cada  
respuesta.  
  
Para cada respuesta, vas a identificar lo siguiente: Entre 1 y 2 temas principales mencionados en  
la respuesta. Si hay más de un tema, deben estar separados por una coma.  
  
Aquí te dejo unos ejemplos del formato:  
  
Respuesta: En Colombia debemos cambiar nuestra visión social del voto voluntario para mejorar la  
educación.  
Tema: Voto voluntario, Educación  
  
Respuesta: Deberíamos mejorar la infraestructura de transporte público para reducir el tráfico en  
las ciudades.  
Tema: Infraestructura, Transporte público  
  
Respuesta: Es importante mantener la salud mental como una prioridad en nuestras políticas públicas.  
Tema: Salud mental, Políticas públicas  
  
La salida debe tener la estructura de los ejemplos anteriores, es decir:  
Tema:
```

Fuente: Elaboración propia

La siguiente figura presenta el proceso para generar los temas mediante la integración del contexto y las respuestas:

**Figura 6.** Proceso general para asignar tópicos a respuestas.



Fuente: Elaboración propia

El desarrollo y ejecución del modelo de clasificación de tópicos mediante la API de OpenAI comienza con la carga del archivo de contexto que tiene las instrucciones

para el modelo. La función *clasificacion\_topicos* recibe un mensaje como entrada y define un *prompt* que incluye la respuesta del usuario, solicitando al modelo la extracción de temas principales. Luego, se construye una lista de mensajes que integra el rol del sistema, con el contenido del contexto, y el rol del usuario, con el contenido del prompt. Estos mensajes se envían a la API de OpenAI mediante el método *ChatCompletion.create*, utilizando el modelo GPT-3.5-turbo para procesar la entrada y devolver una lista de tópicos identificados a la respuesta enviada.

### 3.4.2. Modelado de Sentimiento utilizando Prompt Engineering y LLMs

En este apartado, se diseña y construye un modelo de análisis de sentimientos para determinar la polaridad de las respuestas recopiladas.

El proceso de modelado se hizo mediante la conexión de API de OpenAI, permitiendo el uso de modelos generativos para calcular la polaridad de las respuestas a partir del prompts. Este enfoque aprovecha la capacidad de estos modelos para comprender el texto en diferentes contextos, proporcionando así una evaluación de la polaridad emocional expresada por los usuarios en sus respuestas.

En la Figura 7 y 8, se sintetiza la función utilizada para el cálculo de la polaridad, así como el diagrama ilustrativo para el mismo

---

---

**Figura 7.** Función para calcular sentimiento usando API OpenAI.

```
def sentimiento(text):
    prompt = f"Evalúa el sentimiento de este texto en español y devuelve una puntuación de polaridad como un número entre -1 y 1, donde -1 es muy negativo, 0 neutro y 1 muy positivo: '{text}'. Solo devuelve el número sin texto adicional."
    try:
        response = openai.ChatCompletion.create(
            model="gpt-3.5-turbo",
            messages=[
                {"role": "system", "content": "Eres un asistente útil."},
                {"role": "user", "content": prompt}
            ],
            max_tokens=10,
            n=1,
            stop=None,
            temperature=0.2
        )
        result = response.choices[0].message['content'].strip()
        return float(result)
    except ValueError:
        # Si no se puede convertir a float, intenta extraer el número del texto
        try:
            return float(''.join(filter(lambda x: x.isdigit() or x == '.' or x == '-', result)))
        except ValueError:
            return 0.0
    except openai.error.OpenAIError as e:
        # Manejo de errores de OpenAI
        print(f"Error de OpenAI: {e}")
        return None
```

Fuente: Elaboración propia

---

---

---

**Figura 8.**Proceso general para asignar sentimiento a respuestas.



Fuente: Elaboración propia

---

Para este caso la función *sentimiento* recibe un mensaje como entrada y define un *prompt* que incluye la respuesta del usuario, solicitando al modelo evaluar el sentimiento del texto entre -1 y 1. Este mensaje se envían a la API de OpenAI mediante el método *ChatCompletion.create*, utilizando el modelo GPT-3.5-turbo para procesar la entrada y devolver un valor numérico del sentimiento de la respuesta enviada.

### 3.4.3. Agrupamiento y visualización

Una de las tareas más importante de análisis del discurso, es el proceso de agrupar por algún nivel de similitud respuestas dentro del conjunto de datos recolectados. Una vez se ha desarrollado e implementado algún algoritmo de agrupamiento, pueden venir otros procesos de uso, uno de los cuales es la visualización de los grupos además de poder extraer conceptos, temas, entidades comunes a cada grupo. El proceso de visualización para este trabajo comienza con particionamiento de la base de datos en función de la primera pregunta del Ciclo 1, donde los usuarios responden si quieren cambiar, mantener o mejorar algo. Se realizó así para analizar objetivamente las distribuciones de los temas según la primera respuesta. Luego se realiza la generación de embeddings a partir de respuestas de usuarios. Los embeddings, que son representaciones vectoriales derivadas del texto a analizar, capturan las características semánticas y contextuales de las respuestas para posteriormente ser utilizadas en diferentes enfoques, como es el caso de esta etapa.

Una vez obtenidos estos vectores, se realiza un modelado de clustering para segmentar las respuestas. Se empleo la técnica de Kmeans de la librería Sklearn, el cual es un algoritmo que agrupa los datos buscando separarlos en N grupos de varianza igual, minimizando un criterio llamada Inercia, o la suma de cuadrados dentro del grupo. La métrica de distancia utilizada por defecto es la euclidiana para medir la similitud entre los puntos y los centroides. La selección de K óptimo se realiza utilizando el método del codo, el cual permite identificar el punto de inflexión en la curva de suma de distancias (intra-cluster), facilitando así una segmentación más precisa de los datos de embeddings.

Después del proceso de clustering se aplica una técnica de reducción de dimensionalidad llamada Análisis de Componentes principales (PCA, por sus siglas en inglés). El PCA es una técnica de reducción de dimensionalidad que transforma un conjunto de variables en uno más pequeño de variables no correlacionadas, también llamadas componentes principales (Khurana, 2023; Drikvandi & Lawal, 2023). Esta técnica se usa en conjuntos de altas dimensiones en los datos ya que reduce el número de dimensiones o variables mientras conserva la mayor parte de la variabilidad de los datos originales. Este proceso mejora la eficiencia computacional buscando no perder información relevante de los datos, el cual puede ser configurado por el usuario (Khurana, 2023; Drikvandi & Lawal, 2023). Para este caso, el PCA transforma los vectores de embeddings de alta dimensión en un espacio de menor dimensión. Este proceso es un paso importante en este apartado, ya que va a permitir representar visualmente los datos en espacios bidimensionales y tridimensionales utilizando 2 y 3 componentes principales respectivamente, conservando las relaciones estructurales y semánticas entre los puntos de datos.

Con los clústeres definidos y las dimensiones reducidas, se procede a la generación de etiquetas para cada clúster mediante el uso de modelos generativos de OpenAI. Para esto se utiliza una función que recibe las etiquetas individuales de cada respuesta generada en el apartado de modelado de tópicos y crea una macro etiqueta o tema principal para cada cluster que se está analizando

En la Figura 9 y 10, se presenta la función utilizada para el cálculo de las macro etiquetas que será aplicada a cada cluster así como parte del contexto utilizado en para instruir al modelo generativo.

---

**Figura 9.** Función para construir macro etiquetas a partir de API de OpenAI.

```
def clasificacion_etiquetas(mensaje: str):
    with open('contexto_visualizacion.txt', 'r') as file:
        contexto = file.read()
    prompt = f"""
Macroetiqueta: {mensaje} \n
"""
    mensajes = [
        {
            "role": "system",
            "content": contexto
        },
        {
            "role": "user",
            "content": prompt
        }
    ]

    # Crea mensaje
    response = openai.ChatCompletion.create(
        model="gpt-3.5-turbo",
        messages=mensajes,
        max_tokens=50
    )
    return response['choices'][0]['message']['content']
```

Fuente: Elaboración propia

---

---

**Figura 10.** Contexto utilizado para función de macro etiquetas.

'contexto\_visualizacion.txt':

Como experto en análisis de texto, se te encarga evaluar una serie de etiquetas de temas principales de conversaciones previamente creadas. Estas etiquetas reflejan temas individuales que deben agruparse bajo una macroetiqueta representativa. Tu tarea consiste en identificar y agrupar estas etiquetas para formar una macroetiqueta que represente el conjunto de temas individuales.

Para cada grupo de etiquetas, vas a identificar lo siguiente:

La macroetiqueta que representa el conjunto de temas individuales.  
Aquí te dejo unos ejemplos del formato:

Etiquetas: Educación, Regulación de mecanismos de control  
Macroetiqueta: Educación y Regulación

Etiquetas: Mentalidad centralista, Equilibrio global  
Macroetiqueta: Gobernanza y Equilibrio Global

Etiquetas: Participación de científicos en políticas públicas, Educación, Ciencia y tecnología  
Macroetiqueta: Ciencia y Políticas Públicas

Para la salida de la respuesta sigue las siguientes los ejemplos anteriores, es decir:  
Macroetiqueta:

Fuente: Elaboración propia

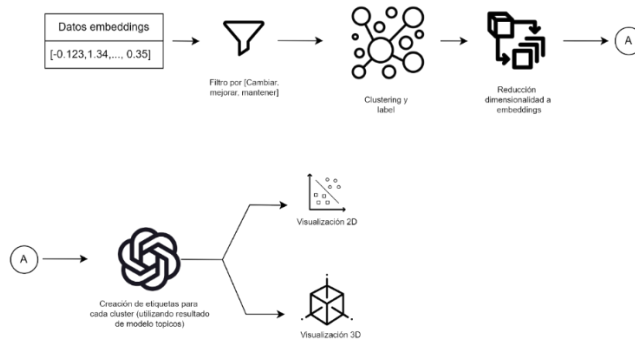
---

Una vez se tiene las etiquetas de cada cluster calcula a partir de la función `clasificación_etiquetas`, se construyen las visualizaciones interactivas en 2D y 3D utilizando librerías de Python como Plotly. Estas visualizaciones permiten a los usuarios analizar dinámicamente los datos, aplicando filtros interactivos por las etiquetas que se desean analizar. La capacidad de interactuar con estas visualizaciones de los datos de texto de opinión proporciona a los equipos de análisis o interesados una comprensión de la distribución espacial de los temas, encontrando relaciones y patrones que podrían no ser evidentes mediante métodos de análisis tradicionales.

El proceso general para construir las visualizaciones 2D y 3D se presenta mediante un diagrama ilustrativo a continuación (ver Figura 11):

---

**Figura 11.** Proceso general para agrupamiento y visualización.



Fuente: Elaboración propia

---

#### 3.4.4. Asistente conversacional

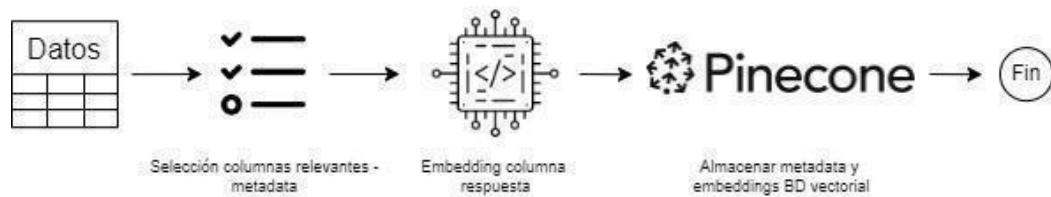
Para este trabajo se construyó un prototipo práctico de asistente conversacional que permite interactuar con la base de datos de encuestas de opiniones. La relevancia está en la capacidad de convertir datos no estructurados en conversaciones accesibles, permitiendo a los usuarios obtener información y analizar resultados de manera intuitiva. A diferencia de los métodos tradicionales de NLP, que limitan la interacción dinámica con los datos debido a su enfoque estático y preprocesamiento fijo este asistente emplea técnicas avanzadas para superar estas limitaciones. El proceso para la creación de este prototipo se basa en técnicas de Retrieval Augmented Generation (RAG, por sus siglas en inglés). Estas técnicas combinan métodos de recuperación de información y modelos generativos para construir respuestas contextualizadas de acuerdo con las necesidades de los usuarios. Los embeddings y los modelos generativos desempeñan un papel importante para este proceso, permitiendo que el asistente recupere información precisa y genere respuestas coherentes a las consultas de los usuarios. Sin embargo, estos modelos pueden sufrir de alucinación, generando contenido ficticio o inexacto. Para mitigar este fenómeno, se introduce contexto real proveniente de la base de datos, mejorando la precisión y la fiabilidad de las respuestas generadas.

El uso de técnicas RAG en asistentes conversacionales en este trabajo busca mejorar la capacidad de estos sistemas para abordar preguntas complejas y dar respuestas detalladas y precisas en el análisis del discurso. Esta técnica empleada (RAG) se basa en la creación de embeddings para representar semánticamente los fragmentos de texto de opiniones recuperados, almacenados en una base de datos vectorial que facilita las búsquedas, de manera que sean más eficientes y precisas. Tras recuperar los fragmentos pertinentes, estos sirven para guiar un modelo generativo, como Llama3, un modelo de lenguaje desarrollado por Meta especializado en generar texto coherente y relevante basado en el contexto proporcionado. Este enfoque permite construir respuestas contextualizadas y coherentes, ajustando las respuestas a diversas consultas de los usuarios. En el

ámbito del análisis de discurso, esta capacidad ofrece insights detallados de las opiniones ciudadanas, superando las limitaciones de los métodos tradicionales de NLP que tienen una interacción estática con los datos.

El proceso de selección de las columnas relevantes, creación de embeddings y almacenamiento de estos datos en una base de datos vectorial se ilustra en la Figura 12.

**Figura 12.** Proceso de creación y almacenamiento de embeddings.

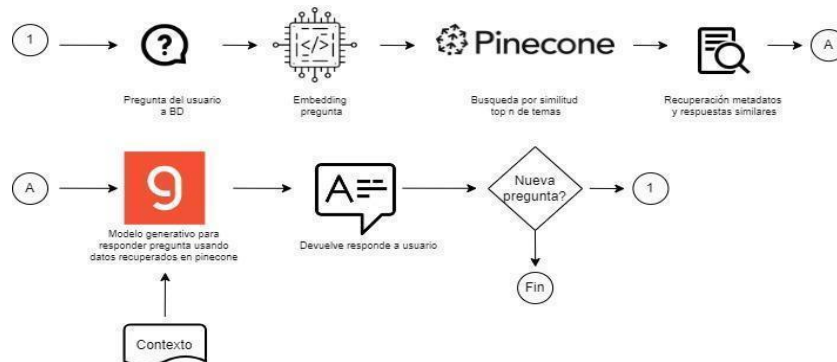


Fuente: Elaboración propia

Este proceso se corre y almacena una única vez con la información de la encuesta, en caso de venir ediciones posteriores o en caso tal se quiera actualizar la información deberá volver a correrse a demanda.

Luego, cuando se va a interactuar con el asistente, el flujo del proceso desde que el usuario interesado ingresa la pregunta hasta que se responde por la IA se presenta en la Figura 13.

**Figura 13.** Proceso general de funcionamiento del asistente.



Fuente: Elaboración propia

Cada vez que el usuario formula una pregunta, el sistema convierte la pregunta en un embedding, que luego se envía a Pinecone (sistema de almacenamiento de la

base de datos vectorial) para realizar una búsqueda por similitud entre los temas. Pinecone recupera metadatos y respuestas similares, que son utilizados por el modelo generativo para elaborar una respuesta contextualizada. Este flujo se repite con cada nueva consulta, asegurando que todos los pasos se completen para proporcionar una respuesta precisa y relevante al usuario.

Para obtener las respuestas, la función utiliza el modelo generativo LLama3, desarrollado por Meta, distinto a los utilizados en otras técnicas presentadas en este trabajo. Aunque sigue utilizando un método de llamada de API, permite configurar el rol que el modelo debe asumir. En el contexto de este trabajo, se instruye al modelo para que actúe como *asesor político*, analizando las opiniones ciudadanas recuperadas de la base de datos vectorial para proponer políticas públicas. A continuación, la Figura 14 presenta la función que crea el contexto.

---

**Figura 14.** Función para crear contexto completo.

```
def create_prompt(similarities: List[dict]) -> str:
    base_prompt = """
    Actúa como un asesor político importante.
    Tu función es analizar las respuestas y opiniones de ciudadanos sobre diversos temas abiertos.
    Los ciudadanos han indicado si consideran que se debe "Cambiar", "Mejorar" o "Mantener" el
    estado actual, y han proporcionado una explicación detallada de su respuesta.
    Tu tarea es agrupar las respuestas en tres categorías: "Cambiar", "Mejorar" o "Mantener", y
    diseñar una política pública en un párrafo máximo para cada categoría basada en las opiniones y
    explicaciones de los ciudadanos.
    Estas son las opiniones de los ciudadanos:
    {reasons}
    """
    reasons = ""

    for response in similarities:
        reasons += f"""
        {response["metadata"]["c1_preg"]}:          {response["metadata"]["c1_resp"]}          Razón:
        {response["metadata"]["c1_razon"]}
        """
    return base_prompt.format(reasons=reasons)
```

Fuente: Elaboración propia

---

La salida de esta función crea el contexto completo que será ingresado a la función que hace el llamado al modelo generativo Llama3, como se presenta a continuación (ver Figura 15).

---

**Figura 15.** Función para hacer llamado al modelo con contexto completo.

```
def get_response(query: str) -> str:
    # Get the prompt
    similarities = get_similar_text(query)
    prompt = create_prompt(similarities)

    client = Groq(api_key=os.getenv("GROQ_API_KEY"))
    completion = client.chat.completions.create(
        model="llama3-70b-8192",
        messages=[
            {"role": "user", "content": prompt},
        ],
        temperature=1,
        max_tokens=1024,
        top_p=1,
        stop=None,
    )
    return completion.choices[0].message.content
```

Fuente: Elaboración propia

---

### 3.4.5. Modelado de clasificación de texto

Dentro del conjunto de datos, se observa que las respuestas de los participantes contienen un indicador verbal (C1\_VA) asignado manualmente por el entrevistador, el cual clasifica las respuestas según emociones, valores, fines o consecuencias negativas. En este apartado, se construyen modelos de clasificación para predecir la etiqueta del indicador verbal utilizando tres aproximaciones con diferentes técnicas de procesamiento de datos seguidos por clasificadores tradicionales. Estas aproximaciones permitirán examinar y comparar la eficacia de técnicas avanzadas de procesamiento de lenguaje natural en la predicción de indicadores verbales, buscando mejorar la precisión del modelo final.

- *Aproximación 1:* Clasificador de Texto para el Indicador Verbal del Ciclo 1 – Mediante el uso de vector embeddings, configuración original de Fajardo Becerra (2023).

Para este primer enfoque, se utilizó la misma configuración de referencia, tanto en modelos de clasificación como en la estructuración de los datos. En esta oportunidad, el objetivo es, construir clasificadores empleando esta configuración, pero en lugar de utilizar datos tabulares, se utiliza la técnica de vector embeddings para la codificación de los datos. Los vectores de embeddings fueron construidos mediante el uso de la API de OpenAI.

Una vez obtenido el vector de embeddings, se procedió a utilizar los mismos clasificadores de la configuración de referencia con el fin de comparar el rendimiento de los modelos mediante métricas de clasificación como precisión, recall, F1-score y accuracy.

A continuación, se presenta una vista rápida del conjunto de datos utilizado en el trabajo de Fajardo Becerra (2023) para construir los modelos de clasificación. En el trabajo de referencia estos datos tabulares fueron procesados de manera tradicional mediante la codificación categórica de las variables textuales y la codificación numérica de las variables cuantitativas (ver Tabla 4). Una vez codificados estos datos, se procedió a entrenar los clasificadores.

**Tabla 4** Conjunto de datos seleccionado para modelo de clasificación de texto

| texto_1                                            | cambiar mejorar mantener | género              | grupo etario    | Dpto   | escolaridad         | región | etiqueta |
|----------------------------------------------------|--------------------------|---------------------|-----------------|--------|---------------------|--------|----------|
| voto voluntario obligatorio revisar abstención...  | cambiar                  | No sabe/No responde | De 37 a 47 años | Bogotá | No sabe/No responde | Centro | NUL      |
| colombiano agenc interés individual colectivo ...  | cambiar                  | No sabe/No responde | De 37 a 47 años | Bogotá | No sabe/No responde | Centro | NUL      |
| visión social educación capacidad persona talen... | cambiar                  | No sabe/No responde | De 37 a 47 años | Bogotá | No sabe/No responde | Centro | VA       |

Fuente: Conjunto de datos configurado en Fajardo Becerra (2023)

Para el presente trabajo, se realizó una concatenación de todos los datos tabulares presentados en la tabla. Posteriormente, se aplicó el procesamiento de embeddings para construir un vector único, el cual fue utilizado para construir los modelos de clasificación.

La función programada para la construcción del vector de embeddings sigue la siguiente lógica (ver Figura 16).

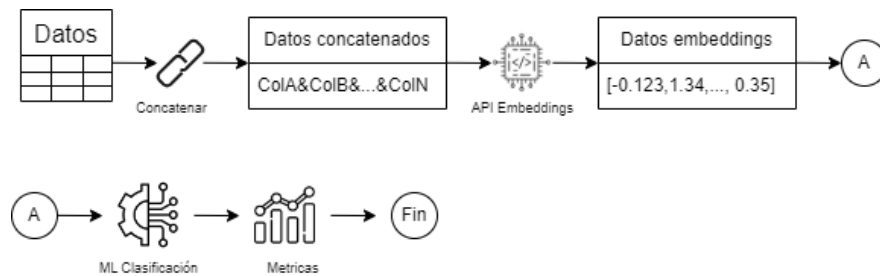
**Figura 16.** Función para construir embeddings a partir de API OpenAI.

```
def obtener_embeddings(texto):
    response = openai.Embedding.create(
        input=[texto],
        model="text-embedding-ada-002"
    )
    embeddings = response['data'][0]['embedding']
    return embeddings
```

Fuente: Elaboración propia

A continuación, se presenta un diagrama ilustrativo (ver Figura 17) de todo el proceso de esta primera aproximación.

**Figura 17.** Proceso general de clasificación con aproximación 1.



Fuente: Elaboración propia

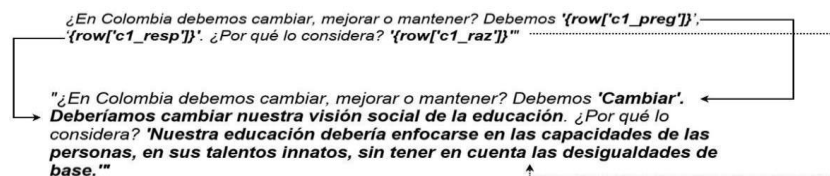
La configuración de los hiperparámetros de los modelos de clasificación se presenta en los anexos de este trabajo.

- *Aproximación 2:* Clasificador de texto para el Indicador Verbal del Ciclo 1 - Mediante el uso modelos generativos, Prompt Engineering y vector embeddings

Para este segundo enfoque, se construyó un texto unificado a partir de las respuestas de los usuarios utilizando modelos generativos. Se aplicaron técnicas de embeddings para el procesamiento del texto y se utilizaron los mismos modelos de clasificación de la aproximación anterior. El objetivo de esta aproximación es mejorar la unificación y consistencia del texto, con el fin de aumentar la precisión de los modelos de clasificación.

En principio se seleccionan las variables que contienen las respuestas textuales de los participantes y sus razones correspondientes al ciclo 1. Esto se debe a que la etiqueta que se desea predecir está directamente asociada a estas respuestas. Para unificar estas respuestas, se crea una función que concatena las respuestas de cada participante en el ciclo 1 (ver Figura 18), generando un texto más coherente

**Figura 18.** Ejemplo de construcción de texto a partir de respuestas.



Fuente: Elaboración propia

La Figura 18, consolida las concatenaciones **c1\_preg**, **c1\_resp** y **c1\_raz** que corresponden a las respuestas que dio cada usuario. Una vez que se tiene el texto base para cada participante, se aplica una función para mejorar el texto concatenado. Esta función procesa el texto anteriormente mostrado, utilizando técnicas de prompt engineering y mediante API se envía al modelo GPT-3.5 Turbo para optimizar la respuesta, manteniendo la integridad de la respuesta original de cada usuario. El uso de modelos generativos como GPT facilita la optimización en la representación y comprensión del texto, además permite una capacidad de generalización y adaptación que supera las limitaciones de las aproximaciones tradicionales de NLP. La capacidad de estos modelos generativos para capturar matices y contextos complejos es importante para mejorar la precisión tanto del entendimiento de los usuarios como para el uso de modelación (ver Figura 19).

---

**Figura 19.** Función para mejorar texto concatenado usando LLMs de OpenAI.

```
def mejorar_respuesta(respuesta):
    prompt = f"Reescribe el siguiente texto para mejorar la claridad y coherencia sin
    agregar información nueva: {respuesta}"
    response = openai.ChatCompletion.create(
        model="gpt-3.5-turbo",
        messages=[
            {"role": "system", "content": "You are a helpful assistant."},
            {"role": "user", "content": prompt}
        ],
        max_tokens=500,
        n=1,
        stop=None,
        temperature=0.5
    )
    return response.choices[0].message['content'].strip()
```

Fuente: Elaboración propia

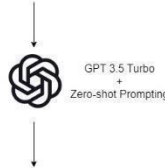
---

La Figura 20, muestra la salida de la función, tomando el mismo ejemplo anterior.

---

**Figura 20.** Ejemplo de uso de salida de función para transformar respuesta.

*"¿En Colombia debemos cambiar, mejorar o mantener? Debemos '**Cambiar**'.  
**Deberíamos cambiar nuestra visión social de la educación.** ¿Por qué lo  
considera? '**Nuestra educación debería enfocarse en las capacidades de las  
personas, en sus talentos innatos, sin tener en cuenta las desigualdades de  
base.**'"*



*"¿En Colombia deberíamos cambiar, mejorar o mantener? Considero que  
deberíamos Cambiar. Es importante modificar nuestra percepción social de la  
educación. Creo que la educación debería centrarse en las capacidades y talentos  
naturales de las personas, sin importar las desigualdades iniciales."*

Fuente: Elaboración propia

---

La respuesta mejorada se utiliza como entrada para la función de *obtener\_embedding*. Esta función convierte el texto en un vector numérico, que en este caso tiene 1536 dimensiones, representando todo el contenido del texto ingresado por respuesta. El modelo utilizado para esta tarea es el text-embedding-ada-002 de OpenAI, que se conecta mediante API (ver Figura 21).

---

**Figura 21.** Función para obtener embeddings.

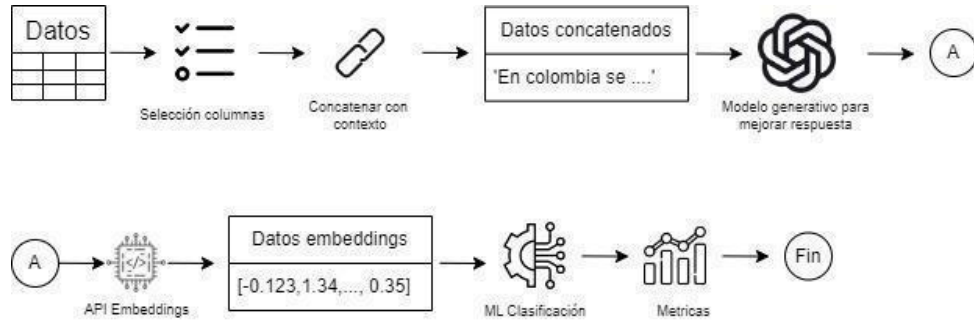
```
def obtener_embeddings (texto):  
    response = openai.Embedding.create(  
        input=[texto],  
        model="text-embedding-ada-002"  
    )  
    embeddings = response['data'][0]['embedding']  
    return embeddings
```

Fuente: Elaboración propia

---

Con los vectores generados para cada respuesta, se procede a construir los modelos de clasificación. A continuación, se presenta un diagrama ilustrativo (ver Figura 22) del proceso de esta segunda aproximación.

**Figura 22.** Proceso general de clasificación con aproximación 2.

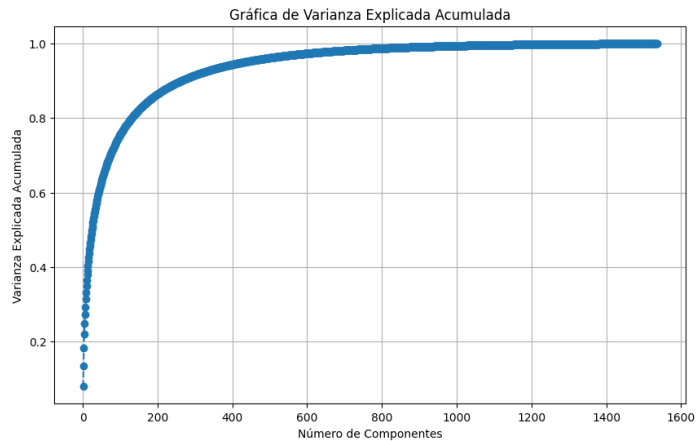


Fuente: Elaboración propia

- *Aproximación 3:* Clasificador de texto para el Indicador Verbal del Ciclo 1 - Mediante el uso de modelos generativos, prompt engineering, vector embeddings y reducción de dimensionalidad

Para este tercer enfoque, se utilizó la misma lógica del anterior, aunque con un paso adicional. Se implementó una reducción de dimensionalidad en el vector de embeddings, utilizando PCA y trabajando con 200 dimensiones. Esta selección se basó en un análisis de la variabilidad explicada, determinando que con 200 dimensiones se alcanzaba un porcentaje mayor al 80% de la variabilidad explicada. En ese sentido, se pasó de trabajar con 1536 dimensiones por cada respuesta, a 200 dimensiones. La Figura 23, muestra la varianza explicada de PCA.

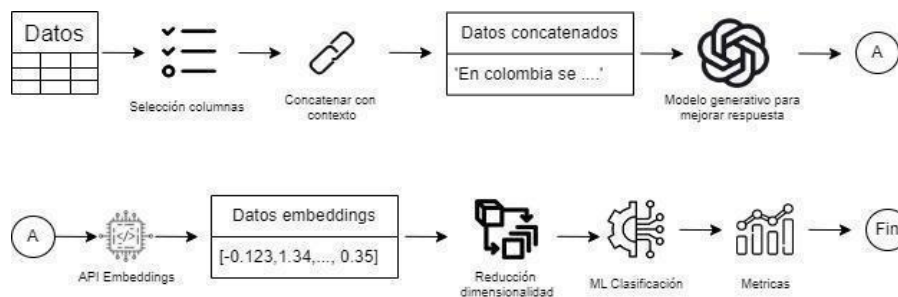
**Figura 23.** Varianza explicada en función del número de componentes principales.



Fuente: Elaboración propia

La implementación de esta reducción, especialmente para este contexto, facilita el manejo y procesamiento de los embeddings, buscando mejorar la precisión de los modelos de clasificación al eliminar redundancias y reducir el ruido en los datos. Esta aproximación es especialmente efectiva en el procesamiento de lenguaje natural, donde la alta dimensionalidad de los datos de texto está presente (Zou et al., 2006). A continuación, se presenta un diagrama ilustrativo (Figura 24) del proceso de esta tercera aproximación.

**Figura 24.** Proceso general de clasificación con aproximación 3.



Fuente: Elaboración propia

### 3.5. EVALUACIÓN

La evaluación de las técnicas empleadas en el capítulo de modelación, así como los resultados obtenidos, se presentarán con detalle en el capítulo 4 de este trabajo.

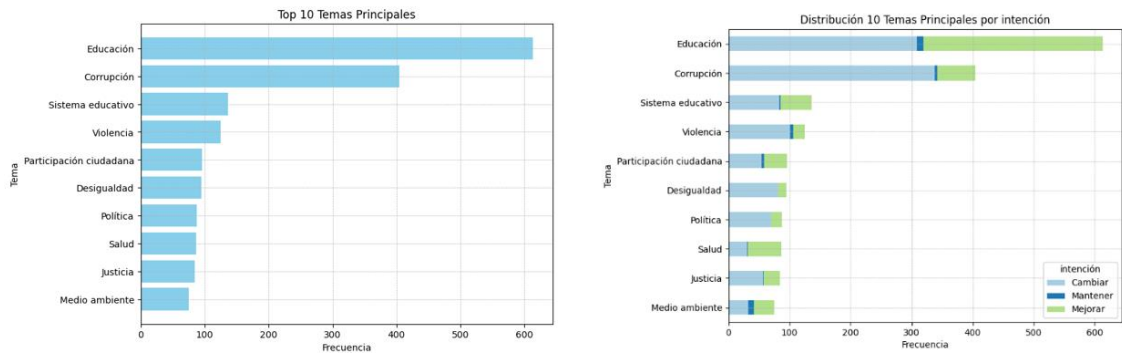
## 4 RESULTADOS

En este capítulo se presentan los análisis resultantes de la aplicación de técnicas avanzadas de modelos de lenguaje al análisis de discurso. Los resultados expuestos proporcionan una visión de las percepciones y emociones de los participantes encuestados. Cabe señalar que la interpretación de estos datos debe complementarse con la experiencia y conocimientos de equipos expertos, como el personal del proyecto "TQHC" y especialistas en políticas públicas, comunidades y humanidades. La colaboración interdisciplinaria enriquecerá la comprensión y aplicación de los hallazgos, asegurando una perspectiva más integral y contextualizada

### 4.1. MODELADO DE TÓPICOS UTILIZANDO PROMPT ENGINEERING Y LLMS

En este apartado se presentan los resultados del modelado de tópicos aplicado al análisis de discurso en el contexto del programa "TQHC", empleando técnicas avanzadas de prompt engineering y LLMs. Inicialmente, se realizó un análisis de las respuestas de los participantes para identificar los temas más mencionados en sus opiniones. Las técnicas empleadas permitieron no solo extraer los temas principales, sino también identificar patrones y relaciones que no son inmediatamente evidentes en el texto sin procesar. Este enfoque permitió una comprensión estructurada del contenido de las respuestas, proporcionando una visión de los asuntos de interés y preocupación para los ciudadanos. A continuación, se presenta la Figura 25, que posibilita la comprensión del modelado de tópicos, donde se puede apreciar la diversidad y complejidad de las percepciones recogidas en el programa.

**Figura 25.** Análisis de tópicos y temas principales.



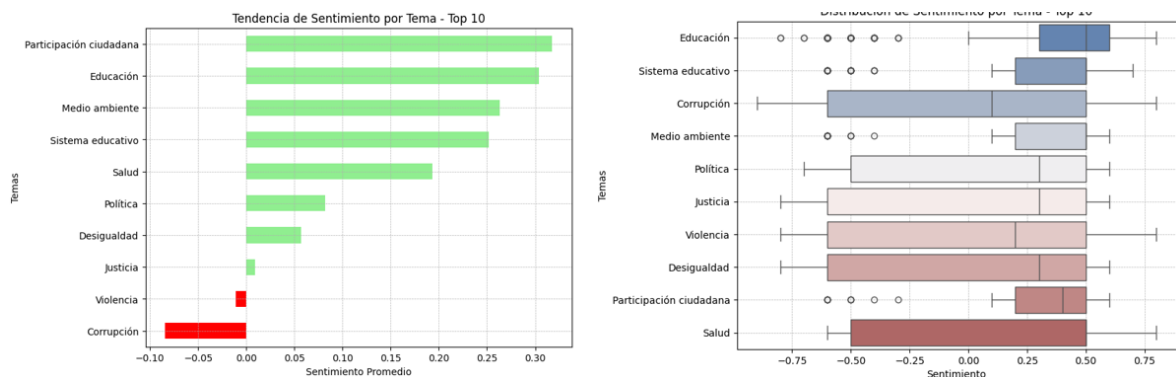
Fuente: Elaboración propia

Respecto a la frecuencia de los temas más recurrentes, en la parte superior izquierda de la Figura 25, se muestra que "Educación" y "Corrupción", son los temas con mayores menciones con un porcentaje de 5% y 3,2% respecto a todos los temas mencionados respectivamente, indicando que son temas con una alta prioridad y

preocupación entre los participantes. En lo que concierne al “*Sistema educativo*”, aunque esté relacionado con “*Educación*”, se destaca como un tema separado, señalando aspectos específicos dentro del ámbito educativo que los participantes desean abordar. Temas como “*Violencia*” (1%) y “*Participación ciudadana*” (0,8%), también tienen una presencia notable dentro de todo el conjunto de datos, sugiriendo que son aspectos clave en las preocupaciones y propuestas de los ciudadanos. Temas como *Desigualdad*, *Política*, *Salud*, *Justicia* y *Medio ambiente*, aunque menos frecuentes, representan áreas de interés considerable para los participantes, indicando una diversidad de preocupaciones. La gráfica de la derecha muestra estos temas según la intención expresada por los participantes: “*Cambiar*”, “*Mantener*” o “*Mejorar*”. La mayoría de las respuestas sobre “*Educación*” y “*Corrupción*” están asociadas con la intención de cambio, lo que indica una percepción negativa en estos ámbitos. Por otro lado, temas como “*Salud*” y “*Medio ambiente*” presentan una distribución más equilibrada entre las intenciones de “*Mantener*” y “*Mejorar*”, sugiriendo percepciones variadas entre los participantes. Este análisis por las variables sociodemográficas presentes en la base de datos permite entender cómo diferentes grupos demográficos priorizan y perciben estos temas, aportando una perspectiva más completa sobre las prioridades y percepciones en distintos segmentos de la población.

La Figura 26, consolida la tendencia de sentimientos a partir de los 10 temas principales identificados en el programa.

**Figura 26.** Análisis de tema por sentimientos top 10 temas generales.



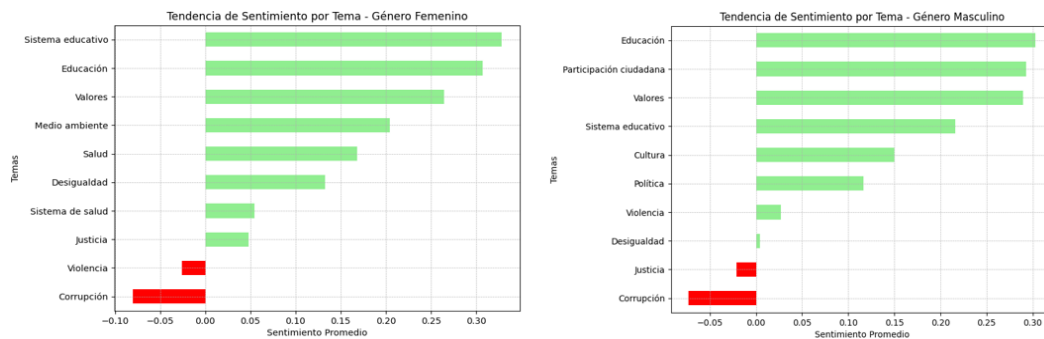
Fuente: Elaboración propia

A través de la Figura 26, se revela que “*Participación ciudadana*”, “*Educación*”, “*Medio ambiente*” y “*Sistema educativo*” tienen una tendencia de sentimiento positiva, sugiriendo una percepción favorable y reconocimiento de avances en estos ámbitos. En contraste, “*Violencia*” y “*Corrupción*” presentan una tendencia de sentimiento negativo, lo que sugiere insatisfacción y preocupación, posiblemente vinculadas a experiencias personales negativas o percepciones de ineficacia en

políticas públicas. La distribución de sentimientos por tema muestra la variabilidad en las respuestas. Aunque “Educación” y “Sistema educativo” tienen una tendencia positiva, también presentan una amplia dispersión de sentimientos, reflejando opiniones diversas. Esta variabilidad en las percepciones se contextualiza con las variables sociodemográficas de la base de datos, permitiendo entender cómo diferentes grupos demográficos experimentan estos temas. Estos resultados resaltan las áreas donde se percibe tanto necesidad urgente de cambio como oportunidades de mejora.

A partir de la Figura 27, se distribuyen los temas de acuerdo con el género de los participantes.

**Figura 27.** Análisis por género.



Fuente: Elaboración propia

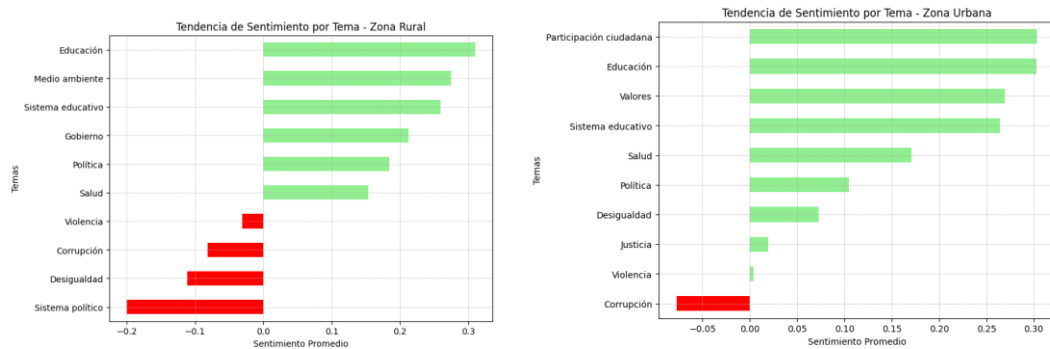
Con respecto al género “Femenino”, los temas como “Sistema educativo”, “Educación”, “Valores”, “Medio ambiente” y “Salud” tienen una tendencia de sentimiento positivo, indicando una percepción favorable hacia estos aspectos. Los temas de “Violencia” y “Corrupción” presentan una tendencia de sentimiento negativo, reflejando insatisfacción, posiblemente relacionada con experiencias personales o una percepción de mala en políticas públicas.

Para el caso del género “Masculino”, se observa una tendencia positiva en temas como “Educación”, “Participación ciudadana”, “Valores” y “Sistema educativo”, sugiriendo una valoración positiva en estos ámbitos. Sin embargo, también se observa una tendencia negativa en los temas de “Corrupción” y “Justicia”, aunque la magnitud es menor en comparación con el género femenino. Comparando los resultados de cada género, se puede observar que tanto hombres como mujeres tienen percepciones similares en cuanto a los temas con tendencias positivas, como lo son “Educación” y “Sistema educativo”. No obstante, las mujeres muestran una preocupación más pronunciada en temas como “Violencia” y “Corrupción” en comparación con los hombres, quienes tienen una percepción negativa más moderada en estos temas. Este análisis resalta las diferencias de género en la percepción y valoración de los temas, proporcionando una comprensión más

detallada de las dinámicas de género en el discurso colectivo. Además, identifica las áreas específicas donde las preocupaciones de hombres y mujeres divergen.

Por otro lado, la Figura 28 aporta un entendimiento de los temas, de acuerdo con la zona territorial de los participantes.

**Figura 28.** Análisis por zonas.



Fuente: Elaboración propia

En las zonas rurales, los temas de “Educación”, “Medio ambiente”, “Sistema educativo”, “Gobierno” y “Política” tienen una tendencia de sentimiento positivo, sugiriendo una percepción favorable hacia estos temas, reflejando expectativas de mejora y reconocimiento de avances en estas áreas. Por el contrario, temas como “Violencia”, “Corrupción”, “Desigualdad” y “Sistema político” presentan una tendencia de sentimiento negativo, reflejando insatisfacción y preocupación, posiblemente debido a experiencias de injusticia, inseguridad en la administración pública. De igual manera, en las zonas urbanas, los temas como “Participación ciudadana”, “Educación”, “Valores” y “Sistema educativo” revelan una tendencia de sentimiento positivo, indicando una valoración constructiva y posiblemente un mayor compromiso cívico y percepción de oportunidades en estos ámbitos. Por el contrario, los temas de “Corrupción” y “Violencia” muestran una tendencia de sentimiento negativo, aunque la magnitud de la negatividad es menor en comparación con las zonas rurales, lo que podría indicar una menor incidencia y/o una percepción diferente de estos problemas.

Comparando los resultados entre zonas rurales y urbanas, se observa que ambas zonas tienen percepciones positivas similares en temas como “Educación” y “Sistema educativo”. No obstante, las zonas rurales muestran una preocupación más marcada en temas de “Desigualdad” y “Sistema político”, mientras que las zonas urbanas se observa una valoración positiva de la “Participación ciudadana” y “Valores”. Esta variación en las percepciones por ubicación geográfica proporciona una visión detallada de cómo los habitantes de diferentes áreas experimentan y valoran distintos temas en el discurso, resaltando las diferencias en sus preocupaciones y prioridades. Al contextualizar estos hallazgos con las variables

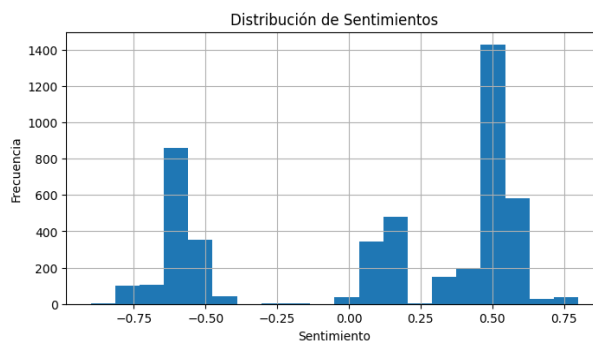
sociodemográficas se obtiene una comprensión más dinámica de los territorios en la percepción de temas relevantes dentro del discurso colectivo.

El análisis del modelado de tópicos utilizando técnicas de prompt engineering y LLMs en el programa sugieren variaciones en las percepciones de los participantes según sea el tema. Los resultados en general indican una frecuencia de menciones a temas como “Educación” (5%) y “Corrupción” (3,2%), con una marcada intención de cambio en estos ámbitos. Además, se observa una tendencia de sentimientos positivos en temas como “Sistema educativo” y “Participación ciudadana”, especialmente en zonas urbanas, mientras que temas como “Violencia” y “Corrupción”, son temas que generan sentimientos negativos más pronunciados en las zonas rurales y entre las mujeres. La capacidad de los LLMs para entender y procesar el contexto ha sido importante para detectar estas variaciones, revelando insights valiosos sobre las preocupaciones y expectativas de los ciudadanos en diferentes regiones y grupos demográficos.

## 4.2. MODELADO DE SENTIMIENTO UTILIZANDO PROMPT ENGINEERING Y LLMS

Para este trabajo se llevó a cabo un análisis de sentimientos para evaluar las emociones expresadas en las respuestas de los participantes. Utilizando técnicas de NLP y LLMs, se cuantifican los sentimientos con valores entre -1 y 1, reflejando la polaridad negativa o positiva de las respuestas (ver Figura 29). Los resultados del análisis incluyen diversas variables sociodemográficas, permitiendo segmentar los resultados según género, edad, grupo étnico y región. Este enfoque facilita la identificación de patrones emocionales y diferencias contextuales en las percepciones de los encuestados. Este estudio proporciona una comprensión de las emociones en las percepciones de los ciudadanos encuestados en el programa.

**Figura 29.** Análisis por otras categorías.

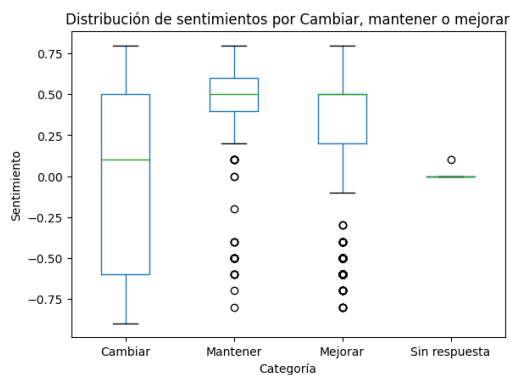


Fuente: Elaboración propia

La Figura 29, presenta el histograma de sentimientos en las respuestas de los participantes del programa. Se observan tres picos en la gráfica: uno cerca de -0.6, otro alrededor de 0.25, y el último en 0.5. Esto indica que las respuestas tienden a agruparse en torno a sentimientos negativos moderados, neutros positivos, y positivos altos. La frecuencia más alta se encuentra en el rango de 0.5, sugiriendo una tendencia de sentimientos positivos en las respuestas. Sin embargo, la presencia del pico en el rango negativo también destaca que hay una cantidad de respuestas con sentimientos negativos, reflejando una diversidad en las percepciones y emociones de los participantes.

Ahora se analizará cómo se distribuyen los sentimientos de acuerdo a la categoría "cambiar", "mejorar", "mantener", que es la primera respuesta que responden los encuestados sobre la situación actual del país (ver Figura 30).

**Figura 30.** Distribución del análisis de sentimientos.

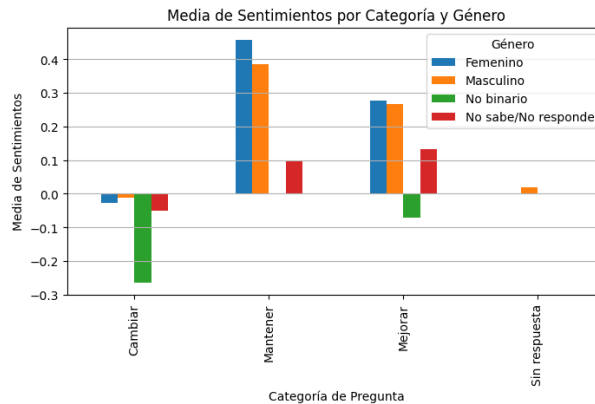


Fuente: Elaboración propia

Las respuestas en la categoría "cambiar" muestran una amplia distribución de sentimientos, representados en una mediana cercana a cero, pero con valores que se extienden hacia el rango negativo. Esto sugiere que los encuestados asocian la necesidad de cambio con sentimientos negativos, reflejando insatisfacción y un deseo de transformación en aspectos problemáticos. Por su parte en la categoría "mantener", las respuestas presentan una distribución centrada y una mediana que se inclina hacia el lado positivo de la distribución de sentimientos. Esto indica una percepción de satisfacción con los aspectos que los encuestados consideran deben mantener, reflejando estabilidad y aprobación en estos temas específicos. Finalmente, las respuestas en la categoría "mejorar", muestran una mediana positiva, con una mayor dispersión de valores hacia ambos extremos del espectro. Esto sugiere que, aunque hay una tendencia a reconocer aspectos positivos, también existe una conciencia sobre áreas que necesitan refinamiento. La categoría "Sin Respuesta" presenta pocos datos, pero con valores atípicos que podrían reflejar respuestas únicas o excepcionales.

A continuación, se examinan las variables sociodemográficas para comprender cómo diferentes grupos perciben y responden emocionalmente a los temas discutidos (ver Figura 31). Se analizan las respuestas segmentadas por género, grupo etario, región y grupo étnico. Se buscó una visión detallada de las emociones en el discurso y cómo estas varían según las características demográficas de los participantes.

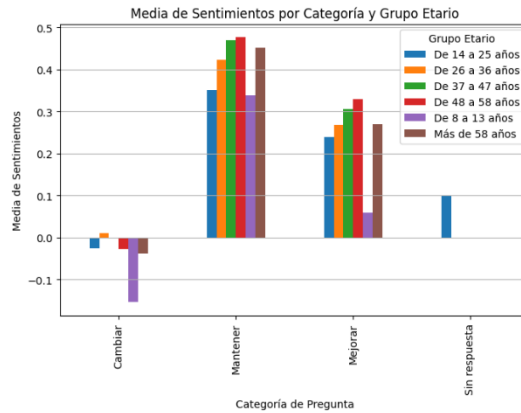
**Figura 31.** Valor medio sentimiento por género y categoría de pregunta 1.



Fuente: Elaboración propia.

En la categoría *"cambiar"*, el género masculino y no binarias muestran sentimientos más negativos, reflejando una mayor insatisfacción con los aspectos que consideran necesitan cambios, mientras que el género femenino tiene una percepción menos negativa. En *"mantener"*, todos los géneros presentan sentimientos positivos, especialmente el masculino y femenino, indicando satisfacción con el estado actual de ciertos aspectos. En la categoría *"mejorar"*, *masculino* y *femenino* expresan sentimientos positivos, sugiriendo que perciben aspectos positivos, pero reconocen la necesidad de mejoras; sin embargo, las personas no binarias tienen sentimientos negativos, lo que podría indicar un escepticismo sobre las posibilidades de mejora. Estas diferencias reflejan cómo los distintos géneros perciben y responden emocionalmente a las necesidades de cambio, estabilidad y mejora en el contexto del discurso social en Colombia.

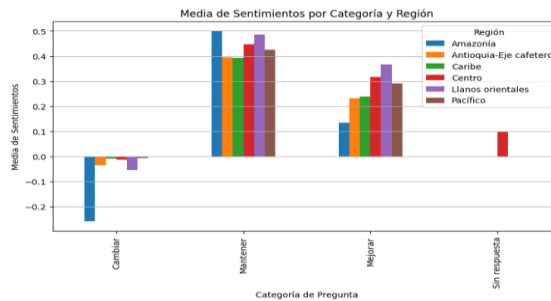
**Figura 32.** Valor medio sentimiento por grupo etario y categoría de pregunta 1.



Fuente: Elaboración propia.

De acuerdo con los resultados obtenidos en la categoría "*cambiar*", los grupos de entre 37 a 47 años y 48 a 58 años muestran medias de sentimiento negativas, indicando mayor insatisfacción en estos grupos de edad. Por el contrario, otros grupos etarios presentan sentimientos más neutros. En "*mantener*", todos los grupos etarios exhiben sentimientos positivos, con los grupos de 26 a 36 años y más de 58 años destacándose por su alta satisfacción, reflejando una percepción de estabilidad y aprobación en estos aspectos. En la categoría "*mejorar*", todos los grupos muestran sentimientos positivos, aunque con una variabilidad menor, sugiriendo un consenso general sobre la necesidad de mejoras, especialmente en los grupos de 26 a 36 años y de 37 a 47 años. La categoría "*sin respuesta*", tiene pocos datos, mientras que los encuestados de 8 a 13 años muestran una media positiva. Este análisis presenta cómo las percepciones y emociones varían según la edad, destacando la importancia de considerar el contexto etario en el análisis de las opiniones y sentimientos. A continuación, se realiza la misma distribución de los datos con respecto a la región (ver Figura 33).

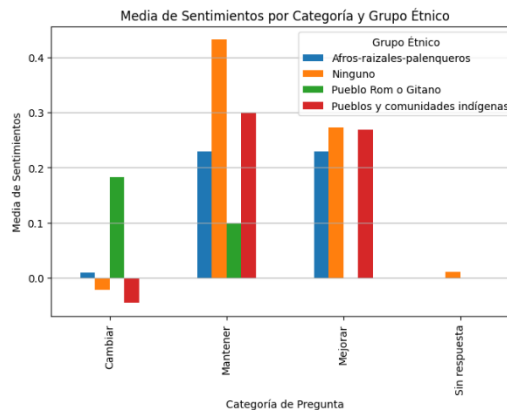
**Figura 33.** Valor medio sentimiento por Región y categoría de pregunta 1.



Fuente: Elaboración propia.

A partir de la evidencia presentada en la categoría "cambiar", la región "Amazonia", muestra una media de sentimientos notablemente negativa, indicando una mayor insatisfacción con respecto a otras regiones que presentan sentimientos más neutros. En "mantener", todas las regiones exhiben sentimientos positivos, con la "Amazonia" y el "Caribe" destacándose por su alta satisfacción. En la categoría "mejorar", los sentimientos son también positivos en todas las regiones, pero con una variabilidad menor, lo que sugiere la necesidad de mejoras, especialmente en las regiones de "Antioquia"-Eje Cafetero" y "Llanos orientales". La categoría "sin respuesta" muestra una media de sentimientos neutra "Centro", aunque este dato es aislado. Este análisis destaca cómo las percepciones y emociones varían según la región. La Figura 34, se basa en graficar cada una de las categorías con respecto a su grupo étnico.

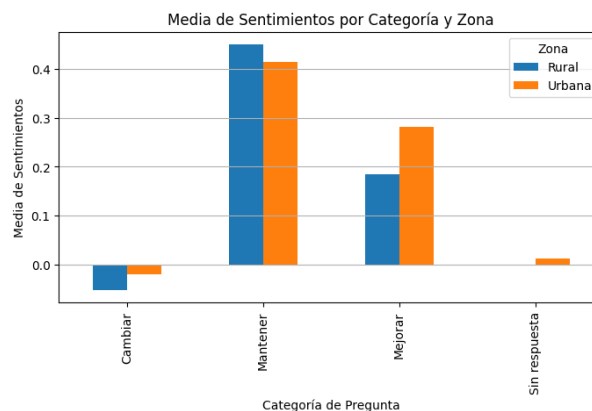
**Figura 34.** Valor medio sentimiento por grupo étnico y categoría de pregunta 1.



Fuente: Elaboración propia.

En esta nueva segmentación, en la categoría "cambiar", los sentimientos son generalmente neutros o ligeramente negativos, con el grupo "Pueblo Rom o Gitano" mostrando una percepción más negativa, lo que indica una mayor insatisfacción. En "mantener", todos los grupos presentan sentimientos positivos, con los "Afros-raizales-palenqueros" destacándose por su alta satisfacción. La categoría "mejorar" muestra sentimientos positivos en todos los grupos, aunque con variaciones en los pueblos y comunidades indígenas y en aquellos que no pertenecen a ningún grupo específico, sugiriendo un reconocimiento de aspectos positivos junto con la necesidad de mejoras. Este análisis revela cómo las percepciones y emociones varían entre los grupos étnicos, subrayando la importancia de considerar estas diferencias en el análisis del discurso colectivo. Ahora se realiza el mismo análisis según zona (ver Figura 35).

**Figura 35.** Valor medio sentimiento por zona y categoría de pregunta 1.



Fuente: Elaboración propia.

Tanto los encuestados de zonas rurales como urbanas tienen sentimientos neutros o levemente negativos en la categoría "*cambiar*", reflejando una percepción moderada de la necesidad de cambio. En la categoría "*mantener*", ambos grupos muestran un valor medio de sentimientos positivos alto, con una corta diferencia en las zonas rurales, indicando una satisfacción con el estado actual de ciertos aspectos. En "*mejorar*", los sentimientos positivos son más pronunciados en las zonas urbanas, destacando un mayor reconocimiento de la necesidad de mejoras en estas áreas. Estas diferencias reflejan cómo las percepciones y emociones varían entre las zonas rurales y urbanas.

Finalmente, este análisis de sentimientos aplicado a las respuestas del proyecto "TQHC", revela una diversidad de percepciones y emociones entre los colombianos. Los encuestados expresan una satisfacción con aspectos que consideran deben mantenerse, especialmente en regiones como la "Amazonía" y entre grupos etarios mayores, destacando la valoración de estabilidad y continuidad en ciertas áreas. Sin embargo, hay una notable insatisfacción y percepción de necesidad de cambio en categorías como "*cambiar*" y "*mejorar*", con sentimientos negativos más pronunciados en ciertas regiones y entre adultos de mediana edad. Al analizar el discurso sobre lo que se busca mejorar, se identifican áreas positivas donde los encuestados reconocen avances y éxitos, así como aspectos negativos donde perciben fallos y oportunidades de mejora. Este análisis busca mejorar la comprensión de las emociones en las conversaciones de los colombianos, destacando la importancia de considerar las variables sociodemográficas para una interpretación más completa y contextualizada. La combinación de lo positivo y negativo en las mejoras deseadas proporciona un panorama integral de las aspiraciones y preocupaciones de la sociedad colombiana.

### 4.3. AGRUPAMIENTO Y VISUALIZACIÓN

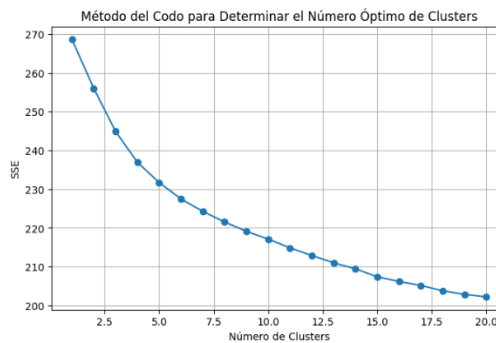
Los resultados de agrupamiento y visualización se enfocan principalmente en brindar al usuario tenga herramientas para analizar los datos de una manera visual

y, que le permitan encontrar patrones al momento al analizar datos de texto de una manera diferente, de forma que pueda encontrar patrones e insights valiosos en este tipo de datos. Los resultados se dividen en 3 partes de acuerdo con la intención principal dada en la respuesta: “Cambiar”, “mejorar” o “mantener”.

Para cada una de las intenciones, se realiza el proceso de clustering con su selección de K para cada caso. A continuación, se presenta el resultado de la selección de K óptimo para las respuestas asociadas a “cambiar”

---

**Figura 36.** Resultado método del codo para clustering de categoría “cambiar”.



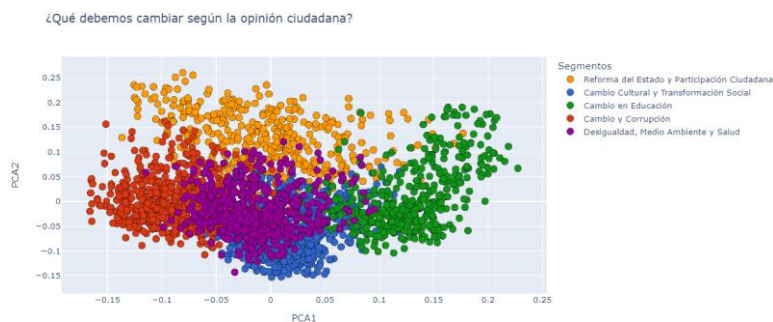
Fuente: Elaboración propia.

---

La figura anterior, muestra que la curva que comienza a aplanarse alrededor de 5 a 10 clústeres. Esto sugiere que el número óptimo de clústeres podría estar en ese rango. Para fines de este trabajo, se eligieron 5 clusters para el análisis de “cambiar”, ya que parece ser un buen equilibrio entre complejidad y adecuación del modelo. La Figura 37, representa gráficamente la pregunta, **¿Qué deberíamos cambiar? – Análisis 2D.**

---

**Figura 37.** Grafica 2D de clustering y etiquetas a respuesta de ¿Qué deberíamos cambiar?



Fuente: Elaboración propia.

---

La gráfica 37 sobre lo que los ciudadanos opinan con respecto a qué se debe cambiar, muestra cómo los temas de “*reforma del estado*”, “*cambio cultural*”, “*educación*”, “*corrupción*” y “*desigualdad*” se agrupan en segmentos interrelacionados. Los clusters de “*reforma del estado*”, “*corrupción*” y “*desigualdad*” se superponen considerablemente, sugiriendo que estos temas son percibidos de manera interconectada por los ciudadanos. Esta superposición indica la necesidad de abordar estas áreas con estrategias integradas en las políticas públicas, ya que no son vistas como cuestiones aisladas. Por otro lado, el cluster de “*educación*” se distingue de los demás, lo que implica que los ciudadanos ven la educación como un área que requiere soluciones específicas y diferenciadas. La densidad de puntos en el centro de la gráfica, representando la intersección de múltiples temas, destaca la importancia de enfoques interseccionales en las políticas públicas. Esto subraya la necesidad de considerar cómo las percepciones sobre “*participación ciudadana*”, “*cultura*”, “*corrupción*” y “*salud ambiental*” están relacionadas. Para diseñar políticas públicas efectivas y alineadas con las necesidades ciudadanas, es importante analizar estas conexiones, permitiendo la creación de estrategias transversales que aborden las complejidades de las relaciones temáticas observadas en la opinión pública.

A continuación, se presenta gráficamente la pregunta, **¿Qué deberíamos cambiar? – Análisis 3D (ver Figura 38).**

**Figura 38.** Grafica 3D de clustering y etiquetas a respuesta de ¿Qué deberíamos cambiar?

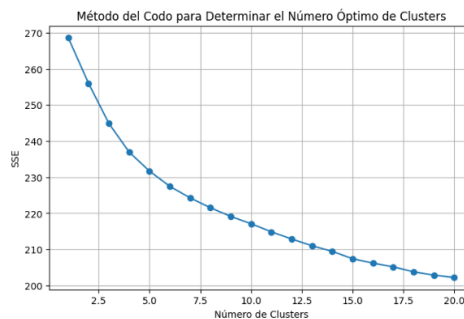


Fuente: Elaboración propia.

La Figura 38, de visualización 3D de las opiniones ciudadanas, muestra interrelaciones entre los temas de “*reforma del estado*” y “*corrupción*”, sugiriendo que las políticas públicas deben abordar estos asuntos de manera conjunta. Esta percepción de conexión indica que las estrategias integradas podrían reforzar tanto la transparencia como la participación ciudadana. En particular, el cluster referente al cambio en “*educación*” resalta la uniformidad de las percepciones ciudadanas en este tema, implicando que las políticas educativas deberían ser diseñadas de manera específica y prioritaria. Por otra parte, la dispersión observada en el cluster de “*desigualdad*”, “*medio ambiente*” y “*salud*” sugiere que las percepciones sobre estos temas son diversas y complejas. En consecuencia, las políticas públicas en este ámbito deben ser adaptativas y multitemáticas, es decir, capaces de abordar una variedad de preocupaciones y perspectivas ciudadanas. La representación 3D de las opiniones ciudadanas proporciona una vista que permite comprender mejor la complejidad y las relaciones de las percepciones sobre temas del programa.

Continuando hacia las otras categorías, se modela la pregunta **¿Qué deberíamos mejorar? – Análisis 2D Y 3D**. A continuación, se presenta el resultado de la selección de K óptimo para las respuestas asociadas a “*mejorar*”

**Figura 39.** Resultado método del codo para clustering de categoría “mejorar”.

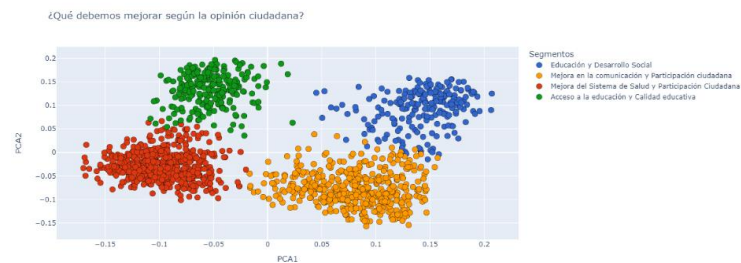


Fuente: Elaboración propia.

El gráfico anterior, muestra una disminución en SSE hasta aproximadamente 4 clusters, en el cual la pendiente se vuelve menos pronunciada. Este cambio en la tasa de disminución sugiere que 4 clusters optimizan el equilibrio entre la complejidad del modelo y la reducción del SSE.

La Figura 40, representa gráficamente la pregunta, **¿Qué deberíamos mejorar? – Análisis 2D**

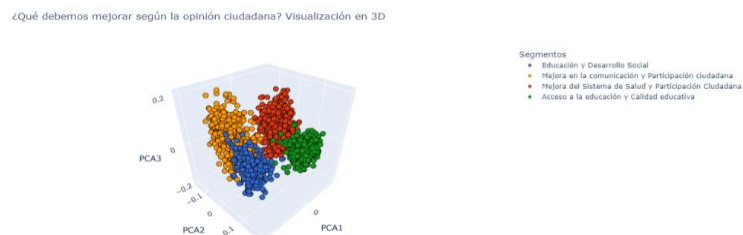
**Figura 40.** Grafica 2D de clustering y etiquetas a respuesta de ¿Qué deberíamos mejorar?



Fuente: Elaboración propia

El análisis de las opiniones ciudadanas sobre las áreas que necesitan mejora revela cuatro segmentos principales: 1) “educación” y “desarrollo social”, 2) “mejora en la comunicación” y “participación ciudadana”, 3) “mejora del sistema de salud” y “participación ciudadana”, 4) “acceso a la educación” y “calidad educativa”. El cluster de “educación” y “desarrollo social” indica una demanda por mejoras en este ámbito, reflejando expectativas de más desarrollo y apoyo educativo. La mejora en la “comunicación” y “participación ciudadana”, muestra diversidad de opiniones, con un consenso sobre la necesidad de fortalecer estos aspectos. El cluster correspondiente a la “mejora del sistema de salud” y “participación ciudadana”, sugiere que estos temas están vinculados, sugiriendo políticas integradas. El segmento de “acceso a la educación” y “calidad educativa” destaca la demanda por mejorar tanto el acceso como la calidad en el ámbito educativo.

**Figura 41.** Grafica 3D de clustering y etiquetas a respuesta de ¿Qué deberíamos mejorar?

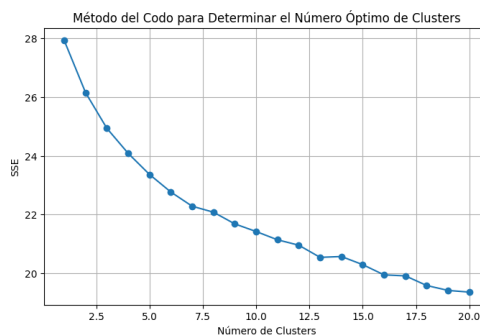


Fuente: Elaboración propia

La gráfica 3D sobre las áreas que deberían mejorarse, según las opiniones ciudadanas, muestra cuatro segmentos principales: 1) “educación” y “desarrollo social”, 2) “mejora en la comunicación” y “participación ciudadana”, 3) “mejora del sistema de salud” y “participación ciudadana”, 4) “acceso a la educación” y “calidad educativa”. El agrupamiento de “educación” y “desarrollo social”, indica una fuerte demanda por avances en este ámbito, reflejando expectativas de desarrollo y apoyo educativo más sólido. La dispersión en el segmento de “mejora en la comunicación” y “participación ciudadana” sugiere una variedad de opiniones, con un acuerdo sobre la necesidad de fortalecer estos aspectos. El segmento de “mejora del sistema de salud” y “participación ciudadana” sugiere una conexión entre estos temas, implicando que las políticas deben abordarse de manera conjunta. El “acceso a la educación” y “calidad educativa”, presenta la necesidad de mejorar tanto el acceso como la calidad en el ámbito educativo. Esta tipo de visualización 3D (ver Figura 41) busca facilitar la comprensión de cómo las percepciones ciudadanas sobre las áreas que requieren mejora están interconectadas y estructuradas, ofreciendo una perspectiva clara de las prioridades en el contexto de desarrollo social y participación comunitaria.

Finalmente, se modela la pregunta **¿Qué deberíamos mantener? – Análisis 2D Y 3D**. A continuación, se presenta el resultado de la selección de K óptimo para las respuestas asociadas a “mantener”

**Figura 42.** Resultado método del codo para clustering de categoría “mantener”.



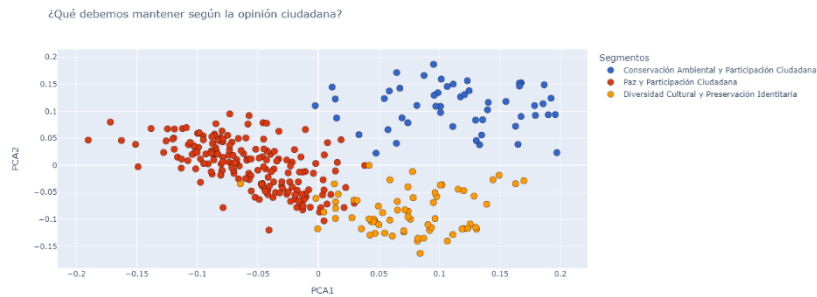
Fuente: Elaboración propia.

El gráfico anterior indica que utilizar 3 clusters ofrece una segmentación adecuada de los datos, optimizando el balance entre la simplificación del modelo y la disminución de la SSE, sin incurrir en una complejidad adicional innecesaria.

La Figura 43, representa gráficamente la pregunta, **¿Qué deberíamos mantener? – Análisis 2D**.

---

**Figura 43.** Grafica 2D de clustering y etiquetas a respuesta de ¿Qué deberíamos mantener?.



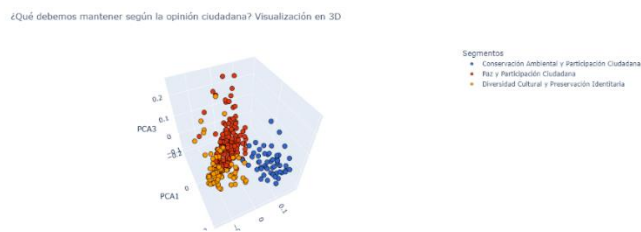
Fuente: Elaboración propia.

---

La gráfica 2D, sobre las áreas que los ciudadanos opinan que se deben mantener, revela tres segmentos principales: 1) “conservación ambiental” y “participación ciudadana”, 2) “paz” y “participación ciudadana”, 3) “diversidad cultural” y “preservación identitaria”. El segmento de “*conservación ambiental*” y “*participación ciudadana*” sugiere una percepción sobre la necesidad de conservar estos aspectos, reflejando un compromiso hacia la sostenibilidad y la participación en asuntos ambientales. Por su parte, el segmento de “*paz*” y “*participación ciudadana*” indica una variedad de opiniones con un tema en común que subraya la importancia de mantener los procesos de paz y la participación cívica. Finalmente, la “*diversidad cultural*” y “*la preservación identitaria*”, muestra que las percepciones sobre estos temas están conectados, destacando la necesidad de preservar conjuntamente la identidad cultural y la biodiversidad.

---

**Figura 44.** Grafica 3D de clustering y etiquetas a respuesta de ¿Qué deberíamos mantener?.



Fuente: Elaboración propia.

---

A través de la Figura 44, se puede observar la visualización 3D de las áreas que los ciudadanos consideran que se deben mantener, para la cual se identifican tres segmentos principales: 1) “conservación ambiental” y “participación ciudadana”, 2) “paz” y “participación ciudadana”, 3) “diversidad cultural” y “preservación identitaria”. Esta representación 3D busca una comprensión más dinámica de cómo las percepciones ciudadanas sobre lo que debe mantenerse están interconectadas y estructuradas, proporcionando una perspectiva de las áreas que los ciudadanos consideran relevantes para mantener en el contexto social y cultural actual.

El análisis de las visualizaciones en 2D y 3D sobre las áreas que los ciudadanos consideran que deben cambiarse, mejorarse o mantenerse revela percepciones y prioridades conectadas. Los segmentos muestran temas como la “*conservación ambiental*”, “*participación ciudadana*”, “*educación*” y la “*preservación cultural*”, son vistos de manera conjunta en algunos casos y de forma diferenciada en otros. Mediante el proceso de clustering, se pudieron identificar y analizar estas relaciones, algo que no habría sido posible sin esta técnica. Estas visualizaciones destacan la necesidad de enfoques integrados y específicos en las políticas públicas. Este análisis proporciona una base para diseñar estrategias que reflejen las complejidades y relaciones de las prioridades ciudadanas.

#### **4.4. ASISTENTE CONVERSACIONAL**

En este trabajo, se desarrolló un asistente conversacional utilizando técnicas de RAG para facilitar la interacción con la base de datos de encuestas de opiniones ciudadanas. El asistente de este trabajo está dirigido a investigadores, analistas de políticas públicas, gestores comunitarios y académicos interesados en comprender mejor las opiniones ciudadanas. Puede ser utilizado por entidades gubernamentales, ONG y centros de investigación para diseñar y evaluar políticas públicas (dada su configuración inicial). La capacidad de configurar el rol del asistente permite su adaptación a diversas tareas y enfoques, como para este trabajo fue el de asumir el rol de *un asesor político para analizar respuestas y proponer políticas públicas*.

Sin embargo, la configuración flexible del asistente, que incluye la capacidad de ajustar su rol según la tarea, es un factor potencial para diversas aplicaciones. Esta adaptación permite a los usuarios personalizar la interacción según sus necesidades específicas, ya sea para el análisis de discurso, la elaboración de políticas, la investigación académica o resúmenes personalizados.

El asistente corre en máquina local a través de la terminal del IDE en el que se desarrolló, aunque también se puede llamar y ejecutar via comand line. Los resultados siguen la siguiente estructura de funcionamiento:

1. Se activa, y le solicita que “*Ingrese la pregunta*”.
2. El usuario realiza una pregunta, por ejemplo: *¿Cómo podemos mejorar el sistema de salud?*

3. El asistente internamente ejecuta el proceso detallado en el apartado 3.4.5 y devuelve la respuesta teniendo en cuenta las intenciones “Cambiar”, “Mejorar” y “Mantener”. La respuesta que devuelve es la creación de una idea de política pública según las opiniones de los ciudadanos en cada una de las intenciones. Por ejemplo, continuando con el ejemplo anterior: *¿Cómo podemos mejorar el sistema de salud?* El asistente respondió:

---

**Figura 45.** Respuesta del asistente conversacional.

Como asesor político, he agrupado las respuestas de los ciudadanos en tres categorías: "Cambiar", "Mejorar" y "Mantener" (no hay respuestas en esta última categoría). A continuación, presento una política pública para cada categoría basada en las opiniones y explicaciones de los ciudadanos.

**Cambiar**  
La política pública para esta categoría se centrará en transformar radicalmente el sistema de salud actual. Debe priorizarse la equidad en la atención médica y garantizar que el sistema de salud no sea negociable. Se debe establecer un sistema de salud que priorice la vida y la atención médica de calidad sobre los intereses económicos. Se debe revisar y reformar la legislación actual para asegurar que el sistema de salud sea más justo y accesible para todos. Se debe garantizar que los recursos se distribuyan de manera equitativa y transparente, eliminando intermediarios que obstaculizan el acceso a la atención médica.

**Mejorar**  
La política pública para esta categoría se enfocará en mejorar la calidad y eficiencia del sistema de salud actual. Debe priorizarse la mejora de la prestación del servicio de salud, garantizando que los recursos lleguen a los prestadores que brindan atención médica de calidad. Se debe trabajar en la capacitación y recursos adecuados para los profesionales de la salud. Se debe también reforzar la estructura del sistema de salud, reduciendo la burocracia y los intermediarios, y garantizando que las IPS estén en buen estado financiero. Se debe priorizar la vida y la atención médica de calidad, especialmente en enfermedades graves, y asegurar que el estado tenga convenios adecuados con farmacéuticas para facilitar el acceso a medicamentos vitales.

Es importante destacar que no hay respuestas en la categoría "Mantener", lo que sugiere que la mayoría de los ciudadanos consideran que el sistema de salud actual necesita algún tipo de cambio o mejora.

Fuente: Elaboración propia<sup>1</sup>

---

4. El asistente vuelve a su estado inicial para que el usuario vuelva a realizar otra pregunta.

Otro ejemplo de pregunta y respuesta con el asistente: *¿Qué se opina sobre la educación?*

---

---

<sup>1</sup> En el Anexo 2, se presenta la salida original del asistente vía terminal.

**Figura 46.** Respuesta del asistente conversacional ante otra pregunta.

A continuación, presento las categorías y políticas públicas correspondientes basadas en las opiniones y explicaciones de los ciudadanos:

**Cambiar**

La política pública para este grupo se enfocaría en revisar y reformar el sistema educativo actual para hacerlo más accesible, inclusivo y enfocado en el desarrollo de habilidades y valores humanos. Se necesitaría un enfoque más holístico en la educación, que vaya más allá de la mera transmisión de conocimientos y se centre en la formación de ciudadanos críticos, reflexivos y comprometidos con la sociedad. Esto podría incluir la revisión de la financiación y el acceso a la educación, así como la promoción de una cultura de respeto y paz en el aula.

**Mejorar**

La política pública para este grupo se enfocaría en fortalecer y mejorar la calidad de la educación, garantizando que todos los estudiantes tengan acceso a una educación de alta calidad y relevancia. Se necesitaría invertir en la formación de docentes, la modernización de la infraestructura educativa y la promoción de programas que fomenten la innovación y el desarrollo de habilidades. También sería importante garantizar que la educación sea accesible y asequible para todos, sin importar su nivel socioeconómico. Esto podría incluir la creación de programas de becas y ayudas financieras para estudiantes de bajos ingresos.

**Mantener**

No hay respuestas que coincidan con esta categoría, ya que la mayoría de los ciudadanos consideran que la educación necesita cambiar o mejorar en algún aspecto. Sin embargo, si hubiera respuestas que sugieran mantener el estado actual de la educación, la política pública para este grupo se enfocaría en fortalecer y consolidar los avances y logros alcanzados en la educación, garantizando que el sistema educativo siga siendo sólido y accesible para todos.

Fuente: Elaboración propia<sup>2</sup>

El desarrollo del asistente conversacional utilizando RAG, permite una interacción flexible y dinámica con la base de datos de opiniones ciudadanas del programa "TQHC". Los resultados muestran la capacidad del asistente para manejar una amplia variedad de preguntas y configuraciones, proporcionando respuestas contextualizadas que reflejan las percepciones y prioridades de la ciudadanía. Esta capacidad de adaptación y personalización del asistente para diferentes roles y tareas facilita la extracción de insights valiosos, permitiendo a los investigadores y formuladores de políticas analizar múltiples dimensiones de las opiniones ciudadanas. El uso de estas técnicas avanzadas para la construcción del asistente ha sido clave para convertir datos no estructurados en información accionable, buscando aplicarla en el análisis de discurso y para este caso potenciando el diseño de estrategias que respondan a las necesidades y expectativas de la población.

#### **4.5. MODELADO DE CLASIFICACIÓN DE TEXTO**

En este apartado se presentan los resultados de los modelos de clasificación utilizando los diferentes enfoques mencionados en durante este trabajo, vector embeddings, prompt engineering, PCA y modelos generativos. Estos resultados se comparan con los presentados por Fajardo Becerra (2023), quien utilizó tres

<sup>2</sup> En el Anexo 3, se presenta la salida original del asistente vía terminal.

modelos clasificadores con técnicas tradicionales de NLP. Los mismos modelos se presentan en este estudio. En general, para cada modelo clasificador se compararon cuatro aproximaciones: una con la configuración de referencia de Fajardo Becerra (2023) y las otras tres con las siguientes configuraciones:

- Aproximación de referencia: Configuración de los datos y cada clasificador según Fajardo Becerra (2023),
- Aproximación 1: **Vector Embeddings sobre configuración de datos de referencia utilizado por Fajardo Becerra, 2023**
- Aproximación 2: **Vector Embeddings + Prompt Engineering.**
- Aproximación 3: **Vector Embeddings + PCA + Prompt Engineering**

Los indicadores verbales utilizados en la configuración de referencia y para este trabajo corresponden a BI (13%), CN (13%), NUL(13%) y VA (22%) ya que corresponden a la mayoría de los registros y fueron los empleados para el probar los modelos de clasificación. A continuación, se presentan los resultados del modelo de regresión logística para cada aproximación:

**Tabla 5** Resultado Regresión Logística con diferentes aproximaciones

| Modelo/Aproximación                             | Clase | Precision | Recall | F1-Score | Accuracy |
|-------------------------------------------------|-------|-----------|--------|----------|----------|
| Regresión Logística - de Fajardo Becerra, 2023. | BI    | 0,32      | 0,41   | 0,36     | 0,51     |
|                                                 | CN    | 0,35      | 0,42   | 0,39     |          |
|                                                 | NUL   | 0,27      | 0,28   | 0,28     |          |
|                                                 | VA    | 0,71      | 0,61   | 0,66     |          |
| Regresión Logística – Aproximación 1            | BI    | 0,38      | 0,18   | 0,24     | 0,58     |
|                                                 | CN    | 0,42      | 0,25   | 0,31     |          |
|                                                 | NUL   | 0,38      | 0,2    | 0,26     |          |
|                                                 | VA    | 0,64      | 0,87   | 0,74     |          |
| Regresión Logística - Aproximación 2            | BI    | 0,48      | 0,35   | 0,41     | 0,57     |
|                                                 | CN    | 0,48      | 0,42   | 0,45     |          |
|                                                 | NUL   | 0,39      | 0,32   | 0,35     |          |
|                                                 | VA    | 0,64      | 0,75   | 0,69     |          |
| Regresión Logística - Aproximación 3            | BI    | 0,42      | 0,3    | 0,35     | 0,57     |
|                                                 | CN    | 0,49      | 0,37   | 0,42     |          |
|                                                 | NUL   | 0,46      | 0,28   | 0,35     |          |
|                                                 | VA    | 0,62      | 0,78   | 0,69     |          |

El modelo Regresión Logística - de Fajardo Becerra (2023) tiene una precisión de 0.51, con buen recall para las clases BI y CN, pero un f1-score bajo para NUL, sugiriendo un equilibrio subóptimo entre precisión y recall. Por su parte, el enfoque

Regresión Logística - Aprox. 1 mejora la precisión a 0.58, con un rendimiento bueno en la clase VA (f1-score de 0.74), aunque con un recall bajo para BI, CN y NUL. Por su parte el modelo Regresión Logística - Aprox. 2., mantiene una precisión de 0.57, mejorando la precisión y el recall para las clases BI y CN, aunque sigue teniendo un rendimiento subóptimo para NUL. Finalmente, el modelo Regresión Logística – Aprox. 3 también tiene una precisión de 0.57, mostrando un buen equilibrio entre precisión y recall, especialmente en la clase VA.

Se encontró que el modelo Regresión Logística – Aprox. 1 fue el de mayor rendimiento en términos de precisión (0.58) y es efectivo en la clase VA, siendo la mejor opción para aplicaciones donde esta clase es crítica. Para un enfoque más equilibrado, los modelos Regresión Logística – Aprox. 2. y Regresión Logística – Aprox. 3 son buenas alternativas, con una precisión de 0.57 y un mejor balance entre precisión y recall en la mayoría de las clases.

Ahora, se presentan los resultados del modelo de random forest para cada aproximación:

**Tabla 6** Resultado Random Forest con diferentes aproximaciones

| <b>Modelo/Aproximación</b>                       | <b>Clase</b> | <b>Precision</b> | <b>Recall</b> | <b>F1-Score</b> | <b>Accuracy</b> |
|--------------------------------------------------|--------------|------------------|---------------|-----------------|-----------------|
| Random Forest- de Fajardo Becerra, D. P. (2023). | BI           | 0,24             | 0,48          | 0,32            | 0,36            |
|                                                  | CN           | 0,29             | 0,38          | 0,33            |                 |
|                                                  | NUL          | 0,21             | 0,4           | 0,27            |                 |
|                                                  | VA           | 0,78             | 0,32          | 0,45            |                 |
| Random Forest - Aproximación 1                   | BI           | 0,4              | 0,01          | 0,02            | 0,56            |
|                                                  | CN           | 0,57             | 0,02          | 0,03            |                 |
|                                                  | NUL          | 0                | 0             | 0               |                 |
|                                                  | VA           | 0,56             | 0,99          | 0,71            |                 |
| Random Forest - Aproximación 2                   | BI           | 0,25             | 0             | 0,01            | 0,54            |
|                                                  | CN           | 0,71             | 0,02          | 0,04            |                 |
|                                                  | NUL          | 0,6              | 0,03          | 0,05            |                 |
|                                                  | VA           | 0,54             | 0,99          | 0,7             |                 |
| Random Forest- Aproximación 3                    | BI           | 0,67             | 0,02          | 0,03            | 0,54            |
|                                                  | CN           | 0                | 0             | 0               |                 |
|                                                  | NUL          | 1                | 0,02          | 0,04            |                 |
|                                                  | VA           | 0,54             | 1             | 0,7             |                 |

El modelo Random Forest - de Fajardo Becerra (2023), tiene una precisión de 0.36, con buen recall para las clases BI, CN y NUL, pero un f1-score bajo para NUL y VA, sugiriendo un bajo equilibrio entre precisión y recall. Por su parte, el enfoque

Random Forest – Aprox.1 mejora la precisión a 0.56, aunque presenta un recall extremadamente bajo para BI, CN y NUL, pero con un alto rendimiento para la clase VA (f1-score de 0.71). Por su parte el modelo Random Forest – Aprox.2 mantiene una precisión similar (0.54), mejorando la precisión para la clase CN, aunque con un bajo rendimiento en las clases BI y NUL. Finalmente, el mejor rendimiento lo presenta el modelo Random Forest – Aprox. 3, con una precisión de 0.54. Este enfoque logra un buen equilibrio entre precisión y recall, especialmente en la clase VA, pero tiene problemas de detección en la clase CN.

Se encontró que el modelo Random Forest - Aprox. 1 fue el de mayor rendimiento en términos de precisión, siendo superior con 0.56, especialmente en la clase VA. Este modelo es la mejor opción para aplicaciones donde la clase VA es crítica. Aunque se debe tener cuidado con las clases donde el recall es bajo.

Ahora, se presentan los resultados del modelo de Naives Bayes para cada aproximación:

**Tabla 7** Resultado Naives Bayes con diferentes aproximaciones

| Modelo/Aproximación                     | Clase | Precision | Recall | F1-Score | Accuracy |
|-----------------------------------------|-------|-----------|--------|----------|----------|
| Naives Bayes - de Fajardo Becerra 2023. | BI    | 0,25      | 0,55   | 0,34     | 0,39     |
|                                         | CN    | 0,29      | 0,55   | 0,38     |          |
|                                         | NUL   | 0,22      | 0,18   | 0,2      |          |
|                                         | VA    | 0,76      | 0,35   | 0,48     |          |
| Naives Bayes - Aproximación 1           | BI    | 0,26      | 0,43   | 0,32     | 0,41     |
|                                         | CN    | 0,24      | 0,42   | 0,3      |          |
|                                         | NUL   | 0,29      | 0,3    | 0,29     |          |
|                                         | VA    | 0,74      | 0,43   | 0,54     |          |
| Naives Bayes - Aproximación 2           | BI    | 0,26      | 0,4    | 0,31     | 0,39     |
|                                         | CN    | 0,36      | 0,45   | 0,4      |          |
|                                         | NUL   | 0,24      | 0,51   | 0,33     |          |
|                                         | VA    | 0,74      | 0,33   | 0,46     |          |
| Naives Bayes - Aproximación 3           | BI    | 0,27      | 0,11   | 0,16     | 0,48     |
|                                         | CN    | 0,28      | 0,46   | 0,35     |          |
|                                         | NUL   | 0,4       | 0,15   | 0,22     |          |
|                                         | VA    | 0,58      | 0,69   | 0,63     |          |

El modelo Naives Bayes - de Fajardo Becerra 2023 tiene una precisión de 0.39, con buen recall para las clases BI y CN, pero un f1-score bajo para NUL, sugiriendo un bajo equilibrio entre precisión y recall. Por otro lado, el enfoque Naives Bayes – Aprox. 1 mejora un poco la precisión a 0.41, con una mejor precisión y recall general, y un f1-score mejorado para la clase VA. El modelo Naives Bayes – Aprox. 2.

mantiene una precisión similar 0.39, pero mejora el recall para NUL y el f1-score para CN, indicando una mejor detección de casos positivos y negativos en estas clases. Finalmente, el mejor rendimiento lo presenta el modelo Naives Bayes – Aprox. 3, con una precisión de 0.48. Este enfoque logra un mejor equilibrio entre precisión y recall, especialmente en la clase VA.

Se encontró que el modelo Naives Bayes – Aprox. 3 fue el de mayor rendimiento, siendo superior en precisión y equilibrio de métricas. Este modelo mejora la detección en clases críticas como VA, siendo la mejor opción para aplicaciones donde esta clase es importante. Por otro lado, si se busca un enfoque más sencillo sin PCA, el modelo Naives Bayes – Aprox. 1 es una buena alternativa, mejorando el trabajo inicial.

## 5 CONCLUSIONES Y TRABAJOS FUTUROS

### 5.1 CONCLUSIONES

En este trabajo se ha explorado y aplicado técnicas avanzadas LLMs para apoyar diferentes tareas de análisis de discurso. Este trabajo incluyó la implementación de modelos para el análisis de sentimientos y de tópicos, así como técnicas de clustering para agrupar temas y visualizar relaciones. También se desarrolló un asistente conversacional utilizando técnicas de recuperación y generación de respuestas con técnicas RAG y se evaluaron modelos de clasificación de texto con aproximación de embeddings para la predicción de etiquetas de datos. Dentro de la exploración de modelos aplicadas al análisis de discurso, se puede concluir que:

1. El modelado de tópicos mediante técnicas de prompt engineering y LLMs en el programa "TQHC" ha revelado diversas percepciones y prioridades entre los participantes, diferenciadas por tema de las respuestas y variables sociodemográficas. Los datos indican una frecuencia elevada de menciones a temas como Educación y Corrupción, con una clara intención de cambio en estos ámbitos. Además, se identifican tendencias de sentimientos positivos en temas como Sistema educativo y Participación ciudadana en zonas urbanas, mientras que temas como Violencia y Corrupción generan sentimientos negativos más intensos en las zonas rurales y entre las mujeres. Este análisis proporciona una comprensión de las dinámicas y preocupaciones presentes en el discurso, destacando las diferencias en las percepciones según el contexto demográfico y geográfico de los participantes.
2. El análisis de sentimientos aplicado a las respuestas del proyecto "TQHC" permite observar una comprensión de las emociones en las respuestas de los ciudadanos. Utilizando técnicas avanzadas de NLP y LLMs, se cuantificaron los sentimientos, lo que permitió identificar patrones emocionales y diferencias contextuales en función de variables sociodemográficas como género, edad, grupo étnico y región. Los resultados indican una tendencia general hacia la insatisfacción, especialmente en la categoría "*cambiar*", donde se observaron sentimientos negativos, sugiriendo una percepción de necesidad de transformación en aspectos problemáticos. Por otro lado, la categoría "*mantener*" muestra una mayor satisfacción, con sentimientos positivos, indicando estabilidad y aprobación en ciertos temas. Este análisis no sólo muestra la importancia de considerar las variables demográficas para una interpretación más completa y contextualizada, sino que también destaca la aplicabilidad y buen funcionamiento de los modelos de LLMs aplicados a través de análisis de sentimientos en el análisis de discurso.

3. El uso de técnicas de agrupamiento o clustering y visualización en 2D y 3D ha permitido observar las percepciones interconectadas de los ciudadanos respecto a qué cambiar, mejorar o mantener. Estas visualizaciones han mostrado cómo temas como la conservación ambiental, la participación ciudadana, la educación y la preservación cultural pueden ser percibidos de manera conjunta o diferenciada. El proceso de clustering ha sido relevante para identificar y analizar estas relaciones, proporcionando una comprensión de las relaciones y diversidades en las opiniones ciudadanas. Este enfoque facilita la interpretación de las complejidades y matices de las prioridades ciudadanas en el análisis de discurso.
4. La implementación del asistente conversacional utilizando técnicas de RAG en el programa "TQHC" ha demostrado su capacidad para manejar diversas preguntas y configuraciones, generando respuestas contextualizadas que reflejan las percepciones y prioridades de los ciudadanos. La adaptabilidad del asistente a diferentes roles y tareas ha permitido extraer insights detallados, proporcionando a investigadores e interesados una herramienta para explorar y comprender las múltiples dimensiones de las opiniones ciudadanas. Los resultados obtenidos evidencian que el asistente permite interactuar con los datos a través de lenguaje natural, facilitando una comprensión de las necesidades y expectativas expresadas por los ciudadanos en la encuesta.
5. En cuanto a los modelos de clasificación para los indicadores verbales, las comparaciones entre los diferentes modelos y aproximaciones muestran variaciones en la precisión, el recall y el f1-score, destacando las fortalezas y debilidades de cada enfoque. Los resultados demuestran que la incorporación de técnicas avanzadas como vector embedding, prompt engineering y PCA puede mejorar el rendimiento de los modelos, proporcionando una mayor capacidad para capturar y clasificar las percepciones ciudadanas de manera más precisa. Esta capacidad de ajuste y mejora en los modelos de clasificación muestran la importancia de seguir explorando estas técnicas para obtener una comprensión más detallada de las opiniones y prioridades de los ciudadanos en las diferentes tareas de clasificación que se tengan o incluso en análisis de discurso.
6. La exploración y aplicación de técnicas avanzadas de modelos grandes de lenguaje en el análisis de discurso ha demostrado su capacidad para extraer conocimiento detallados de los datos de texto, permitiendo a los interesados tener una mejor comprensión de las percepciones y emociones de los participantes. Además, la implementación de modelos avanzados y técnicas de NLP ha resultado ser una herramienta valiosa no solo para la clasificación y análisis de temas y sentimientos, sino también para la visualización y comprensión de las relaciones entre diversas percepciones y prioridades

ciudadanas. La capacidad de interactuar en lenguaje natural con la base de datos mediante técnicas de RAG añade un valor adicional, permitiendo consultas personalizadas y mejorando la accesibilidad a los datos. Así, el trabajo realizado en esta tesis resalta el potencial de estas tecnologías para apoyar diversas tareas de análisis de discurso y clasificación.

## **5.2. TRABAJOS FUTUROS**

A continuación, se presentan varias líneas de trabajo a ser contempladas en trabajos futuros:

1. Desarrollar técnicas de personalización para asistentes conversacionales que ajusten las respuestas a las características individuales del usuario, utilizando datos sociodemográficos para ofrecer respuestas más relevantes.
2. Ampliar el análisis de sentimientos y tópicos incorporando datos multimodales, como texto, audio y video, para obtener una comprensión más completa de las percepciones y emociones a través de diferentes formatos de encuestas.
3. Extender el uso de técnicas de clustering y visualización a otros contextos, como salud pública, educación o investigación de mercados, para validar y generalizar los enfoques desarrollados.
4. Optimizar técnicas de prompt engineering para mejorar la efectividad de los LLMs en tareas específicas de análisis de discurso, experimentando con diferentes estructuras de prompts.

## 6 REFERENCIAS

- Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., & McGrew, B. (2023). GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Arseniev-Koehler, A. (2021). Theoretical foundations and limits of word embeddings: What types of meaning can they capture? *Sociological Methods & Research*, 0049-1241.
- Asudani, D. S., Nagwani, N. K., & Singh, P. (2023). Impact of word embedding models on text analytics in deep learning environment: A review. *Artificial Intelligence Review*, 56(9), 10345-10425.
- Babbie, E. (2013). *The practice of social research*. Cengage Learning.
- Baldassarre, M. T., Caivano, D., Fernandez Nieto, B., Gigante, D., & Ragone, A. (2023). The social impact of generative AI: An analysis on ChatGPT. In *Proceedings of the 2023 ACM Conference on Information Technology for Social Good* (pp. 363-373).
- Baroni, M., Dinu, G., & Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 238-247).
- Barua, S. (2024). Exploring autonomous agents through the lens of large language models: A review. *arXiv preprint arXiv:2404.04442*.
- Benara, V., Singh, C., Morris, J. X., Antonello, R., Stoica, I., Huth, A. G., & Gao, J. (2024). Crafting interpretable embeddings by asking LLMs questions. *arXiv preprint arXiv:2405.16714*.
- Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8), 1798-1828.
- Bhardwaj, A., Khanna, P., & Kumar, S. (2020). Generative model for NLP applications based on component extraction. *Procedia Computer Science*, 167, 918-931.
- Boleda, G. (2020). Distributional semantics and linguistic theory. *Annual Review of Linguistics*, 6(1), 213-234.
- Brookshear, J. G. (1993). *Teoría de la computación: lenguajes formales, autómatas y complejidad*. Boston: Addison Wesley.

- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877-1901.
- Crossley, S. A., Allen, L. K., Kyle, K., & McNamara, D. S. (2014). Analyzing discourse processing using a simple natural language processing tool. *Discourse Processes*, 51(5-6), 511-534.
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877-1901.
- Dong, L., Yang, N., Wang, W., Wei, F., Liu, X., & Zhou, M. (2019). Unified language model pre-training for natural language understanding and generation. *arXiv preprint arXiv:1905.03197*.
- Drikvandi, R., & Lawal, O. (2023). Sparse principal component analysis for natural language processing. *Annals of Data Science*, 10(1), 25-41.
- Fajardo Becerra, D. P. (2023). Definición de una metodología para análisis de discurso basado en lingüística computacional y técnicas de aprendizaje de máquina [Tesis de Maestría, Universidad EAFIT]. Repositorio Institucional Universidad EAFIT. <http://hdl.handle.net/10784/32412>
- Franceschelli, G., & Musolesi, M. (2024). Reinforcement learning for generative AI: State of the art, opportunities and open research challenges. *Journal of Artificial Intelligence Research*, 79, 417-446.
- Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., & Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.
- Ghazizadeh, E., & Zhu, P. (2020). A systematic literature review of natural language processing: Current state, challenges and risks. In *Proceedings of the Future Technologies Conference* (pp. 634-647). Cham: Springer International Publishing.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., & Bengio, Y. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27.
- Gozalo-Brizuela, R., & Garrido-Merchán, E. C. (2023). A survey of generative AI applications. *arXiv preprint arXiv:2306.02781*.
- Gruetzemacher, R. (2022). Overview of natural language processing. *Harvard Business School*.

- He, Y., Huang, F., Jiang, X., Nie, Y., Wang, M., Wang, J., & Chen, H. (2024). Foundation model for advancing healthcare: Challenges, opportunities, and future directions. *arXiv preprint arXiv:2404.03264*.
- Huber, P., & Carenini, G. (2022). Towards understanding large-scale discourse structures in pre-trained and fine-tuned language models. *arXiv preprint arXiv:2204.04289*.
- Johnson, S. J., Murty, M. R., & Navakanth, I. (2024). A detailed review on word embedding techniques with emphasis on word2vec. *Multimedia Tools and Applications*, 83(13), 37979-38007.
- Joty, S., Carenini, G., Ng, R., & Murray, G. (2019). Discourse analysis and its applications. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*.
- Jurafsky, D., & Martin, J. H. (2021). *Speech and language processing*. Pearson.
- Kamkarhaghighi, M., & Makrehchi, M. (2017). Content tree word embedding for document representation. *Expert Systems with Applications*, 90, 241-249.
- Khandale, K. B., & Mahender, C. N. (2020). Natural language processing based rule based discourse analysis of Marathi text. In *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)* (pp. 356-362). IEEE.
- Khurana, D., Koli, A., Khatter, K., & Singh, S. (2023). Natural language processing: State of the art, current trends and challenges. *Multimedia Tools and Applications*, 82(3), 3713-3744.
- Kingma, D. P., & Welling, M. (2019). An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4), 307-392.
- Knoth, N., Tolzin, A., Janson, A., & Leimeister, J. M. (2024). AI literacy and its implications for prompt engineering strategies. *Computers and Education: Artificial Intelligence*, 6, 100225.
- Korzynski, P., Mazurek, G., Krzyrkowska, P., & Kurasinski, A. (2023). Artificial intelligence prompt engineering as a new digital competence: Analysis of generative AI technologies such as ChatGPT. *Entrepreneurial Business and Economics Review*, 11(3), 25-37.
- Kvale, S. (2007). *Doing interviews*. SAGE Publications.
- Lenci, A. (2018). Distributional models of word meaning. *Annual Review of Linguistics*, 4(1), 151-171.

- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33, 9459-9474.
- Manakul, P., Liusie, A., & Gales, M. J. (2023). SelfcheckGPT: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*.
- Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT Press.
- Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press.
- Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4), 1093-1113.
- Megahed, F. M., Chen, Y. J., Zwetsloot, I., Knoth, S., Montgomery, D. C., & Jones-Farmer, L. A. (2023). AI and the future of work in statistical quality control: Insights from a first attempt to augmenting ChatGPT with an SQC knowledge base (ChatSQC). *arXiv preprint arXiv:2308.13550*.
- Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., & Gao, J. (2024). Large language models: A survey. *arXiv preprint arXiv:2402.06196*.
- Miner, G., Elder IV, J., Fast, A., Hill, T., & Nisbet, R. (2012). *Practical text mining and statistical analysis for non-structured text data applications*. Academic Press.
- Mochtak, M. (2024). *War narratives in post-conflict societies: Keeping the past alive in the former Yugoslavia*. Taylor & Francis.
- Naseem, U., Razzak, I., Khan, S. K., & Prasad, M. (2021). A comprehensive survey on word representation models: From classical to state-of-the-art word representation language models. *Transactions on Asian and Low-Resource Language Information Processing*, 20(5), 1-35.
- Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., & Mian, A. (2023). A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*.
- Öhman, E., & Metcalfe, A. G. (2021). Japanese beauty marketing on social media: Critical discourse analysis meets NLP. In *Proceedings of the Workshop on Natural Language Processing for Digital Humanities* (pp. 131-137).
- O'Neill, L., Anantharama, N., Buntine, W., & Angus, S. D. (2021). Quantitative

- discourse analysis at scale—AI, NLP and the transformer revolution. *Monash University, SoDa Laboratories*, 2021-12.
- Ororbia, A., & Kifer, D. (2022). The neural coding framework for learning generative models. *Nature Communications*, 13(1), 2064.
- Rush, A. M., Chopra, S., & Weston, J. (2015). A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 379-389).
- Sajjad, H., Alkhouli, T., Vogel, S., & Ney, H. (2013). Statistical models for unsupervised, semi-supervised, and supervised machine translation. *Computational Linguistics*, 39(2), 203-236.
- Sharma, P., Kumar, M., Sharma, H. K., & Biju, S. M. (2024). Generative adversarial networks (GANs): Introduction, taxonomy, variants, limitations, and applications. *Multimedia Tools and Applications*, 1-48.
- Shinn, N., Cassano, F., Labash, B., Gopinath, A., Narasimhan, K., & Yao, S. (2023). Reflexion: Language agents with verbal reinforcement learning. *arXiv preprint cs.AI/2303.11366*.
- Sun, Y., & Ortiz, J. (2024). Rapid review of generative AI in smart medical applications. *arXiv preprint arXiv:2406.06627*.
- Taha, K., Yoo, P. D., Yeun, C., & Taha, A. (2024). Text classification: A review, empirical, and experimental evaluation. *arXiv preprint arXiv:2401.12982*.
- Vardanyan, E., Hunanyan, S., Galstyan, T., Minasyan, A., & Dalalyan, A. S. (2024). Statistically optimal generative modeling with maximum deviation from the empirical distribution. In *Forty-first International Conference on Machine Learning*.
- Varma, K. K. (2019). Deconstructing word embedding models. *arXiv preprint arXiv:2406.06627*.
- Vaswani, A., Saber, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*.
- Vicente, M., Barros, C., Peregrino, F. S., Agulló, F., & Lloret, E. (2015). La generación de lenguaje natural: Análisis del estado actual. *Computación y Sistemas*, 19(4), 721-756.
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., & Zhou, D. (2022).

- Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35, 24824-24837.
- Wu, T., Jiang, E., Donsbach, A., Gray, J., Molina, A., Terry, M., & Cai, C. J. (2022). Promptchainer: Chaining large language model prompts through visual programming. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts* (pp. 1-10).
- Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T., Cao, Y., & Narasimhan, K. (2024). Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36.
- Zhang, S. J., Florin, S., Lee, A. N., Niknafs, E., Marginean, A., Wang, A., & Drori, I. (2023). Exploring the MIT mathematics and EECS curriculum using large language models. *arXiv preprint arXiv:2306.08997*.
- Zhang, B., Chang, K., & Li, C. (2024). Simple techniques for enhancing sentence embeddings in generative language models. *arXiv preprint arXiv:2404.03921*.
- Zhou, Y., Muresanu, A. I., Han, Z., Paster, K., Pitis, S., Chan, H., & Ba, J. (2022). Large language models are human-level prompt engineers. *arXiv preprint arXiv:2211.01910*.
- Zou, H., Hastie, T., & Tibshirani, R. (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15(2), 265-286.

## ANEXOS

**Anexo 1.** Hiperparámetros de los modelos de Clasificador de Texto para el Indicador Verbal del Ciclo 1

| <b>Modelo</b>       | <b>Configuración de referencia hiperparámetros</b>      | <b>Configuración optimizada hiperparámetros</b>                                                                                        |
|---------------------|---------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------|
| Regresión logística | Multi_class = 'multinomial',<br>Solver = 'lbfgs', C = 5 | Solver = 'saga', C = 5                                                                                                                 |
| Árbol de decisión   | Criterion = 'entropy'                                   | min_samples_split: 2,<br>min_samples_leaf: 10,<br>max_features: None,<br>max_depth: 10, criterion:<br>'entropy'                        |
| Random Forest       | N_estimators = 10                                       | n_estimators: 50,<br>min_samples_split: 10,<br>min_samples_leaf: 1,<br>max_features: 'sqrt',<br>max_depth: 30, criterion:<br>'entropy' |
| KNN                 | N_neighbors = 5, Weights<br>= 'uniform'                 | weights: 'distance', p: 1,<br>n_neighbors: 15,<br>leaf_size: 50, algorithm:<br>'ball_tree'                                             |
| MLP                 | Alpha = 1.5, Max_iter =<br>500                          | solver: 'adam', max_iter:<br>500, learning_rate:<br>'constant',<br>hidden_layer_sizes:<br>(100,), alpha: 1.0,<br>activation: 'tanh'    |

## Anexo 2. Salida del asistente conversacional de la pregunta: *¿Cómo podemos mejorar el sistema de salud?*

```
C:\Users\eljimenez\Desktop\Personal\1. Tesis de Maestria\Tesis 2024\Modelacion\Asistente>python -m main
Ingrese la pregunta:¿Como podemos mejorar el sistema de salud?
Como asesor político, he agrupado las respuestas de los ciudadanos en tres categorías: "Cambiar", "Mejorar" y "Mantener"
(no hay respuestas en esta última categoría). A continuación, presento una política pública para cada categoría basada
en las opiniones y explicaciones de los ciudadanos.

**Cambiar**

La política pública para esta categoría se centrará en transformar radicalmente el sistema de salud actual. Debe priorizarse la equidad en la atención médica y garantizar que el sistema de salud no sea negociable. Se debe establecer un sistema de salud que priorice la vida y la atención médica de calidad sobre los intereses económicos. Se debe revisar y reformar la legislación actual para asegurar que el sistema de salud sea más justo y accesible para todos. Se debe garantizar que los recursos se distribuyan de manera equitativa y transparente, eliminando intermediarios que obstaculizan el acceso a la atención médica.

**Mejorar**

La política pública para esta categoría se enfocará en mejorar la calidad y eficiencia del sistema de salud actual. Debe priorizarse la mejora de la prestación del servicio de salud, garantizando que los recursos lleguen a los prestadores que brindan atención médica de calidad. Se debe trabajar en la capacitación y recursos adecuados para los profesionales de la salud. Se debe también reforzar la estructura del sistema de salud, reduciendo la burocracia y los intermediarios, y garantizando que las IPS estén en buen estado financiero. Se debe priorizar la vida y la atención médica de calidad, especialmente en enfermedades graves, y asegurar que el estado tenga convenios adecuados con farmacéuticas para facilitar el acceso a medicamentos vitales.

Es importante destacar que no hay respuestas en la categoría "Mantener", lo que sugiere que la mayoría de los ciudadanos
```

## Anexo 3. Salida del asistente conversacional de la pregunta: *¿Qué se opina sobre la educación?*

```
C:\Users\eljimenez\Desktop\Personal\1. Tesis de Maestria\Tesis 2024\Modelacion\Asistente>python -m main
Ingrese la pregunta: Que se opina sobre la educacion
A continuación, presento las categorías y políticas públicas correspondientes basadas en las opiniones y explicaciones de los ciudadanos:

**Cambiar**

La política pública para este grupo se enfocaría en revisar y reformar el sistema educativo actual para hacerlo más accesible, inclusivo y enfocado en el desarrollo de habilidades y valores humanos. Se necesitaría un enfoque más holístico en la educación, que vaya más allá de la mera transmisión de conocimientos y se centre en la formación de ciudadanos críticos, reflexivos y comprometidos con la sociedad. Esto podría incluir la revisión de la financiación y el acceso a la educación, así como la promoción de una cultura de respeto y paz en el aula.

**Mejorar**

La política pública para este grupo se enfocaría en fortalecer y mejorar la calidad de la educación, garantizando que todos los estudiantes tengan acceso a una educación de alta calidad y relevancia. Se necesitaría invertir en la formación de docentes, la modernización de la infraestructura educativa y la promoción de programas que fomenten la innovación y el desarrollo de habilidades. También sería importante garantizar que la educación sea accesible y asequible para todos, sin importar su nivel socioeconómico. Esto podría incluir la creación de programas de becas y ayudas financieras para estudiantes de bajos ingresos.

**Mantener**

No hay respuestas que coincidan con esta categoría, ya que la mayoría de los ciudadanos consideran que la educación necesita cambiar o mejorar en algún aspecto. Sin embargo, si hubiera respuestas que sugieran mantener el estado actual de la educación, la política pública para este grupo se enfocaría en fortalecer y consolidar los avances y logros alcanzados en la educación, garantizando que el sistema educativo siga siendo sólido y accesible para todos.
```