ANÁLISIS COMPARATIVO DE LA PERCEPCIÓN MEDIÁTICA DE LA REFORMA A LA SALUD EN COLOMBIA USANDO TÉCNICAS NLP

Bryan Steven Hernández Moreno
Samuel Andrés Coronado Cobos
José Luis González Ipuz

Proyecto Aplicado para optar al título de
Magister en Ciencia de Datos

Director
Abel Álvarez Bustos

Codirector
Carlos Ernesto Ramírez Ovalle

FACULTAD DE INGENIERÍA Y CIENCIAS
MAESTRÍA EN CIENCIA DE DATOS
SANTIAGO DE CALI, MAYO 29 DE 2025

FICHA RESUMEN

TÍTULO DEL PROYECTO: ANÁLISIS COMPARATIVO DE LA PERCEPCIÓN MEDIÁTICA DE LA REFORMA A LA SALUD EN COLOMBIA USANDO TÉCNICAS NLP

1. ÁREA DE TRABAJO: Ciencia De Datos
2. TIPO DE PROYECTO: Aplicado
3. ESTUDIANTE(S): Bryan Steven Hernández, Samuel Coronado, y José González
4. CORREO ELECTRÓNICO: bryanshm@javerianacali.edu.co, scoronadoc@javerianacali.edu.co, gjoselo360@javerianacali.edu.co
5. DIRECCIÓN Y TELEFONO: Carrera 15 # 27-33, Celular:3126222332
6. DIRECTOR: Abel Álvarez Bustos
7. VINCULACIÓN DEL DIRECTOR: Profesor de planta de la Facultad de Ingeniería y Ciencias
8. CORREO ELECTRÓNICO DEL DIRECTOR: abel.alvarez@javeriancali.edu.co
9. CO-DIRECTOR: Carlos Ernesto Ramírez Ovalle
10. GRUPO O EMPRESA QUE LO AVALA: Ninguna
11. OTROS GRUPOS O EMPRESAS: Ninguna
12. PALABRAS CLAVE: NLP, Análisis de sentimiento, Machine Learning, Salud, Reforma.
13. FECHA DE INICIO: 26 de junio del 2024
14. DURACIÓN ESTIMADA (En meses): 8
15. RESUMEN:

Este estudio aplicó técnicas de ciencia de datos y procesamiento de lenguaje natural (NLP) para analizar la percepción mediática sobre la reforma a la salud en Colombia (2022-2024), abordando una brecha en la literatura al examinar diferencias regionales en la cobertura periodística. Partiendo del rol del periodismo en la formación de opinión pública especialmente en temas críticos como la salud, se recolectaron 1.401 noticias mediante web scraping de fuentes confiables (SCImago) en las regiones Andina, Caribe y Pacífica, siguiendo criterios de inclusión rigurosos (periodo 2022-2024, idioma español, relevancia temática). Los datos se preprocesaron con técnicas de NLP (tokenización, lematización, eliminación de stopwords y publicidad) y se depuraron mediante análisis estadístico (excluyendo 39 noticias atípicas por IQR). Para el análisis, se implementaron modelos de similitud (TF-IDF, Doc2Vec, MPNet) y clasificación de sentimientos (BETO, RoBERTa y ChatGPT-4o), este último como contraste. Los modelos fine-tuned (BETO: 91.29% accuracy; RoBERTa: 89.18%) superaron significativamente a ChatGPT-4o (67.29%), demostrando la importancia del ajuste especializado para contextos periodísticos en español. El etiquetado manual (26.43% del corpus) permitió validar los resultados, destacando tendencias regionales: neutralidad en la cobertura Andina (asociada a enfoques institucionales), mayor positividad en el Caribe y predominio de narrativas

negativas en el Pacífico (vinculadas a críticas locales). Los hallazgos confirman que: Las diferencias geopolíticas y socioculturales moldean narrativas mediáticas, pese a cierta homogeneidad discursiva intrarregional (validada por métricas de similitud). El fine-tuning de modelos de NLP es crucial para análisis de sentimientos en dominios especializados, siendo BETO óptimo para español. La metodología propuesta integrando web scraping, NLP y visualización interactiva (Power BI) ofrece un marco replicable para estudios de percepción mediática en políticas públicas.

TABLA DE CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

1. INTRODUCCIÓN

El periodismo desempeña un papel fundamental en la formación de la opinión pública a nivel global, ya que proporciona información veraz y oportuna sobre los determinantes sociales [1]. Esta función es esencial para la construcción de una sociedad democrática que protege los derechos y promueve la salud, la cultura, la educación y el desarrollo. La manera en que se realiza la cobertura periodística, especialmente en temas de interés público, influye significativamente en la percepción y comprensión de la información difundida a través de los diversos medios, siendo de particular importancia en lo que respecta a la salud de la comunidad.

En Colombia, durante el año 2022, la introducción de una propuesta de reforma sanitaria por parte del gobierno nacional generó un intenso debate en las plataformas mediáticas; esto permitió que los diversos medios periodísticos a nivel nacional manifestaran su percepción sobre la reforma, hasta el momento, no se ha encontrado en la literatura ningún análisis que permita comprender cómo los medios de comunicación de distintas regiones del país abordan este tema.

Ante este escenario, la ciencia de datos se erige como una solución innovadora para conocer y comprender la percepción de los medios periodísticos respecto a la reforma. Para abordar sistemáticamente la percepción mediática, se implementaron técnicas avanzadas de procesamiento del lenguaje natural (NLP).

Este enfoque incluyó la recolección automatizada de noticias relevantes mediante web scraping en fuentes confiables de las regiones Andina, Caribe y Pacífico. Los textos recolectados se sometieron a un preprocesamiento que contempló la eliminación de stopwords, tokenización y lematización, con el fin de garantizar la coherencia y calidad de los datos. Posteriormente, se llevó a cabo un análisis estadístico para entender la estructura y composición de la base de datos. Se seleccionaron e implementaron modelos de similitud de documentos como TF-IDF, Doc2Vec y MPNet, así como modelos de análisis de sentimientos basados en BETO, RoBERTa y ChatGPT-4o. La evaluación de estos modelos se efectuó utilizando métricas de distancia o similitud, como la similitud del coseno y la distancia euclidiana. Adicionalmente, se realizó un proceso de etiquetado manual de las noticias para generar un conjunto de referencia que permitiera entrenar y validar los modelos supervisados, lo cual facilitó el uso de métricas de rendimiento como precisión, recall y F1-score en la evaluación del desempeño. Estos resultados permitieron visualizar patrones y diferencias en la percepción de la reforma sanitaria en Colombia.

2. CONTEXTUALIZACIÓN DEL PROYECTO

2.1 PLANTEAMIENTO DEL PROBLEMA

En el año 2022, el gobierno colombiano presentó una ambiciosa propuesta de reforma al sistema nacional de salud. Esta iniciativa ha suscitado un amplio debate tanto en la esfera política como en los medios de comunicación, dada su intención de transformar profundamente aspectos sensibles del modelo de aseguramiento, financiación, atención primaria, y organización institucional del sector salud. Si bien esta investigación no se enfoca en el análisis técnico de dicha reforma, resulta imprescindible contextualizar el porqué de su carácter controversial y el rol que desempeñan los medios de comunicación en su recepción y representación pública [2], [3].

La reforma se desarrolla en un entorno marcado por profundas desigualdades regionales históricas. Colombia presenta disparidades significativas en términos de acceso a servicios de salud, calidad de atención, infraestructura, presencia del Estado y capital humano, especialmente entre regiones como la Andina, Caribe y Pacífica [3], [4]. Tales condiciones estructurales influyen directamente en la forma en que las comunidades perciben los cambios propuestos, y en cómo estos son reflejados y amplificados por los medios de comunicación regionales.

Los medios, lejos de ser actores neutrales, constituyen dispositivos activos de construcción simbólica y discursiva. A través de sus narrativas, contribuyen a moldear la percepción pública de las políticas estatales, reforzando o disputando su legitimidad. Esta capacidad de influencia está mediada por factores ideológicos, intereses editoriales, estructuras económicas y marcos culturales que varían entre regiones del país [5], [6]. En consecuencia, es razonable suponer que la cobertura mediática sobre la reforma no sea homogénea, sino que presente diferencias sustantivas en cuanto al tono, el enfoque temático y la valoración implícita o explícita del contenido.

A pesar de la relevancia de esta problemática, se identifica una notoria ausencia de estudios sistemáticos y comparativos que analicen, desde una perspectiva computacional y reproducible, cómo los medios de diferentes regiones del país han abordado esta reforma. En particular, no se han aplicado metodologías basadas en técnicas de procesamiento de lenguaje natural (NLP, por sus siglas en inglés) que permitan modelar de forma automatizada el sentimiento, la opinión y la similitud entre los discursos mediáticos regionales [7], [8]. Esta brecha académica limita la capacidad de comprender cómo las narrativas mediáticas influyen en la aceptación o rechazo de políticas públicas, así como su potencial efecto en la polarización social y en la percepción pública de legitimidad gubernamental.

Desde el enfoque de la ciencia de datos aplicada a fenómenos sociales, este vacío constituye una oportunidad para implementar modelos computacionales que permitan extraer y analizar patrones discursivos con rigurosidad metodológica y alcance cuantitativo. La aplicación de

herramientas de NLP a textos periodísticos posibilita la identificación de estructuras latentes de opinión, la comparación de enfoques narrativos entre regiones y la visualización de convergencias o divergencias discursivas, contribuyendo así a una comprensión más integral del debate público.

En suma, la inexistencia de análisis comparativos automatizados sobre la percepción mediática regional en torno a la reforma a la salud representa un vacío significativo tanto en el ámbito académico como en el desarrollo de políticas públicas basadas en evidencia. Esta investigación, al aplicar técnicas avanzadas de NLP a contenidos periodísticos provenientes de las principales regiones del país, busca aportar una mirada novedosa y fundamentada al entendimiento de las dinámicas discursivas que atraviesan este proceso reformista de alto impacto.

2.2 FORMULACIÓN DEL PROBLEMA

Para abordar la problemática planteada, es fundamental responder a un interrogante clave: ¿Cómo puede el uso de técnicas avanzadas de ciencia de datos y procesamiento de lenguaje natural (NLP) y análisis de sentimientos mejorar la comprensión de la percepción mediática sobre la reforma a la salud en Colombia, e identificar si existe similitud o disparidad en la cobertura mediática?

2.3 SISTEMATIZACIÓN

Adicionalmente, se complementa esta investigación con las siguientes preguntas: ¿Cuáles son las fuentes de datos mediáticos más relevantes y confiables para recolectar información con web scraping, sobre la reforma a la salud en cada región? ¿Cuáles son las técnicas más eficaces para limpiar y preprocesar datos textuales provenientes de medios de comunicación? ¿Cómo se pueden evaluar los resultados de los modelos de análisis de sentimiento, modelado de temas y modelos de similitud, aplicados a la cobertura mediática sobre la reforma?

3. OBJETIVOS

3.1 OBJETIVO GENERAL

Aplicar técnicas de ciencia de datos y procesamiento de lenguaje natural (NLP) para analizar sistemáticamente la percepción mediática sobre la reforma a la salud en Colombia, con el fin de mejorar la comprensión pública, la similitud o disparidad de percepciones.

3.2 OBJETIVOS ESPECÍFICOS

- Recolectar datos de medios periodísticos sobre la reforma a la salud de las tres principales regiones de Colombia (Andina, Caribe y Pacífico) mediante técnicas de web scraping, asegurando una muestra estadísticamente significativa, y preprocesar estos datos textuales a través de etiquetado, tokenización, lematización y eliminación de ruido para garantizar la calidad y coherencia del análisis.
- Aplicar modelos de procesamiento del lenguaje natural que categorice el sentimiento y la opinión en los textos periodísticos sobre la reforma a la salud en Colombia y aplicar modelos de similitud de documentos, para comparar y visualizar la similaridad y disparidad de las noticias entre las diferentes regiones del país.
- Evaluar los modelos desarrollados utilizando el conjunto de datos recopilados, ajustando los parámetros para mejorar la precisión.
- Comparar las percepciones sobre la reforma a la salud entre las tres regiones principales de Colombia (Andina, Caribe y Pacífico) utilizando un criterio formal de similaridad, identificando convergencias y divergencias en las opiniones expresadas por los diferentes medios de comunicación y mostrar los resultados a través de técnicas de visualización.

4. MARCO DE REFERENCIA

4.1 MARCO TEÓRICO

A continuación, se presentará una breve descripción de los temas clave que se abordarán en esta investigación. Se proporcionará una definición precisa de cada término, asegurando una connotación apropiada enmarcada en los objetivos de este proyecto. Esta sección del marco teórico busca establecer una base conceptual clara y coherente que respalde el desarrollo de los modelos de ciencia de datos utilizando técnicas de procesamiento del lenguaje natural (NLP) para identificar si existe una sectorización basada en la percepción de la reforma de salud.

El marco teórico está estructurado en cuatro secciones que se desarrollan de manera progresiva. El primero presenta los principios generales del NLP y las etapas de preprocesamiento. El segundo aborda las técnicas de vectorización, los modelos similitud y análisis de texto (BETO, RoBERTa y Sentence-Transformers) empleados para extraer y representar significado semántico. El tercero describe las métricas de similitud y los criterios de evaluación del desempeño. El cuarto, finalmente, vincula estos recursos metodológicos con el contexto aplicado a la reforma a la salud en Colombia y su cobertura mediática mostrando cómo el pipeline propuesto (recolección, limpieza, vectorización, clasificación y agrupamiento de noticias) contribuyo a identificar segmentaciones regionales en la opinión pública y con ello, a cumplir el objetivo central del proyecto aplicado.

4.1.1 Fundamentos de generales

Este primer apartado sienta las bases técnicas indispensables para el resto del trabajo. Se describe cómo el texto inicial se transforma mediante tokenización, lematización, eliminación de stop-words y minería de texto en insumos estructurados que un algoritmo puede interpretar. A continuación, se desarrollan estas definiciones:

4.1.1.1 Procesamiento del Lenguaje Natural (NLP)

El procesamiento de lenguaje natural (NLP) es un área de machine learning que brinda a las computadoras la capacidad de interpretar, manipular y comprender el lenguaje humano. Hoy en día, las organizaciones tienen grandes volúmenes de datos de voz y texto de varios canales de comunicación, como correos electrónicos, mensajes de texto, fuentes de noticias en redes sociales, vídeo, audio y más. Utilizan software de NLP para procesar de forma automática estos datos, analizan la intención o el sentimiento del mensaje y responden en tiempo real a la comunicación humana [9]. Se presentan las definiciones o bases técnicas utilizadas en esta tecnología:

- **Tokenización:** Se refiere al proceso de convertir una secuencia de texto en partes más pequeñas, conocidas como tokens (como caracteres o palabra). Permite a las máquinas

a comprender el lenguaje humano al descomponerlo en trozos, que son más fáciles de analizar [10].

- **Lematización:** Es una técnica de preprocesamiento de textos donde se reducen las formas flexionadas de las palabras en un conjunto de datos de texto a una palabra raíz común o forma de diccionario, también conocida como "lema" en lingüística computacional [11].
- **Eliminación de Stop Words**: Hacen referencia a aquellas palabras que carecen de sentido cuando se escriben solas o sin la palabra clave o keyword, por lo que no aportan de forma significativa al análisis del texto, por lo tanto, se eliminan [12].
- **Análisis de Frecuencia de Palabras**: Técnica que consiste en contar la frecuencia con la que aparecen las palabras en un texto o conjunto de textos. Este análisis permite identificar las palabras más utilizadas, lo que puede ayudar a revelar patrones, temas o asuntos importantes [13].

4.1.1.2 Minería de Texto

La minería de texto es un conjunto de métodos computacionales que persiguen descubrir patrones relevantes y conocimiento accionable dentro de grandes volúmenes de datos textuales no estructurados; para ello resulta imprescindible transformar el lenguaje natural en un formato numérico manejable mediante un flujo de preprocesamiento que incluye la normalización ortográfica, la segmentación léxica o tokenización, la reducción morfológica mediante lematización o stemming, llevando las palabras a su forma base o raíz y finalmente, la vectorización del corpus. Solo tras esta codificación estructurada pueden aplicarse con rigor los algoritmos estadísticos y de aprendizaje automático tales como clustering, clasificación o análisis de asociaciones que permiten extraer conocimiento latente, detectar tendencias y generar valor estratégico a partir del texto procesado [14].

4.1.1.3 Análisis de Sentimientos

Una herramienta fundamental en la minería de texto permite descifrar el tono emocional subyacente en un texto. Esta técnica, de gran utilidad en diversas áreas, clasifica las opiniones expresadas como aprobación, desaprobación, neutralidad, esperanza, preocupación, confianza, desconfianza, entre otros; brindando información valiosa sobre las actitudes y perspectivas presentes en el contenido analizado [15].

4.1.1.4 API (Application Programming Interface)

API es un conjunto de protocolos, reglas e interfaces que permiten la comunicación entre diferentes sistemas de software. Estas interfaces facilitan la recopilación, procesamiento y análisis automatizado de datos textuales desde diversas fuentes mediáticas [16].

4.1.1.5 Lexicón

Se define como con conjunto de palabras o lemas que contiene una orientación semántica del conjunto de palabras, siendo estas positivas o negativas. Con el objetivo de etiquetar un documento, texto, oración u opinión de acuerdo con su orientación semántica o polaridad [17].

4.1.2 Modelos de similaridad y análisis de texto

Se presentan las estrategias que convierten el texto preprocesado en vectores numéricos portadores de significado semántico. Se parte de enfoques clásicos (TF-IDF, Doc2Vec) y se avanza hacia arquitecturas Transformer de última generación (BERT, BETO, RoBERTa, MPNet y GPT-4o), detallando su aporte al análisis de sentimientos y la identificación de patrones regionales. Este bloque demuestra cómo la elección de un modelo afecta la capacidad del sistema para captar matices lingüísticos presentes en las noticias sobre la reforma:

4.1.2.1 Vectorización de textos

La vectorización agrupa un conjunto amplio de técnicas destinadas a traducir texto en vectores numéricos que capturen no solo características léxicas sino también patrones y matices semánticos; estas estrategias van desde enfoques clásicos basados en n-gramas y TF-IDF hasta incrustaciones densas como Word2Vec o GloVe, más recientemente, embeddings contextuales generados por grandes modelos de lenguaje. El propósito fundamental es adecuar los datos textuales al formato exigido por la mayoría de los algoritmos de aprendizaje automático preservando relaciones semánticas latentes. Por ejemplo, el modelo Distributed Memory (DM), aprende un vector de documento prediciendo palabras en contexto al considerar simultáneamente los vectores de las palabras circundantes y el vector que representa al documento en su conjunto [18].

4.1.2.2 TF-IDF

TF-IDF asigna un peso a cada palabra en función de su frecuencia en un documento (TF) y su rareza en el corpus completo (IDF), destacando términos distintivos de cada texto. La fórmula del modelo TF-IDF (Term Frequency-Inverse Document Frequency) es una combinación de dos medidas presentadas a continuación [19]:

1. **TF (Term Frequency):** La frecuencia de un término en un documento específico.
2. **IDF (Inverse Document Frequency):** La frecuencia inversa de un término en todo el conjunto de documentos.

La fórmula del **TF-IDF** en términos generales es:

$$\text{TF-IDF}(i,j) = \text{TF}(i,j) \times \text{IDF}(i)),$$

Formula de TF (Term Frequency) se describe como:

$$\text{TF}(i,j) = \frac{\text{Freq}(i,j)}{\sum_k^{\text{Freq}}(k,j)},$$

Donde está conformado por:

- Freq(i,j) es la cantidad de veces que el término **i** aparece en el documento **j**.
- La suma del denominador es el número total de términos en el documento **j.**

Por otro lado, para la Fórmula de IDF (Inverse Document Frequency) está definida por la siguiente ecuación:

$$\text{IDF}(i) = \log\left(\frac{N}{\text{DF}(i)}\right),$$

Siendo sus componentes:

- N es el número total de documentos en el corpus.
- DF(i) es el número de documentos en los que aparece el término **i**.

Por último, la fórmula Completa de TF-IDF es:

$$\text{TF-IDF}(i,j) = \frac{\text{Freq}(i,j)}{\sum_k^{\text{Freq}}(k,j)} \times \log\left(\frac{N}{\text{DF}(i)}\right).$$

4.1.2.3 Modelo Doc2Vec

El modelo Doc2Vec, es una técnica utilizada en procesamiento de lenguaje natural (PLN) para representar documentos como vectores de características en un espacio vectorial. Fue introducido por Mikolov et al. en 2014 y está diseñado específicamente para capturar el significado semántico de un documento completo, en lugar de trabajar únicamente con palabras individuales como lo hace Word2Vec [20].

4.1.2.3.1 Fundamentos del modelo Doc2Vec

El modelo Doc2Vec tiene como objetivo aprender representaciones vectoriales para documentos completos, capturando patrones semánticos entre palabras y su contexto en el documento. Fue propuesto por Mikolov et al. en 2014 y permite representar documentos de cualquier longitud como vectores en un espacio continuo de baja dimensión [20].

4.1.2.3.2 Funcionamiento básico:

Representación de documentos y palabras: Cada documento es identificado mediante un ID único asociado a un vector. Las palabras dentro del documento también son representadas como vectores aprendidos durante el proceso de entrenamiento [20].
Proceso de aprendizaje: Se utiliza una red neuronal para ajustar los vectores de documentos y palabras en el corpus, de manera que documenten similares se posicionen

más cerca en el espacio vectorial [20].

4.1.2.3.3 Entrenamiento del modelo:

Doc2Vec utiliza redes neuronales para aprender vectores. Asocia un ID único a cada documento y aprende un vector asociado a este ID junto con las representaciones de las palabras en el documento [20].

4.1.2.3.4 Arquitecturas de Doc2Vec, Existen dos variantes principales:

Distributed Memory (DM): Aprende el vector del documento prediciendo palabras en contexto, considerando tanto las palabras como el vector del documento [20].
Distributed Bag of Words (DBOW): Aprende el vector del documento prediciendo palabras individuales en el documento sin considerar su contexto [20]. Se selecciona la variante Distributed Bag of Words (DBOW) de Doc2Vec, debido a su capacidad para generar vectores representativos de documentos completos con menor costo computacional y alta capacidad de generalización. La elección de esta variante se justificó en los siguientes puntos:
Eficiencia en contextos globales: DBOW optimiza los vectores de documentos completos para pre- decir palabras, capturando el contenido semántico de manera efectiva.
Idoneidad para análisis comparativos: Genera representaciones densas que son ideales para evaluar similitudes mediante métricas como la similitud coseno.

4.1.2.4 Arquitectura Transformer

Un Transformer es un modelo de aprendizaje profundo introducido en el año 2017 [21], diseñado para tareas de secuencia a secuencia en procesamiento de lenguaje natural. A diferencia de las redes recurrentes, prescinde por completo de la recurrencia y se basa en un mecanismo de autoatención que asigna pesos variables a cada posición de la entrada, lo que le permite capturar relaciones de largo alcance de manera eficiente y entrenar de forma completamente paralela en GPU. La arquitectura original consta de dos bloques principales: un Encoder, que procesa la secuencia de entrada y la convierte en una representación continua rica en contexto, y un Decoder, que, tomando dicha representación junto con la salida generada hasta el momento, produce la secuencia objetivo token a token [21]. En el diseño estándar se apilan seis capas idénticas de codificador y seis de decodificador, aunque este número puede ajustarse según la aplicación. Gracias a este esquema, los Transformers han logrado resultados de vanguardia en traducción automática, resumen de textos y otras tareas de NLP, y su flexibilidad los ha llevado a extenderse incluso a dominios como visión por computadora y audio.

La innovación clave del Transformer es su atención escalada por producto punto (Scaled Dot-Product Attention) implementada en múltiples cabezas que operan en paralelo, lo que captura dependencias globales entre todas las posiciones de entrada y salida [21]. Cada capa alterna subcapas de atención multi-cabeza con redes feed-forward posición-a-posición, y utiliza

conexiones residuales seguidas de normalización de capa para estabilizar el entrenamiento. Para inyectar información de orden sin recurrir a bucles, se suman codificaciones posicionales (por ejemplo, sinusoidales) a los embeddings de palabras [22]. Esta combinación de componentes elimina las limitaciones de las arquitecturas recurrentes y convolucionales, permitiendo una paralelización masiva y un alcance constante en la modelación de dependencias a largo plazo, lo que se traduce en tiempos de entrenamiento reducidos y una notable mejora de calidad en tareas. Como se observa en la siguiente figura 1.

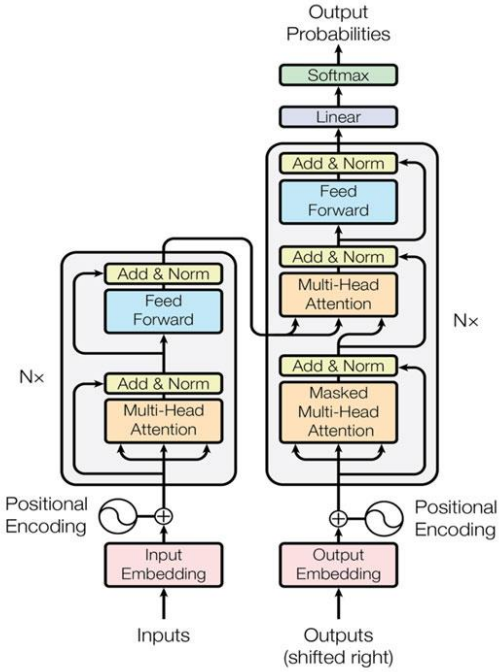

Figura 1. Estructura del modelo Transformers [21].

4.1.2.5 Librerías de Aprendizaje Automático

Son conjuntos de herramientas y bibliotecas de software que facilitan el desarrollo y la implementación de algoritmos de aprendizaje automático algunas de ellas son:

- **Scikit-learn:** Biblioteca de aprendizaje automático en Python que proporciona algoritmos de clasificación y análisis de sentimientos [23].
- **Hugging Face**: Es una biblioteca de software popular en el campo del aprendizaje automático y procesamiento del lenguaje natural (NLP). Hugging Face proporciona diversas herramientas y modelos pre-entrenados que permiten implementar algoritmos avanzados de IA de manera eficiente [24].
- **Gensim:** Biblioteca para modelado de temas y similitud semántica, útil para minería de texto y análisis de grandes volúmenes de documentos [23].

4.1.2.6 Modelo Bert

El modelo BERT fue desarrollado por Google en 2018, es un modelo basado en la

arquitectura Transformer, diseñado para predecir relaciones semánticas entre palabras en un texto considerando el contexto bidireccional. A diferencia de otros modelos previos (como Word2Vec), BERT no asigna un único vector a cada palabra, sino que genera representaciones contextuales, es decir, el significado de una palabra depende de las palabras que la rodean [25].

4.1.2.6.1 Funcionamiento matemático de BERT

El modelo BERT se basa en una serie de capas de transformadores que procesan el texto para generar embeddings ricos en semántica. En este proyecto se utilizó un modelo pre-entrenado de BERT (bert-base- uncased) y a continuación, se describe el proceso matemático subyacente que se realizó:

4.1.2.6.2 Entrada del modelo

Cada texto se divide en tokens mediante un tokenizador específico de BERT. Se agrega un token especial [CLS] al inicio para indicar el comienzo del texto y [SEP] para separar segmentos en textos más largos es de aclarar que máximo se utilizan 512 tokens, más de esa cifra los textos son truncados.
Cada token se convierte en un embedding inicial que combina tres elementos:

- **Embeddings de palabra (*E*)**: Vector que representa el token.
- **Embeddings de posición (*P)***: Vector que indica la posición del token en la secuencia.
- **Embeddings de segmento (*S*)**: Vector que distingue diferentes segmentos de texto en tareas específicas.

La representación final de cada token se calcula con la siguiente ecuación:

$$h_{input} = E + P + S.$$

Este proceso asegura que la representación del texto considere tanto el contenido semántico de las palabras como su posición relativa.

4.1.2.6.3 Capas de transformadores

El modelo BERT aplica varias capas de atención auto-regresiva (*self-attention*) y redes completamente conectadas para generar representaciones contextuales.

4.1.2.6.4 Mecanismo de atención

Cada capa de atención computa una combinación ponderada de las representaciones de las palabras considerando todas las palabras en el texto. La atención auto-regresiva se calcula con la siguiente ecuación:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V,$$

Donde se describe como:

- Q, K, V: Son las matrices de consulta (*query*), clave (*key*) y valor (*value*) derivadas del embedding de entrada.
- d_k: Dimensión de los vectores de consulta (Q).

4.1.2.6.5 Generación de embeddings

El modelo genera un vector contextual para cada token en el texto. En esta implementación, se utilizó el promedio de todos los vectores de tokens (*mean pooling*) para representar el texto completo, en lugar de solo el embedding correspondiente al token [CLS]. Esto mejora la capacidad del modelo para capturar la semántica global del texto.

4.1.2.6.6 Generación de Representaciones por Región

Para cada región, se generó una representación vectorial combinando los embeddings de todas las noticias individuales. En lugar de concatenar todos los textos en una sola entrada, se calcularon los embeddings de cada noticia individual y luego se promedió su representación con la siguiente ecuación:

$$V_{regin} = \sum_{i=1}^{N^V} v_{noticias}$$

Donde está conformado por:

- $\mathbf{v}_{region}$: Representación de la región.
- $\mathbf{v}_{noticias}$: Embedding promedio de la noticia i.
- N: Número total de noticias en la región.

Este enfoque garantiza que la representación regional no sea dominada por textos individuales largos.

4.1.2.7 Similitud entre Regiones

Para comparar las regiones, se calculó la similitud entre los vectores generados mediante la métrica de similitud coseno con la siguiente ecuación:

$$\text{Similitud de Coseno (A, B)} = \frac{A \cdot B}{||A|| \, |B|},$$

Está conformada por:

- $\mathbf{v}_A$ y $\mathbf{v}_B$: Representaciones vectoriales de las regiones A y B.
- $\|\mathbf{v}_A\|$: Magnitud del vector $\mathbf{v}_A$.

4.1.2.7.1 Interpretación de los valores

Un valor próximo a 1 indica una alta similitud semántica, mientras que uno cercano a 0 denota una similitud escasa o inexistente.

4.1.2.8 Modelo MPNet “sentence-transformers/paraphrase-multilingual-mpnet-base-v2”

El modelo "sentence-transformers/paraphrase-multilingual-mpnet-base-v2" es un modelo de transformador basado en MPNet (Masked and Permuted Language Model) diseñado para generar representaciones densas de oraciones en múltiples idiomas. Este modelo forma parte de la arquitectura Sentence-Transformers, lo que le permite capturar relaciones semánticas entre oraciones, mejorando tareas como búsqueda semántica, clasificación de textos y agrupación de documentos [26].
El modelo funciona convirtiendo una oración en un vector de características en un espacio de alta dimensión. Para entender su funcionamiento, es importante desglosar los siguientes pasos clave:

4.1.2.8.1 Tokenización

Dado un texto de entrada, el modelo aplica una tokenización utilizando un tokenizador basado en WordPiece. La oración se divide en subpalabras y se convierte en una secuencia de tokens, Luego, cada token se convierte en un índice correspondiente dentro del vocabulario del modelo.

4.1.2.8.2 Embeddings de Entrada

Cada token se convierte en un vector de embedding de tamaño fijo, extrayendo la representación de una matriz pre-entrenada, se añaden positional embeddings para retener información sobre la posición de las palabras en la oración:

4.1.2.8.3 Paso por la Arquitectura MPNet

MPNet es una variante de BERT que combina enmascaramiento bidireccional (como en BERT) y orden permutado (como en XLNet) [27]. Su entrenamiento sigue dos principios clave:

- **Máscara bidireccional:** Se ocultan ciertas palabras en la oración y el modelo predice su contenido utilizando el contexto bidireccional.
- **Predicción con permutación:** Se introducen diferentes órdenes en la oración para reforzar la capacidad del modelo de capturar relaciones globales dentro del texto.

El modelo usa múltiples capas de transformadores con mecanismos de autoatención, donde cada token se procesa a través de la ecuación de atención. Para obtener una representación única de la oración, se aplica un mecanismo de agregación. En este caso, se usa el **[CLS] token pooling**, extrayendo la representación final del token [CLS] de la última capa del transformador [28]**.**

4.1.2.9 FAISS: Búsqueda Vectorial Aproximada Basada en Similitud

FAISS (Facebook AI Similarity Search) es una biblioteca desarrollada por Meta (anteriormente Facebook) que permite realizar búsquedas rápidas y eficientes de vectores similares en

grandes bases de datos. Su principal función es encontrar los elementos más parecidos entre sí según su representación vectorial, lo cual es especialmente útil en tareas de procesamiento de lenguaje natural, como el análisis de textos, donde cada documento se representa como un vector numérico [29].

Lo que hace FAISS no es entrenar un modelo predictivo, sino facilitar la búsqueda aproximada por similitud entre millones de vectores. En este trabajo, FAISS se utiliza para comparar representaciones vectoriales de textos, como titulares de noticias o agrupaciones regionales, y así identificar qué tan similares son entre sí. Para lograr eficiencia, FAISS utiliza una técnica llamada cuantización de vectores. Esta técnica consiste en reemplazar los vectores originales por versiones más pequeñas o aproximadas llamadas *centroides*. Esto reduce el espacio de almacenamiento y permite hacer búsquedas mucho más rápidas. Por ejemplo, en lugar de comparar directamente todos los textos entre sí, FAISS agrupa vectores similares y solo compara dentro de los grupos más relevantes.
Entre las estrategias de FAISS se destacan:

- Product Quantization (PQ): divide los vectores en partes más pequeñas y las representa de forma más compacta.
- Inverted File Index (IVF): organiza los vectores en clústeres para evitar comparaciones innecesarias.

Estas técnicas hacen posible que FAISS realice búsquedas de alta velocidad incluso cuando se trabaja con cientos de miles de textos o más, manteniendo un buen equilibrio entre precisión y eficiencia.

4.1.2.10 BETO variante "initeautomata/beto-sentiment-analysis"

Es modelo basado en BERT (Bidirectional Encoder Representations from Transformers) pre-entrenado en español. BETO fue desarrollado por el equipo de Hugging Face y está optimizado para tareas de procesamiento de lenguaje natural (NLP) en español [30].
Este modelo ha sido ajustado específicamente para análisis de sentimientos, lo que significa que puede clasificar el sentimiento de un texto en positivo, negativo o neutro. Se entrena con conjuntos de datos en español, asegurando una alta precisión en la interpretación de emociones en textos escritos en este idioma [31].

4.1.2.11 Modelo RoBERTa “pysentimiento/robertuito-sentiment-analysis”

El modelo RoBERTuito-sentiment-analysis es un clasificador de sentimientos en idioma español basado en la arquitectura RoBERTa. En esencia, RoBERTa es un modelo de lenguaje grande de tipo Transformer entrenado originalmente sobre una gran colección de textos de Twitter en español [32]. Su arquitectura sigue la configuración RoBERTa-base, equivalente a BERT-base, con 12 capas de transformador de auto-atención, 12 cabezales de atención en cada capa y un tamaño de representación oculto de 768 dimensiones [33]. Esto supone el orden de 110–125 millones de parámetros entrenables, similar a otros modelos BERT-base en español (como BETO) y a la versión base de RoBERTa en inglés.

Como modelo basado en Transformers, RoBERTuito emplea un codificador bidireccional que

procesa el texto de entrada completo para capturar las dependencias contextuales. Utiliza tokenización sub-palabra (Byte-Pair Encoding, BPE) adecuada para español informal, incorporando tokens especiales para elementos. Luego pasa por las 12 capas transformadoras produciendo representaciones contextuales, y finalmente la representación del token <s> alimenta la capa lineal que emite una probabilidad para cada categoría de sentimiento (POS, NEG o NEU). En consecuencia, RoBERTuito-sentiment-analysis puede asignar una etiqueta de sentimiento (positivo, negativo o neutro) a cada texto, aprovechando la comprensión profunda del contexto lingüístico adquirida en el preentrenamiento [34]. Aunque modelos pre-entrenados como BETO y RoBERTa han sido entrenados con grandes volúmenes de texto en español, su desempeño puede mejorarse significativamente y adaptarse a temas en específico mediante el proceso de fine-tuning. Esta técnica consiste en reentrenar el modelo sobre un conjunto de datos específicos del dominio de interés, permitiendo que se adapte mejor al vocabulario, estilo y contexto particular de la tarea.

4.1.2.12 Fine-Tuning Clásico

El *fine-tuning clásico* es una técnica básica en el aprendizaje por transferencia de conocimiento (*transfer learning*) que consiste en tomar un modelo de lenguaje previamente entrenado sobre grandes volúmenes de texto general como Wikipedia, Common Crawl o libros digitales y reentrenarlo en su totalidad utilizando un conjunto de datos más pequeño y específico, con el objetivo de adaptarlo a una tarea concreta [35]. Este proceso se realiza mediante la retro propagación del error, actualizando todos los pesos y parámetros del modelo base. Esta característica lo convierte en un enfoque más costoso computacionalmente, pero también más potente en términos de capacidad de adaptación. Gracias a ello, el modelo puede captar matices complejos y ajustarse mejor a dominios con vocabulario o estructuras poco frecuentes en los corpus de entrenamiento general.

El principio detrás del fine-tuning clásico es que los modelos de lenguaje grandes han aprendido representaciones lingüísticas generales útiles como gramática, sintaxis, semántica y relaciones contextuales que pueden ser reutilizadas en tareas más especializadas, como la clasificación de sentimientos, el análisis de entidades nombradas o la detección de ciberacoso. En esta metodología, se puede optar por reentrenar el modelo completo o, más comúnmente, ajustar únicamente ciertas capas superiores, mientras se congelan las capas inferiores que contienen conocimientos lingüísticos más generales. Esta estrategia permite una mayor eficiencia computacional y evita sobreajustar el modelo a un conjunto de datos reducido. Al reentrenar estas partes específicas, el modelo adapta sus representaciones a los patrones particulares del nuevo dominio, mejorando su rendimiento en tareas específicas.

4.1.2.13 Modelo GPT-4o "(Generative Pre-trained Transformer 4o) "

Es un modelo de lenguaje autorregresivo basado en la arquitectura Transformer, que representa el estado del arte en procesamiento de lenguaje natural [36]. Su funcionamiento se fundamenta en los siguientes componentes matemáticos:

4.1.2.13.1 Mecanismo de Auto-atención

El núcleo de GPT-4 es el mecanismo de auto-atención, que calcula representaciones contextuales de cada palabra mediante la ecuación presentada a continuación:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V,$$

Donde sus componentes son:

- Q (Queries), K (Keys) y V (Values) son matrices aprendidas durante el entrenamiento.
- dk es la dimensión de las claves, utilizada como factor de escalamiento para estabilizar el cálculo.

4.1.2.13.2 Atención Multi-cabeza

GPT-4 utiliza **atención multi-cabeza** (con $h=12\ h$ a $128h$ cabezas en modelos grandes). Cada "cabeza" de atención aprende patrones diferentes en los datos, y sus resultados se combinan para producir una representación más rica. La fórmula para la atención multi-cabeza es:

$$MultiHead(Q, K, V) = Concat(head_1, \ldots., head_h)W_0.$$

Aquí, cada head realiza un cálculo de atención, y sus resultados se concatenan y se multiplican por una matriz W_0 para obtener la salida final.

4.1.2.13.3 Normalización y Feed-Forward

En cada capa del modelo se incluye normalización de capa (LayerNorm) y una red feed-forward. La normalización de capa se calcula implementando la siguiente ecuación:

$$LayerNorm(x) = \gamma\frac{x-\mu}{\sigma} + \beta,$$

De tal manera que está conformada por:

- μ la media.
- σ es la desviación estándar de la entrada.
- γ y β son parámetros aprendidos durante el entrenamiento.

La red feed-forward utiliza la activación GELU (Gaussian Error Linear Unit), que se define como:

$$GELU(x) = x\phi\ (x),$$

Donde $\phi\ (x)$ es la función de distribución acumulativa normal estándar.

4.1.2.13.4 Preentrenamiento y Fine-tuning Supervisado (SFT)

GPT-4.o se entrena inicialmente mediante un proceso de preentrenamiento no supervisado, utilizando la técnica de máxima verosimilitud. En esta fase, el modelo aprende a predecir el siguiente token dado un contexto [37]. La función de pérdida durante el preentrenamiento está definida por la siguiente ecuación:

$$L(\theta) = -\sum_{t=1}^{T} \log P(x_t | x_{<t}; \theta),$$

Donde sus componentes son X_t que es el token en la posición t, y θ que son los parámetros del modelo.

Posteriormente, GPT-4.o pasa por una fase de fine-tuning supervisado (SFT), en la que se le entrena con ejemplos curados por humanos para mejorar su capacidad de seguir instrucciones, responder preguntas o realizar tareas específicas. Esta etapa permite adaptar el modelo a tareas concretas sin necesidad de reentrenar toda la arquitectura desde cero.

Finalmente, se aplica una etapa adicional conocida como Reinforcement Learning from Human Feedback (RLHF). En esta, se entrena un modelo de recompensa $r\phi\ (x, y)$ con base en evaluaciones humanas y se optimiza el modelo de lenguaje mediante el algoritmo PPO (Proximal Policy Optimization), que tiene la siguiente ecuación de función de pérdida:

$$LPPO(\theta) = E[\min\ (r_t(\theta)\widehat{A_t}, clip(r_t(\theta), 1-\varepsilon, 1+\varepsilon)r_t(\theta)],$$

Donde r_t es la razón de probabilidad entre la política actual y la anterior, $\widehat{A_t}$ es la ventaja estimada, y ϵ es un pequeño valor que garantiza la estabilidad en el entrenamiento [38].

4.1.2.13.5 Características Específicas de GPT-4o

GPT-4o presenta varias mejoras con respecto a sus versiones anteriores:

- Escala: GPT-4o es significativamente más grande, con más de 1 trillón de parámetros.
- Contexto Extendido: A diferencia de versiones anteriores, GPT-4 puede manejar contextos más largos, con un límite de hasta 32,000 tokens.
- Reducción de Alucinaciones: A través de mecanismos avanzados, se han mejorado la coherencia y la reducción de errores, también conocidos como "alucinaciones", en las respuestas generadas.
- GPT-4o es un transformer dedicado a la parte decoder por lo que es tan eficiente en IA generativa.

4.1.3 Métricas de evaluación de un modelo

En el tercer apartado se presentan los criterios utilizados para evaluar el desempeño del pipeline. Se emplearon métricas de similitud y de rendimiento predictivo. Entre las primeras se encuentran la similitud de coseno y la distancia euclidiana, cuyas definiciones ya fueron introducidas en una sección anterior. En este caso, dichas métricas se aplican a las representaciones vectoriales de textos (por ejemplo, vectores TF-IDF), con el fin de comparar niveles de similitud semántica entre documentos relacionados con la reforma a la salud.

Por otro lado, para evaluar la calidad del modelo de clasificación de sentimientos se utilizaron métricas estándar como accuracy, precisión, recall, F1-score y AUC-ROC, que permiten medir la capacidad del modelo para identificar correctamente los distintos tipos de sentimiento presentes en los textos analizados. Estas métricas permiten comparar diferentes enfoques de modelado y asegurar que los hallazgos sobre la cobertura mediática sean estadísticamente válidos, consistentes y reproducibles.

4.1.3.1 Similitud de Coseno

Mide el ángulo entre dos vectores, donde valores cercanos a 1 indican una alta similitud entre ellos. En el contexto del análisis de texto, los vectores son generados a partir del modelo TF-IDF [39]. Su fórmula es:

$$\text{Similitud de Coseno } \backslash(\text{A, B}\backslash)=\frac{A \cdot B}{||A||B|}.$$

Donde se describe:

- A · B es el producto punto entre los vectores A y B.
- II A II es la norma (o longitud) del vector A.
- II B II es la norma (o longitud) del vector B.

Esto permite comparar textos independientemente de su longitud.

4.1.3.2 Distancia Euclidiana

Mide la distancia directa entre los vectores en el espacio, donde valores más bajos indican mayor similitud [39]. Su fórmula es:

$$\text{Distancia Euclidiana}(A, B) = \sqrt{\sum_{i=1}^{n}(A_i - B_i)^2}.$$

De tal manera que sus componentes son:

- A_i y B_i son los elementos del vector A y B, respectivamente, en la i- ésima dimensión.
- La suma recorre todas las dimensiones n del espacio de los vectores.

4.1.3.3 Accuracy (Exactitud)

La accuracy mide la proporción de predicciones correctas que hace el modelo, en relación con la totalidad de ejemplos evaluados. Sin embargo, no es confiable en conjuntos de datos desbalanceados (cuando una clase es mucho más frecuente que las demás) [40]. Se expresa como:

$$Accuracy = \frac{Numero\ de\ predicciones\ correctas}{Total\ de\ predicciones},$$

4.1.3.4 Precisión (Precisión)

La precisión se refiere a la fracción de ejemplos predichos como positivos que realmente son positivos [41]. En un contexto de clasificación binaria se describe con la siguiente ecuación:

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}},$$

Donde:

- TP (True Positives) son los verdaderos positivos,
- FP (False Positives) son los falsos positivos.

4.1.3.5 Recall (Sensibilidad)

La recall mide la fracción de ejemplos verdaderamente positivos que el modelo logra capturar [16]. En clasificación binaria con la fórmula que se muestra a continuación:

$$Recall = \frac{TP}{TP+FN},$$

Donde sus componentes son:

- TP (True Positives) son los verdaderos positivos,
- FN (False Negatives) son los falsos negativos.

4.1.3.6 F1-Score

La F1-Score es la media armónica entre la precisión y la recall. Ayuda a balancear ambos aspectos (precisión y exhaustividad) en una sola métrica, siendo muy utilizada cuando hay un desbalance de clases o cuando se considera tanto la precisión como la sensibilidad igualmente importante [16]. Su ecuación de describe de la siguiente manera:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}.$$

4.1.3.7 AUC-ROC (Area Under the ROC Curve)

La métrica AUC-ROC evalúa cuán bien un modelo clasifica ejemplos positivos y negativos en una tarea de clasificación binaria. La curva ROC muestra la relación entre la tasa de verdaderos

positivos (TPR) y la tasa de falsos positivos (FPR) a diferentes umbrales de decisión. El AUC (área bajo la curva ROC) oscila entre 0 y 1, siendo 1 un rendimiento perfecto [40].

4.1.3.8 Análisis de polaridad en textos

El análisis de polaridad de texto es un problema de clasificación que tiene como objetivo identificar y determinar la actitud expresada en un texto, categorizándola como positiva, negativa o neutra. En el contexto de redes sociales, este proceso permite evaluar las opiniones emitidas por los usuarios sobre un tema específico [42].

4.1.4 Entorno de implantación del proyecto

El cuarto apartado articula la estrategia metodológica con el fenómeno social objeto de estudio: la manera en que los medios cubren la reforma al sistema de salud en Colombia. Para contextualizar esta conexión se presentan, a continuación, los dos conceptos clave que enmarcan el estudio:

4.1.4.1 Reforma a la Salud

La reforma a la salud en Colombia consiste en una serie de cambios estructurales al Sistema de Salud con el objetivo de garantizar el derecho fundamental a la salud de todos los ciudadanos. Propone una reestructuración profunda del sistema sanitario nacional, con el objetivo declarado de garantizar el derecho fundamental a la salud y corregir las falencias estructurales acumuladas desde la implementación de la Ley 100 de 1993 [43]. Entre sus principales medidas se incluyen: la eliminación del sistema de aseguramiento administrado por las EPS, la creación de redes integradas e integrales de prestación de servicios, el fortalecimiento de la atención primaria, la descentralización de recursos hacia entes territoriales y el establecimiento de un sistema de información interoperable [44].

El actual modelo de salud, basado en un esquema de competencia regulada entre entidades promotoras de salud (EPS), ha sido criticado por profundizar las inequidades regionales, restringir el acceso efectivo a servicios, generar barreras administrativas y priorizar lógicas financieras por encima del bienestar del paciente [45]. Diversos estudios han evidenciado brechas significativas en la cobertura, calidad y oportunidad de atención entre zonas urbanas y rurales, así como entre regiones del país [46]. En este sentido, la reforma busca superar estos obstáculos y recuperar el rol central del Estado como garante del derecho a la salud.

No obstante, la iniciativa ha generado un fuerte debate público. Por un lado, sus defensores consideran que representa un cambio necesario hacia un sistema más humano, equitativo y orientado a la prevención; por otro lado, sus detractores advierten que podría politizar la prestación de servicios, poner en riesgo la sostenibilidad financiera del sistema y afectar la calidad de la atención al eliminar actores que han acumulado experiencia operativa [47]. Este escenario ha propiciado una alta polarización mediática y política, donde distintos actores

gobierno, gremios médicos, EPS, usuarios, oposición política manifiestan posturas enfrentadas respecto al rumbo de la política sanitaria del país [48] mientras por otro lado la reforma es percibida por el gobierno de Colombia como una oportunidad para mejorar la calidad y acceso a los servicios de salud, promoviendo la transparencia, la participación social y la eficiencia en la gestión de recursos, con el objetivo de lograr una atención primaria en salud resolutiva y de calidad para toda la población [49].

4.1.4.2 Cobertura Mediática

La cobertura mediática se refiere a la forma en que los medios de comunicación informan sobre un tema, en este caso en la reforma a la salud, basándose en la selección de información, enfoques dados a los eventos, marcos interpretativos y presentación de la información. Esta cobertura puede afectar la percepción de la población al influir en la opinión pública, moldear la atención a ciertos temas y en la forma en que se perciben la reforma propuesta por el gobierno. Los medios pueden destacar aspectos específicos, sesgar la información e influir en la agenda pública, teniendo un impacto significativo en la opinión y comprensión de la audiencia sobre estos temas [50].

4.2 ANTECEDENTES

• Estudios como el de Daemin Park, Dasom Kim y Ah-hyun Park (2024) analizaron la cobertura mediática sobre la enfermería en Corea del Sur entre 2005 y 2022, empleando técnicas avanzadas de procesamiento del lenguaje natural (NLP) y análisis de redes semánticas. Utilizaron métodos como la tokenización, que divide el texto en unidades más pequeñas para facilitar su análisis; el reconocimiento de entidades nombradas (NER), que identifica y clasifica entidades importantes en un texto, como nombres de personas, fechas o lugares; y el análisis de sentimientos, que determina la actitud o emoción expresada en un texto. Este enfoque metodológico permitió identificar patrones de opinión y representación en los medios, Los hallazgos revelaron que la mayoría de las noticias se publicaron en las secciones sociales y locales, mientras que las secciones de economía y política, cruciales para debates profundos sobre políticas públicas, dedicaron una cobertura significativamente menor. Además, en las escasas ocasiones en que se abordaron leyes relacionadas con la enfermería en la sección política, el enfoque predominante fue en los conflictos políticos, sin profundizar en las ventajas o desventajas de dichas legislaciones. Este sesgo en la cobertura mediática podría haber limitado la elaboración de políticas públicas efectivas para mejorar la profesión de enfermería en el país [51].

• En el estudio realizado por Anna Ruelens (2022), enfatiza la importancia de las técnicas de Procesamiento de lenguaje Natural (NLP), para revelar el sentimiento e intentar medir el grado de satisfacción de los usuarios en línea, en cuanto al sistema de salud de los Estados Unidos, ya que estas técnicas permiten trabajar con datos no estructurados y que pueden revelar otro tipo de información oculta que no logran presentar los datos que se consiguen a través de

encuestas, las 2 técnicas que se utilizaron fueron, el análisis de frecuencia de palabras y el análisis de sentimiento, al aplicar estas técnicas el estudio pudo evidenciar, que los temas que más le preocupan a los usuarios son la asequibilidad y la calidad de los servicios de salud, resultados que son similares a los encontrados en datos de tipo encuesta. Este hallazgo es significativo puesto que los costos para extraer y procesar datos en línea son mucho menores que realizar las encuestas [52].

• En el estudio realizado por Miao Chu, Yi Chen, Lin Yan y Junfang Wang (2022), si bien las técnicas de análisis de sentimiento y minería de texto no fueron aplicadas en un contexto de la salud, si no que se enfocaron en analizar la plataforma de viajes Tripadvisor, pero su aporte es valioso puesto que las técnicas de análisis de sentimiento tradicionales se basan en la estadística, sin embargo es difícil clasificar el sentimiento de las personas a través de un contexto, por lo que los autores propusieron un modelo de análisis de sentimiento basado en aprendizaje profundo, utilizando redes neuronales, basado en el modelo de representación de codificador bidireccional de Transformers (BERT), este enfoque no utiliza el modelado de la dimensión temporal, lo cual lo hace atractivo en cuanto a eficiencia del tiempo y también porque puede analizar frases con longitudes un poco más grandes, al utilizar el Modelo BERT, los investigadores alcanzaron una precisión de un 87,29% en los resultados, además se pudo cumplir con los objetivos del proyecto, que era establecer un modelo de clasificación de los comentarios de los usuarios de la plataforma, en donde se encontró una polaridad negativa en dichos comentarios con 683 calificaciones, sin embargo los comentarios positivos tuvieron 90 y los neutros 101, por otra parte también descubrieron que el sentimiento negativo de los usuarios fueron influenciados por otros comentarios negativos previamente [53].

• En el ámbito de la sostenibilidad, destaca el trabajo de Alberto Flores Pastor (2023) de la Universitat Rovira i Virgili, titulado "Análisis de similitud semántica de tweets sobre sostenibilidad utilizando modelos BERT pre-entrenados y técnicas de generación de embeddings". Este estudio se centra en identificar y clasificar tweets relacionados con la sostenibilidad en subtemas específicos mediante técnicas avanzadas de procesamiento de lenguaje natural (NLP). El autor recopiló un conjunto de tweets en español de instituciones turísticas entre 2018 y principios de 2022. Utilizando el modelo de lenguaje BERT, se calcularon las representaciones vectoriales (embeddings) de estos tweets y de frases de referencia relacionadas con la sostenibilidad. Posteriormente, se determinó la similitud entre ellos, lo que permitió asignar subtemas específicos a cada tweet. Este enfoque facilitó un análisis exhaustivo para evaluar la precisión y exhaustividad de los resultados, identificando la frecuencia y relevancia de los subtemas más destacados en los tweets recopilados. Los resultados obtenidos proporcionaron una visión detallada de cómo se aborda la sostenibilidad en el ámbito turístico a través de las redes sociales, permitiendo identificar patrones y tendencias en la comunicación digital sobre este tema. El análisis reveló que ciertos subtemas de sostenibilidad son más prevalentes en las comunicaciones de las instituciones turísticas, lo que refleja las áreas de mayor interés o preocupación dentro del sector. Además, la aplicación

de modelos BERT permitió una clasificación más precisa de los tweets, mejorando la comprensión de cómo se discuten y promueven las prácticas sostenibles en el turismo [54].

• El estudio realizado por Nicholas Gahman y Vinayak Elangovan (2023), nos hace un aporte valioso a la investigación que estamos haciendo, ya que ellos realizan una compilación de los algoritmos que son más utilizados para hacer comparaciones y similitudes entre texto, este tema es clave puesto que para hacer este trabajo se utilizan técnicas de procesamiento de lenguaje natural (PLN), que es lo que ayuda a generar resúmenes de texto, y analizar similitudes de acuerdo a su contenido, los autores hacen una clasificación de esos algoritmos que ayudan a realizar las funciones anteriormente y los clasifican en 3 grupos que son: Los algoritmos estadísticos, redes neuronales y algoritmos basado en corpus o conocimiento, además muestran que algoritmos son más eficiente según la categoría en la que pertenece, finalmente ellos concluyen que hay 4 algoritmos que son útiles para realizar análisis de texto, entre esos se encuentra Semantic BERT + Cosine Similarity, MT-DNN, XLNet y Lin Measure + String Similarity cada uno explicado con sus ventajas y desventajas [55].

• Ahora centrándonos en el contexto de Colombia encontramos el estudio realizado por William Atencia, José Bustillo, Juan Rambal (2020), en su investigación, ellos utilizan las técnicas de procesamiento de lenguajes natural, para analizar los tweets asociados a la política y la polarización en Colombia, su aporte es importante en el proyecto puesto que ellos aplicaron un proceso en el cual nos encontramos muy familiarizados en el mundo de la ciencia de datos y son los procesos ETL (Extracción, Transformación y Cargue de la Información), este proceso les permitió automatizar la recolección de los datos, conectándose directamente a la API de Twitter, evitando hacerlo de manera manual lo cual puede inducir a errores humanos o perdida valiosa de la información, proceso que nos puede servir de guía al momento en que nos encontremos recolectando la información proveniente de los portales web de los medios de comunicación, luego de esto, ellos procedieron a crear un modelo de Red Neuronal, que se encargó de realizar el proceso de clasificación y finalmente utilizaron las Curvas ROA para evaluar la precisión del modelo, el cual alcanzo resultados de entre un 75% y 80%, resultados muy similares a los alcanzados en trabajos hechos en el idioma inglés, ya que estas herramientas tienen un mejor rendimiento en el idioma inglés que para otros idiomas [56].

• Siguiendo en el contexto colombiano, destaca el estudio realizado por Julián David Longas Arteaga (2022) en la Universidad de Antioquia, titulado "Estudio e implementación de un modelo de procesamiento de lenguaje natural que analice la satisfacción de compra mediante el análisis de comentarios de usuarios" Este trabajo implementa técnicas de procesamiento de lenguaje natural (NLP) para evaluar la satisfacción del cliente a partir de comentarios textuales. El autor empleó el modelo BERT (Bidirectional Encoder Representations from Transformers), adaptado al español y entrenado con comentarios reales de clientes colombianos, lo que permitió una comprensión más precisa del lenguaje coloquial utilizado por los consumidores. El proceso seguido incluyó etapas de preprocesamiento de datos, entrenamiento del modelo

y evaluación de su desempeño. Los resultados obtenidos demostraron una alta precisión en la clasificación de sentimientos, alcanzando niveles comparables a estudios realizados en otros idiomas [56].

5. METODOLOGÍA

Con el fin de alcanzar los objetivos al inicio de la investigación, se establecieron etapas metodológicas para desarrollar el proyecto planteado en fases que se describen a continuación.

5.1 SELECCIÓN DE TEXTOS RELEVANTES:

- Para realizar una búsqueda efectiva de la información, se establecieron los siguientes criterios de inclusión mencionados a continuación.
- Criterios de inclusión: Noticias Periodísticas publicadas entre los años 2022 – 2024, Noticias periodísticas publicadas en Colombia, Noticias periodísticas relacionadas con la reforma a la salud actual y no reformas previas al año 2020, Noticias publicadas donde su difusor sea de las regiones Andina, Caribe y pacífica, Noticias publicadas en español, Noticias publicadas en fuentes confiables y pertenecientes a la clasificación SCImago [57].
- Se utilizaron palabras clave como Reforma, Salud, Colombia y regiones geográficas específicas como la Región Andina, Caribe y Pacífica, esto con el fin de identificar artículos periodísticos de interés que nutrieron el proyecto investigativo desarrollado.
- Fuentes de noticias confiables: Se realizó una indagación sobre los portales web y las publicaciones de los medios de comunicación más destacados de cada región para asegurar que la información a utilizar provenga de fuentes fiables; se tuvo en cuenta la clasificación de SCImago Journal & Country Rank (SJR), plataformas que proporcionan indicadores sobre el impacto y la calidad de las publicaciones periodísticas [57].
- La selección de fuentes relevantes es esencial para mitigar el sesgo informativo y asegurar que la cobertura analizada tenga un enfoque balanceado y representativo de las distintas regiones del país.

5.2 RECOLECCIÓN DE LOS DATOS:

- Se llevó a cabo un análisis exploratorio para evaluar la viabilidad de realizar web scraping en los sitios seleccionados. Esta evaluación se centró en identificar posibles medidas anti scraping y la estructura interna del HTML de las páginas web. Se utilizó Python junto con herramientas como BeautifulSoup para explorar los sitios, comprobar la robustez de sus estructuras HTML, y determinar si se podían automatizar las tareas de recolección sin contratiempos [58]. Este análisis fue determinante para asegurar la viabilidad técnica y legal de esta tarea, ya que se trabajó exclusivamente con información de acceso público, la cual no requiere consentimiento para su recopilación ni procesamiento. Además, todos los datos recolectados fueron utilizados

con fines estrictamente académicos, sin intención de explotación comercial ni vulneración de derechos individuales.

- Una vez recolectados los hipervínculos de las noticias, se compiló una lista de cada enlace en un archivo titulado "link-noticias-2024.xlsx". Esta actividad fue fundamental debido a que permitió centralizar todos los enlaces relevantes en un único repositorio facilitando el acceso automatizado a los mismos durante la fase de extracción.
- Con la lista de enlaces compilada, se procedió con la extracción de datos utilizando técnicas de web scraping. Posteriormente, se creó una base de datos que almacenó la fecha de publicación, la URL del sitio web, la región geográfica (Andina, Caribe y Pacífico) y el texto de la noticia con el cual se realizó el análisis.
- La elección de web scraping se basó en la necesidad de manejar grandes volúmenes de datos de manera automatizada y eficiente, evitando errores humanos y asegurando una cobertura actualizada y representativa de la reforma a la salud en Colombia.

5.3 PREPROCESAMIENTO DE DATOS:

- Posterior a la recolección de los datos, se identificaron 1401 noticias relacionadas con la reforma a la salud en Colombia. Continuando con las fases del estudio, se realizó la limpieza de los datos utilizando scripts desarrollados en Python denominados limpiar texto y preprocesar texto.
- La función limpiar texto aplica una limpieza básica, mientras que preprocesar texto, realiza un preprocesamiento que incluye la eliminación de stopwords, la tokenización y la lematización.
- Estos Scripts tienen como función la eliminación del ruido y elementos irrelevantes como, signos de puntuación, emojis, correos electrónicos, tildes, frases publicitarias y espacios redundantes, además de caracteres especiales como $ # ” % & / ( ). Además, se estandarizó el formato textual mediante la conversión a minúsculas y la normalización de palabras acentuadas, garantizando que el texto resultante fuera claro, uniforme y relevante para el desarrollo de la investigación en curso; se eliminaron stopwords, donde se removieron palabras comunes que no aportan valor agregado al análisis (por ejemplo, "la", "de", "a").
- Al realizar la Tokenización, se segmentó el texto en unidades básicas como palabras o frases y se finalizó el proceso realizando lematización, reduciendo las palabras a su forma base para unificar términos similares (por ejemplo, "caminando", "caminé" a "caminar").
- Cada una de estas fases permitió extraer información clave del contenido de las noticias con el propósito de facilitar el análisis posterior, obteniendo un resultado de 1401 noticias.

5.4 ANÁLISIS ESTADÍSTICO

Para llevar a cabo la ejecución de los modelos, fue necesario realizar previamente un análisis de datos que consistió en cargar la información obtenida del preprocesamiento en un script de Python llamado Análisis estadístico (Anexo 2). Este Script tuvo como fin desglosar la cantidad de palabras (tokens) de cada noticia y así realizar una estadística descriptiva; posteriormente, se identificó que la métrica del rango intercuartílico superior era igual a 1352, lo que evidenciaba que 119 noticias se etiquetaban como atípicas, esto debido a que la cantidad de tokens en estas superaba en gran número a la media de los tokens de las demás noticias obtenidas, se determinó no tener en cuenta las 119 noticias atípicas para evitar el sesgo en la información en la muestra objeto del estudio.

5.5 COMPARACIÓN DE SIMILITUD

Para analizar la similitud entre regiones y noticias, se implementaron varios enfoques, el primero de ellos fue basado en el modelo **TF-IDF** (Term Frequency-Inverse Document Frequency), complementado con las métricas de similitud de coseno y distancia euclidiana. Las noticias se agruparon por región, combinando los contenidos de cada región en un solo texto, lo que permitió analizar la similitud entre regiones en lugar de entre noticias individuales. Utilizando el modelo TF-IDF, los textos combinados se transformaron en vectores numéricos, asignando pesos a las palabras según su frecuencia en el documento y su relevancia en el corpus. Posteriormente, se calculó la similitud de coseno entre las representaciones TF-IDF de las regiones, obteniendo valores entre 0 (sin similitud) y 1 (similitud total), y se midió la distancia euclidiana entre los vectores TF-IDF, donde valores más bajos indican mayor similitud. Además, para cada región, se calculó la similitud de coseno y la distancia euclidiana entre las noticias individuales, evaluando la homogeneidad de los contenidos dentro de cada región.

Como segundo enfoque se utilizó el modelo **Doc2Vec** (Anexo 4), en el cual al inicio se realizó el cargue de documentos convirtiéndolos a embeddings para capturar relaciones semánticas entre los elementos cargadas, posteriormente se utilizó una medida de similaridad de coseno para determinar qué tan lejanos o cercanos eran estos vectores; luego se realizó un análisis en las regiones y comparación entre las mismas. Adicionalmente, se implementó una técnica de análisis con hiperparámetros de rejilla para optimizar el modelo de Doc2Vec y con base en los mejores resultados de la comparación de coseno, se decidió utilizar los valores con mejor combinación de parámetros:

vector_size	window	min_count	epochs	promedio_similitud
200	2	1	100	0.908063

Tabla 1. Hiperparámetros optimizados modelo Doc2Vec

Un tercer enfoque fue el modelo **BERT**(en español) (Anexo 5), en el cual se utilizaron las librerías fastembed y faiss para crear los embeddings y comparar los vectores generados,

también fue necesario realizar una revisión para conocer que modelos eran soportados por fastembed escogiendo el modelo "*sentence-transformers/paraphrase-multilingual-mpnet-base-v2*" por características como el multilingüismo [59], la paráfrasis que permite capturar la semántica de las palabras y eficiencia a la hora de generar embeddings de manera más eficiente y rápida [60]; Por otro lado, fue necesario dividir las noticias en fragmentos de 512 tokens debido a las limitaciones inherentes a este tipo de modelos. Se llevaron a cabo dos análisis, el primero comparó todas las particiones entre sí, mientras que el segundo organizó las comparaciones de manera secuencial emparejando el primer fragmento con el primer fragmento de otra noticia, el segundo con el segundo, y así sucesivamente.
Estas técnicas permitieron evaluar las relaciones y diferencias entre las coberturas mediáticas regionales de la reforma a la salud, proporcionando una visión detallada de cómo varía la percepción y el enfoque en distintas áreas del país.

5.6 ETIQUETADO DE LAS NOTICIAS

Con el objetivo de establecer métricas confiables para la posterior evaluación de los modelos de análisis de sentimientos, se llevamos a cabo un proceso de etiquetado manual del 26,43% de las noticias. Para ello, se diseñó una rúbrica compuesta por cinco ítems que permitieron clasificar los textos según distintos criterios asociados al contenido emocional y argumentativo. Esta rúbrica evaluaba las noticias con base en las siguientes dimensiones:

- **Uso de Lenguaje Emocional**: Valoración de la presencia de expresiones con carga subjetiva.
- **Tono y Estilo Periodístico**: Identificación del enfoque narrativo empleado.
- **Contextualización y Evidencia**: Análisis del respaldo mediante datos, citas o fuentes verificables.
- **Evaluación del Ponderado del Léxico**: Determinación de la polaridad general del vocabulario utilizado.
- **Consistencia y Coherencia**: Revisión de la estructura argumentativa y su uniformidad interna.

La regla de decisión definida estableció que el sentimiento predominante sería aquel que apareciera en al menos tres de los cinco ítems evaluados, permitiendo así determinar de manera sistemática si una noticia era clasificada como positiva, negativa o neutral.

5.7 ANÁLISIS DE SENTIMIENTOS

Este análisis se llevó a cabo realizando la implementación de dos modelos para análisis de sentimientos. se escogió el modelo *"finiteautomata/beto-sentiment-analysis"* (Anexo 6) por sus características óptimas para analizar sentimientos en español debido a su entrenamiento especializado en este idioma. A diferencia de modelos generalistas como BERT o XLM-Roberta, BETO ha sido pre entrenado con un amplio corpus en español, lo que le permite capturar mejor

las estructuras lingüísticas y expresiones idiomáticas propias del idioma [61]. Además, ha sido ajustado específicamente con el corpus TASS 2020, que contiene miles de tweets etiquetados manualmente, optimizándolo para analizar textos informales con abreviaciones, emoticones y expresiones coloquiales [42].

Su capacidad de clasificación en tres categorías (positivo, negativo y neutral) proporciona un análisis más detallado del sentimiento del texto en comparación con modelos binarios. Además, al estar disponible en la plataforma Hugging Face, facilita su implementación mediante Transformers y PyTorch, reduciendo la necesidad de entrenar un modelo desde cero [23]. Adicional fue necesario dividir las noticias en fragmentos de 500 tokens debido a las limitaciones de este modelo [23]. A continuación, cada fragmento fue procesado por el *pipeline*, recopilándose las etiquetas (POS, NEU, NEG) y sus probabilidades asociadas. Sobre este conjunto de predicciones parciales se calculó un sentimiento predominante (vía conteo de etiquetas) y una puntuación media (promedio de las probabilidades de clasificación) por noticia. Finalmente, se determinó una etiqueta de sentimiento final para una clasificación uniforme en todo el corpus.

EL segundo modelo fue "pysentimiento/robertuito-sentiment-analysis", basado en la arquitectura RoBERTa y afinado específicamente para textos en español. A diferencia de variantes genéricas de RoBERTa pre-entrenadas en corpus multilingües, RoBERTa ha sido ajustado con extensos conjuntos de datos en español y construcciones sintácticas propias del idioma [39]. Asimismo, su entrenamiento abarca tanto lenguaje formal como registros informales, optimizándolo para el análisis de contenido periodístico y de usuario sin pérdida de fidelidad semántica.

Al igual que con el primer modelo implementado BETO, se empleó una clasificación en tres categorías (positivo, negativo y neutral), permitiendo un desglose más fino de las opiniones expresadas en cada noticia. Dado que el límite de RoBERTa es de 512 tokens por entrada, los textos se fragmentaron en segmentos de hasta 500 tokens antes de su inferencia, asegurando así la integridad de la información y evitando truncamientos indebidos [40]. A continuación, cada fragmento fue procesado por el pipeline, registrando tanto la etiqueta de sentimiento (POS, NEU, NEG) como la probabilidad de la predicción. Sobre este conjunto de inferencias parciales se calculó el sentimiento predominante mediante un conteo de etiquetas y la puntuación media como promedio de las probabilidades asignadas [62] [63].

Por último, se incluyó un modelo de ChatGPT 4o, basado en la arquitectura de transformers con un decoder unidireccional, utilizando un proceso asíncrono a través de la API de OpenAI. Su implementación permitió realizar un análisis adicional de sentimiento en noticias sobre la reforma a la salud en Colombia. Para ello, se diseñó un prompt que instruyó al modelo a clasificar los textos en tres categorías: positivo, negativo y neutro.
Dado que ChatGPT 4o admite hasta 4096 tokens por solicitud, se configuró para procesar las noticias en su totalidad, evitando la fragmentación de los textos. En este estudio, la longitud

promedio de los documentos era de 646 tokens, lo que contrastaba con la capacidad de modelos como BETO, limitado a 512 tokens, que requería dividir los textos antes de analizarlos.

A diferencia de los modelos basados en encoder, como BETO, que procesan el texto en ambas direcciones para capturar el significado completo de una secuencia, los modelos con decoder, como ChatGPT, generan texto de manera secuencial, palabra por palabra, en un solo sentido. Su uso en este análisis se planteó como un enfoque complementario, explorando su desempeño en la clasificación de sentimiento mediante un procesamiento asíncrono de los datos.

5.8 EVALUACIÓN DE LOS MODELOS

El etiquetado manual permitió obtener un conjunto de datos de referencia para calcular las métricas de desempeño de los modelos BETO y RoBERTa, como la precisión, recall y F1-score. Además, estos datos fueron utilizados para ajustar los hiperparámetros de los modelos en el proceso de fine-tuning, con el objetivo de mejorar su capacidad de clasificación de sentimientos en textos sobre la reforma a la salud. La rúbrica detallada utilizada en este proceso se presenta en los anexos.

Como se mencionó en el párrafo anterior, luego de tener las métricas de los modelos BETO y RoBERTa, se realizó el fine-tuning con el fin de adaptar los modelos de lenguaje general al contexto específico de nuestra investigación de análisis de sentimiento de noticias sobre la reforma a la salud en Colombia. Aunque estos están ajustados al modelo español, no están entrenados en el estilo periodístico ni en el vocabulario o matices propios de los medios de comunicación locales. Al ajustar el modelo utilizando el conjunto de noticias etiquetado cuidadosamente según la rúbrica mencionada, se equilibró además la cantidad de noticias por sentimiento, ya que inicialmente se habían etiquetado más noticias de un sentimiento específico, lo que tendía a afectar negativamente los resultados. Gracias a este proceso, logramos que el modelo aprendiera expresiones, matices y tonalidades relevantes para clasificar correctamente el sentimiento. Se partió del modelo base y se añadió una capa de salida para clasificar en positivo, neutro y negativo. Luego se entrenó con un conjunto de datos etiquetado manualmente, utilizando una división del 85% para entrenamiento y el 15% para prueba, lo que garantizó una evaluación equilibrada. Durante el proceso de fine-tuning, se empleó un tamaño de batch de 1, una tasa de aprendizaje de 2e-5 y se entrenó durante 4 épocas, con la implementación de la técnica de *Early Stopping* para prevenir el sobreajuste, permitiendo interrumpir el entrenamiento si no se observaba mejora en el F1 score tras dos épocas. Esto contribuyó a optimizar tanto el tiempo de cómputo como la calidad del modelo, asegurando una buena generalización.

El rendimiento del modelo fue evaluado principalmente con el F1 score, lo que permitió obtener una medida balanceada entre precisión y recall en las clases relevantes para el análisis de las percepciones mediáticas sobre la reforma a la salud en Colombia. Tras completar el

entrenamiento, se generaron métricas adicionales como la matriz de confusión, lo que facilitó una comprensión más profunda de las fortalezas y debilidades del modelo, especialmente en cuanto a la clasificación de los diferentes tonos y enfoques de las noticias relacionadas con la reforma.

A continuación, se presenta un esquema que resume la metodología utilizada en el desarrollo del proyecto.

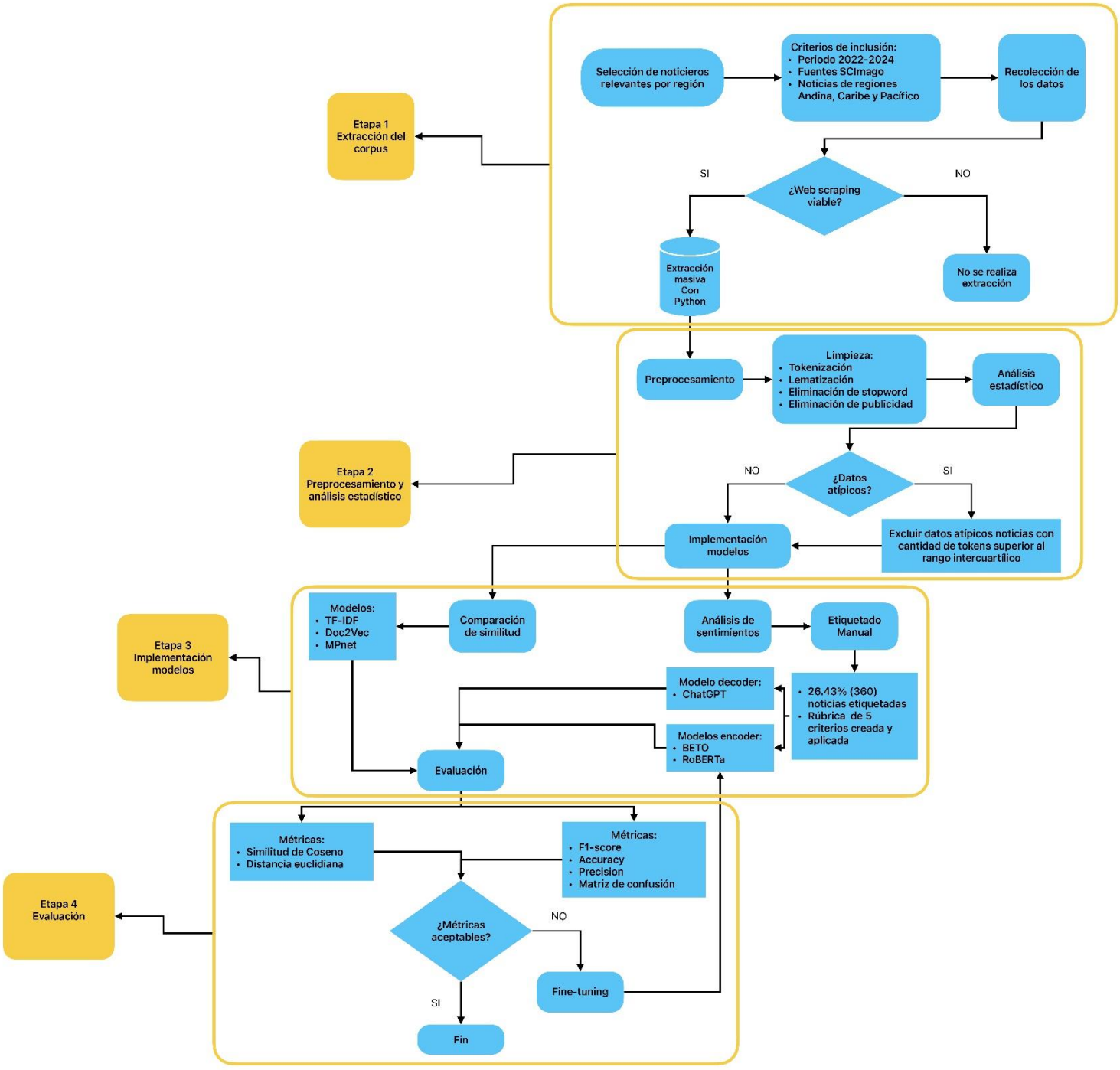

Figura 2. Esquema que resume la metodología en este proyecto

6. ANÁLISIS ESTADÍSTICO DESCRIPTIVO

El análisis descriptivo tuvo el propósito de explorar, resumir y comprender las características principales del conjunto de datos. Haciendo uso de visualizaciones como histogramas, gráficos de barras y nubes de palabras. Este proceso permite identificar patrones, detectar valores atípicos y comprender mejor la estructura de los datos, sentando las bases para los análisis posteriores con modelos avanzados de NPL. ´

6.1 DISTRIBUCIÓN DE TOKENS POR NOTICIAS

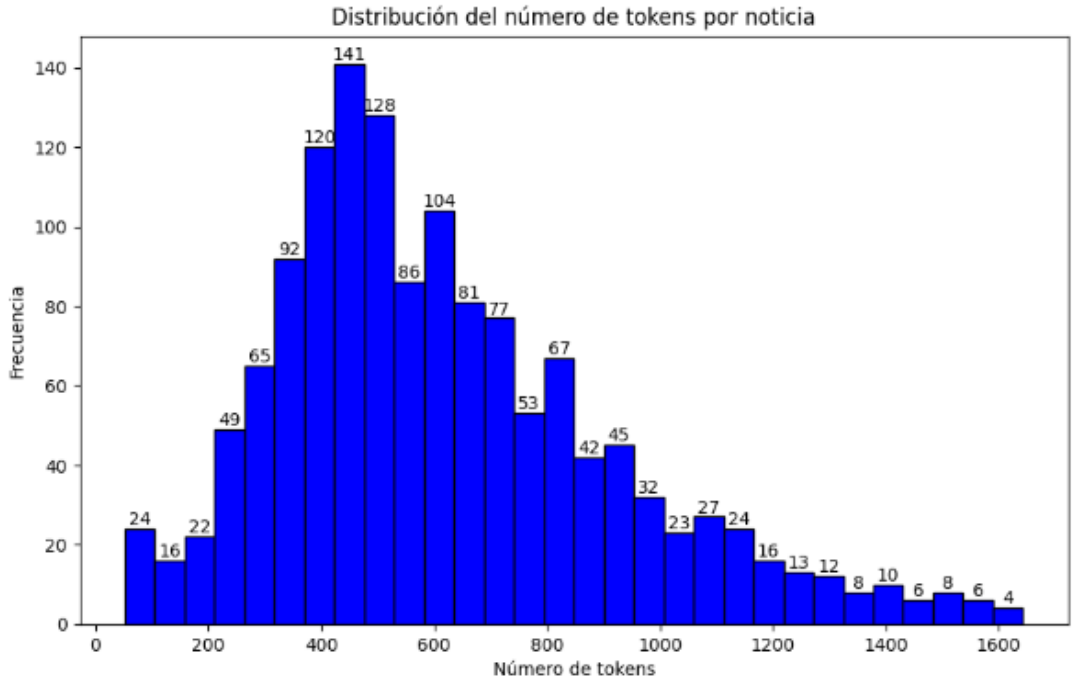

Figura 3. Distribución del número de tokens

La Figura 3. muestra la distribución del número de tokens por noticia. Se observa que la mayoría de las noticias presentan entre 300 y 700 tokens, con un pico máximo en el intervalo de 400 a 500 tokens, donde se concentran 141 noticias. Este rango central abarca una gran proporción del total de noticias, indicando una tendencia hacia textos de longitud moderada. A medida que el número de tokens aumenta, la frecuencia comienza a disminuir gradualmente. No obstante, aún se encuentran cantidades representativas de noticias con longitudes entre 800 y 1000 tokens. Más allá de este umbral, la frecuencia decae de forma más pronunciada, y los casos de noticias con más de 1200 tokens se vuelven mucho más esporádicos. A partir de los 1400 tokens, las frecuencias son notablemente bajas, con apenas unos pocos casos aislados, esta distribución presenta una asimetría positiva típica en colecciones de texto, donde la mayoría de los documentos son relativamente breves, pero existe una cola larga de textos más extensos.

6.2 DISTRIBUCIÓN DE TOKENS POR NOTICIAS Y SEGÚN LA REGIÓN

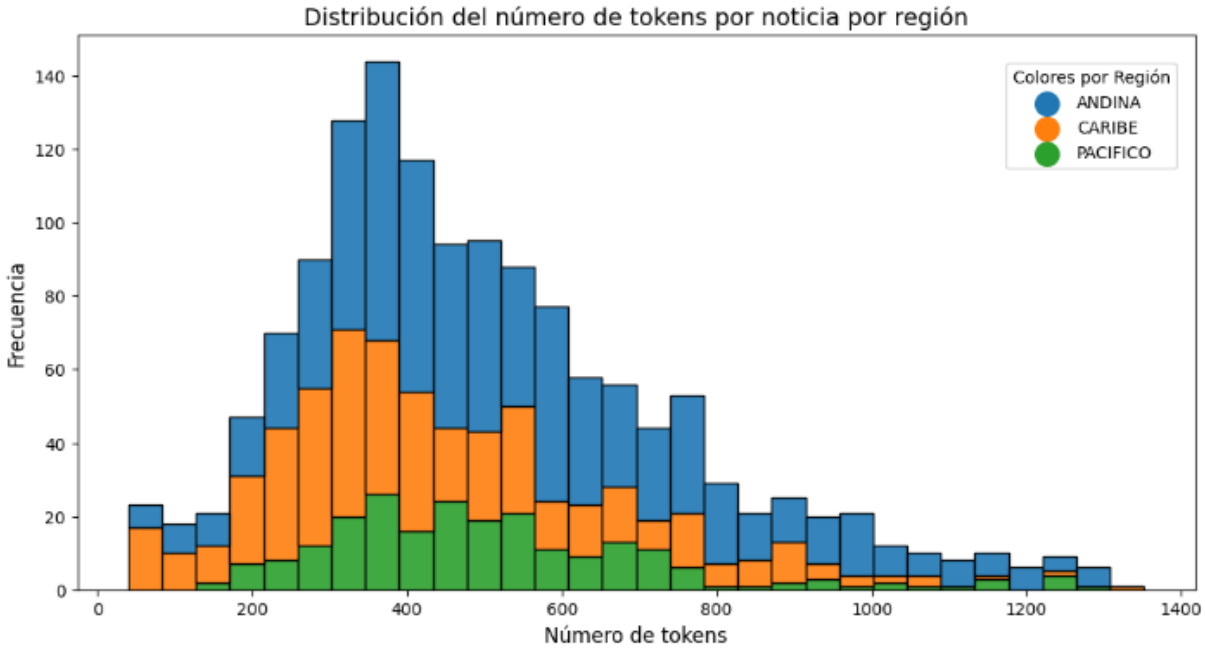

Figura 4. Distribución de tokens por región

El análisis estadístico de tokens por región que aparece en la Figura 4. revela una distribución asimétrica en la cobertura mediática de la reforma a la salud en Colombia, con una mayor concentración de noticias de corta extensión (menos de 800 tokens) y una disminución progresiva en la frecuencia de textos más largos. La región Caribe presenta el mayor volumen de noticias en los rangos más bajos de tokens, mientras que la región Andina muestra una distribución más dispersa con una mayor presencia en valores intermedios, sugiriendo una cobertura más detallada o segmentada. La región Pacífico, aunque con menor representación,

sigue un patrón similar a las demás, pero con una menor frecuencia en general.

6.3 ANÁLISIS DE VALORES ATÍPICOS EN EL NÚMERO DE TOKENS POR NOTICIA

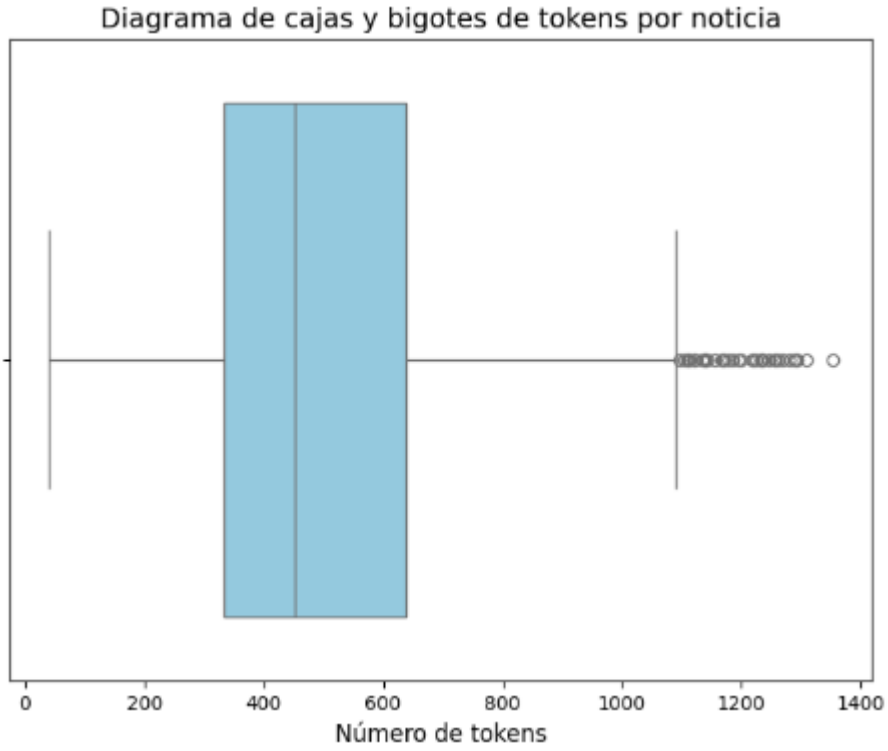

Figura 5. Diagrama de cajas y bigotes de tokens

El análisis estadístico de los tokens por noticia que se observa en la Figura 5. muestra una media de 502 tokens, lo que representa la longitud promedio de las noticias. La mediana es de 451 tokens, indicando que la mitad de las noticias tienen menos de ese número. La desviación estándar es de 244.5 tokens, lo que refleja una variabilidad moderada en la longitud de los textos, con una diferencia significativa entre el mínimo de 40 tokens y el máximo de 1352 tokens. El rango intercuartílico (IQR) es de 305 tokens, con $Q_3 = 639$ y $Q_1 = 334$, lo que permite identificar valores atípicos. Se consideran atípicas las noticias que superan los 1096 tokens, con un total de 39 noticias atípicas, representando aproximadamente el 2.8% del total. Estas noticias alcanzan hasta 1352 tokens, muy por encima de la media. La mayoría de las noticias se encuentran dentro del rango de 334 a 639 tokens.

A partir del análisis estadístico anterior, se decidió eliminar las noticias atípicas en cuanto al número de tokens. Esta decisión se tomó con el objetivo de mejorar el entrenamiento de los modelos, ya que las noticias con un número de tokens excesivamente alto pueden generar problemas debido a las limitaciones en la cantidad máxima de tokens que ciertos modelos pueden procesar. Sin embargo, esta implementación se llevará a cabo más adelante, ya que el

proceso de modelado incluirá varios enfoques, los cuales serán evaluados en el análisis y entrenamiento posterior.

6.4 ANÁLISIS DE CANTIDAD DE NOTICIAS POR REGIÓN

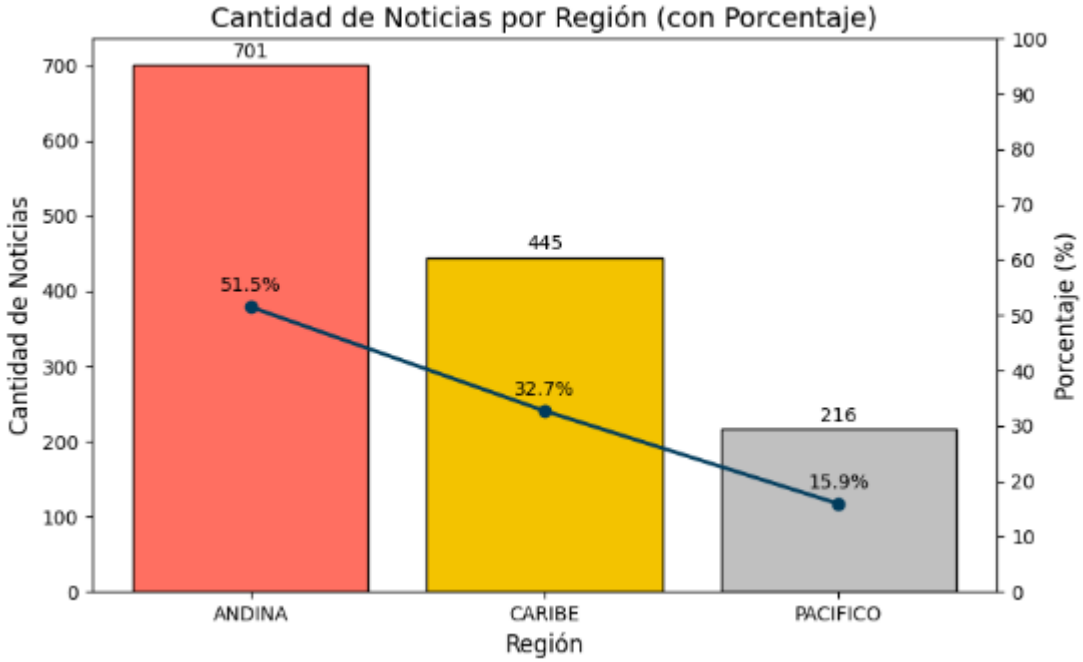

Figura 6. Numero de noticias por región

En la figura 6. se observa que la región Andina concentra la mayor cantidad de noticias, con un total de 701 noticias, lo que representa un 51.5% del total. La región Caribe ocupa el segundo lugar, con 445 noticias, equivalente al 32.7% del total. Finalmente, la región Pacífico cuenta con 216 noticias, representando un 15.9% del total analizado.

Estos resultados muestran una mayor concentración de cobertura mediática en la región Andina, lo cual podría estar relacionado con su alta densidad poblacional y su importancia política, social y económica en el contexto nacional. Esta disparidad en la cantidad de noticias entre regiones podría explicarse por diversos factores, como la mayor presencia de medios de comunicación en centros urbanos de la región Andina, el nivel de relevancia que los eventos relacionados con la reforma a la salud han tenido en estas áreas y la mayor capacidad de estas zonas para generar noticias debido a su población e infraestructura. Un análisis más detallado permitirá evaluar en qué medida estas variables contribuyen a las diferencias observadas en la cobertura regional.

6.4.1 Análisis de palabras más frecuentes por región

Siguiendo con el análisis estadístico, se identificaron las palabras más frecuentes en el contenido de noticias relacionadas con la reforma a la salud en las regiones Andina, Caribe y Pacífico. Para ello, se calcularon las frecuencias absolutas y relativas de aparición de las

palabras, las cuales se presentan gráficamente y se describen en este informe. Este análisis busca resaltar similitudes y diferencias regionales, así como identificar los temas predominantes en el discurso mediático.

6.4.2 Análisis a Nivel General

Figura 7. Palabras más frecuentes en todas las 3 Regiones.

Analizando la Figura 7. se muestra una nube con las 30 palabras más frecuentes entre las tres regiones y entre esas se destacan las palabras "salud", "reforma" y "sistema" como las más frecuentes, sub- rayando la relevancia de la reforma a la salud como eje central en el discurso mediático. Términos como "gobierno" y "proyecto" refuerzan la conexión con las políticas públicas, mientras que las diferencias regionales aportan matices interesantes. La región Andina muestra un enfoque más operativo del sistema de salud, con términos como "eps" y "atención", mientras que en la región Caribe aparecen palabras como "ministro" y "nuevo", reflejando un interés en cambios recientes. Por su parte, la región Pacífico resalta con términos como "debate" y "partido", que apuntan a una mayor atención en las dinámicas políticas y legislativas. Estas variaciones regionales, combinadas con los términos más frecuentes, evidencian los temas prioritarios en la agenda mediática, influenciados por factores políticos, sociales y económicos específicos de cada zona.

6.4.3 Analisis de palabras más frecuentes por región

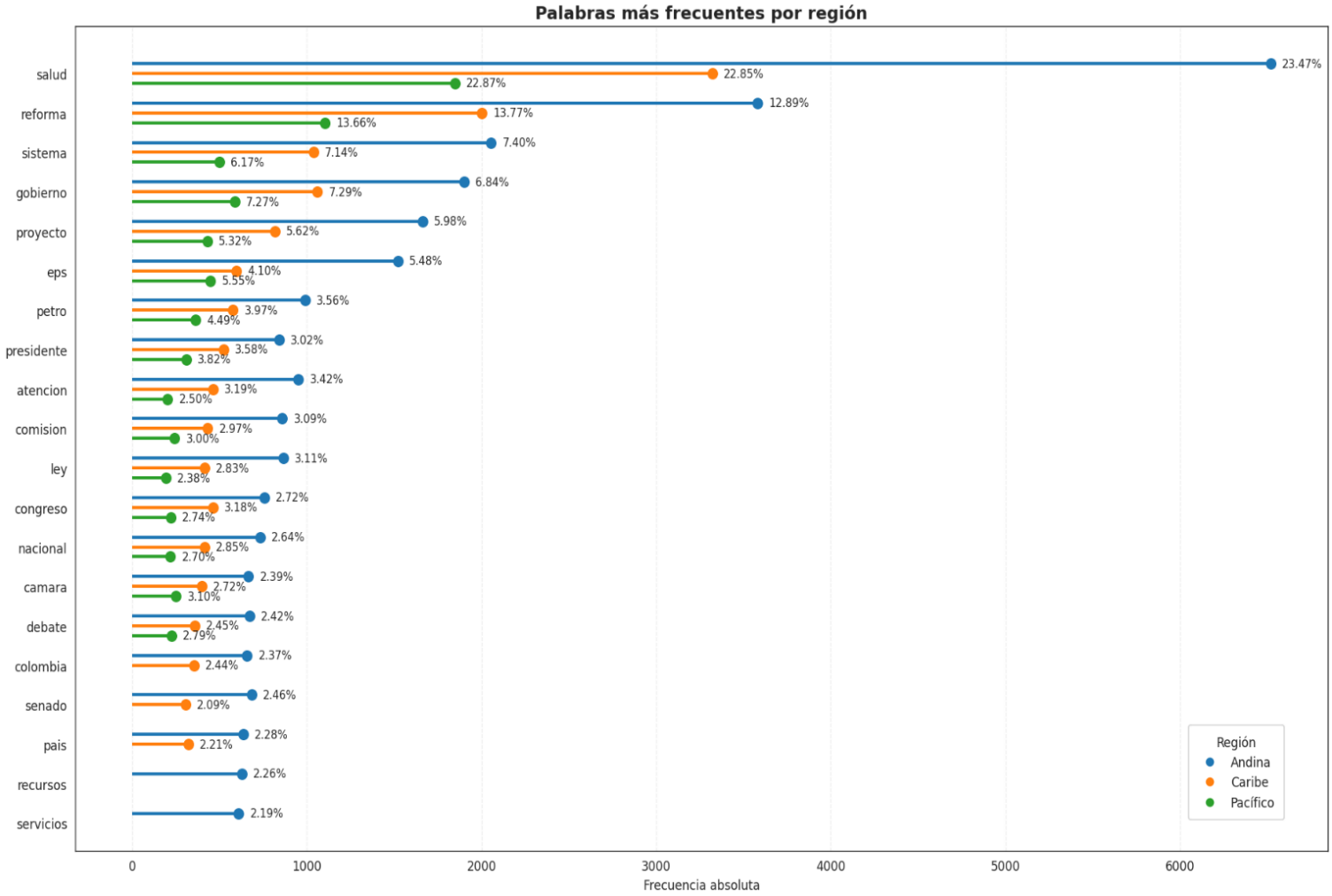

Figura 8. Cantidad de palabras y frecuencia en las 3 Regiones

Como se observa en la Figura 8. Se muestra de manera comparativa tanto las coincidencias como las diferencias en el discurso mediático sobre la reforma a la salud entre las regiones Andina, Caribe y Pacífico. En las tres regiones, las palabras “salud” y “reforma” ocupan los primeros lugares en frecuencia, lo que confirma que la reforma al sistema de salud constituye el eje central del debate público a nivel nacional. Sin embargo, se identifican diferencias relevantes en la intensidad de dicha cobertura: la región Andina presenta la mayor cantidad de menciones, con 6.520 apariciones del término “salud” (23,47%) y 3.580 de “reforma” (12,89%), seguida por la región Caribe con 3.322 (22,85%) y 2.001 (13,77%), y la región Pacífico con 1.846 (22,87%) y 1.102 (13,66%). Esta diferencia en volumen puede interpretarse como una mayor densidad mediática o una cobertura más amplia del tema en la región Andina.

Asimismo, el gráfico permite identificar matices temáticos particulares en cada región. En la región Andina, los términos “eps”, “proyecto”, “atención” y “gobierno” presentan una mayor presencia, lo cual sugiere un enfoque centrado en la implementación operativa de la reforma y en los actores que gestionan el sistema. En contraste, la región Caribe refleja una narrativa

más orientada a la dinámica legislativa y partidista, evidenciada en la relevancia de palabras como “ministro”, “partido”, “cámara” y “congreso”. Por su parte, la región Pacífico se caracteriza por una cobertura más orientada a las discusiones parlamentarias, destacándose términos como “representantes”, “ponencia” y “debate”, aunque con un volumen absoluto menor. Esta diferencia puede deberse a una menor cantidad de noticias analizadas o a una menor visibilidad mediática de la región en los medios de comunicación nacionales.

En general, los resultados muestran que, si bien existe una agenda común centrada en la reforma a la salud, cada región construye una narrativa con énfasis diferenciados: la región Andina resalta los aspectos técnicos y operativos; la Caribe, la confrontación política y legislativa; y la Pacífico, las deliberaciones parlamentarias. Finalmente, el hecho de que las veinte palabras más frecuentes concentren más de una quinta parte del total de palabras analizadas en cada región reafirma la alta relevancia de estos términos en la narrativa mediática sobre la reforma.

6.5 ANÁLISIS ESTADÍSTICO POR EDITORIAL

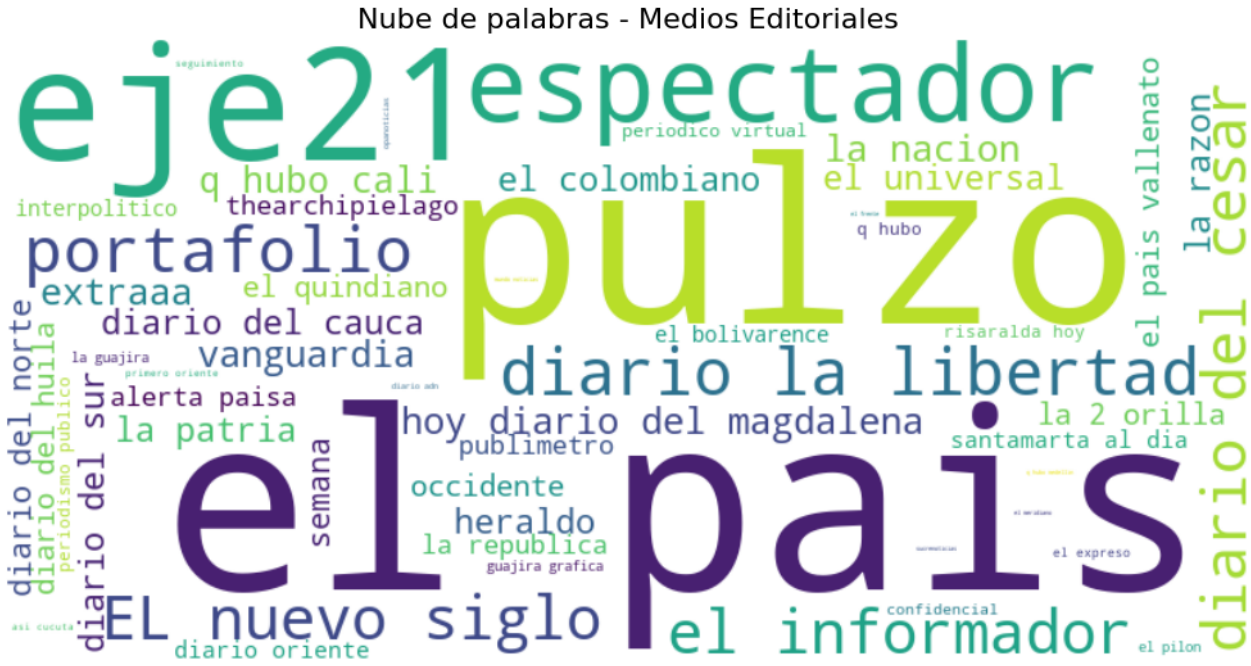

Figura 9. Contribución de los portales de noticias

En la Figura 9. Se muestra el análisis de los 55 portales de noticias relacionados con la reforma a la salud en Colombia, se utilizó una nube de palabras para identificar los medios más relevantes en cuanto a su cobertura del tema. La distribución de los medios de comunicación en esta nube refleja la frecuencia con la que aparecieron en los artículos sobre la reforma, permitiendo observar cuáles son los más prominentes en la difusión de este tema. Entre los portales que más se destacaron se encuentran (El País, Pulzo, Eje21, El Espectador, El Informador, La Nación), los cuales tienen una mayor visibilidad debido a su constante

presencia en las publicaciones.

6.5.1 Editoriales más frecuentes región Andina

Figura 10. Numero de noticias por editorial en la región Andina

En el análisis de la región "ANDINA", se observa en la Figura 10. una clara distribución de las noticias por editorial, destacándose El Espectador, Pulzo, Eje21, Portafolio y El Nuevo Siglo como los principales medios con mayor cobertura sobre la reforma a la salud. El Espectador lidera la lista con casi 70 noticias, seguido de cerca por Pulzo y Eje21. Estos noticieros tienen una presencia significativa en la discusión sobre la reforma, lo que refleja su papel protagónico en la difusión del tema. A medida que se desciende en la lista, se nota una disminución en la cantidad de noticias, con medios como el frente, 180 grados y Q Hubo Medellín que muestran una cobertura más limitada, lo que sugiere una menor implicación en la cobertura del tema.

La Figura adjunta ilustra esta distribución, evidenciando que la mayoría de los noticieros en esta región se encuentran en el rango de 10 a 40 noticias, con un número considerable de medios moderadamente comprometidos con la reforma. Este patrón refleja una participación variada, con un pequeño grupo de medios dominando la cobertura y una mayor cantidad de portales que contribuyen con una menor frecuencia de publicaciones.

6.5.2 Editoriales más frecuentes región Caribe

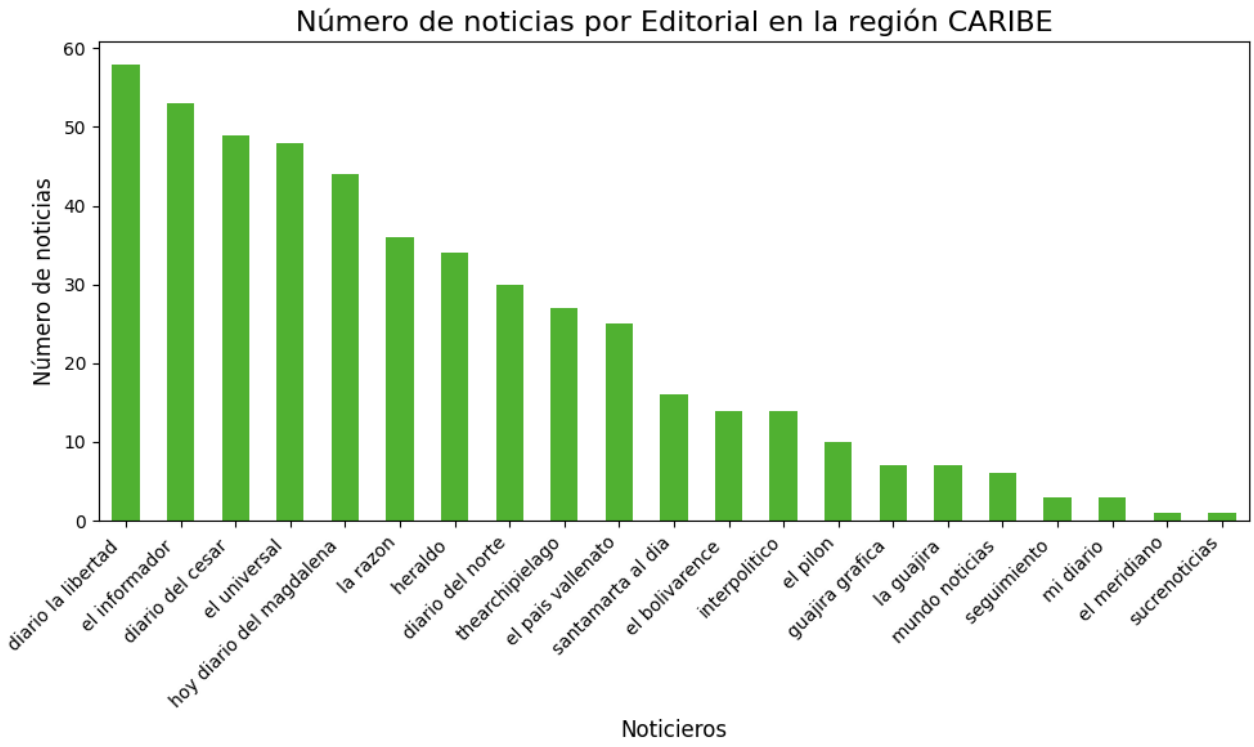

Figura 11. Numero de noticias por editorial en la región Caribe

Para la región Caribe, se observa en la Figura 11. una distribución de noticias por noticiero similar a la de la región Andina, aunque con algunas diferencias en los medios predominantes. Diario La Libertad lidera la cobertura en esta región con un número superior a 50 noticias, seguido de El Informador y Diario del Cesar, que también presentan un volumen significativo de publicaciones. En general, los medios en esta región tienen una cobertura más equilibrada, con varios noticieros que publican un número moderado de noticias sobre la reforma a la salud, como Hoy Diario del Magdalena, La Razón y Diario del Norte. A medida que se avanza en la lista, los medios con menor frecuencia de cobertura incluyen, La Guajira, El Meridiano y Sucrenoticias, que muestran una participación más reducida en la discusión de este tema.

Esta Figura refleja una mayor dispersión en la cobertura, con un grupo pequeño de noticieros que abarcan la mayor parte de las publicaciones, mientras que los demás medios tienen una participación más baja. Esto sugiere que en la región Caribe, aunque hay un interés por la reforma a la salud, la cobertura no está tan concentrada como en la región ANDINA, sino que se distribuye entre varios medios locales y regionales.

6.5.3 Editoriales más frecuentes región Pacifico

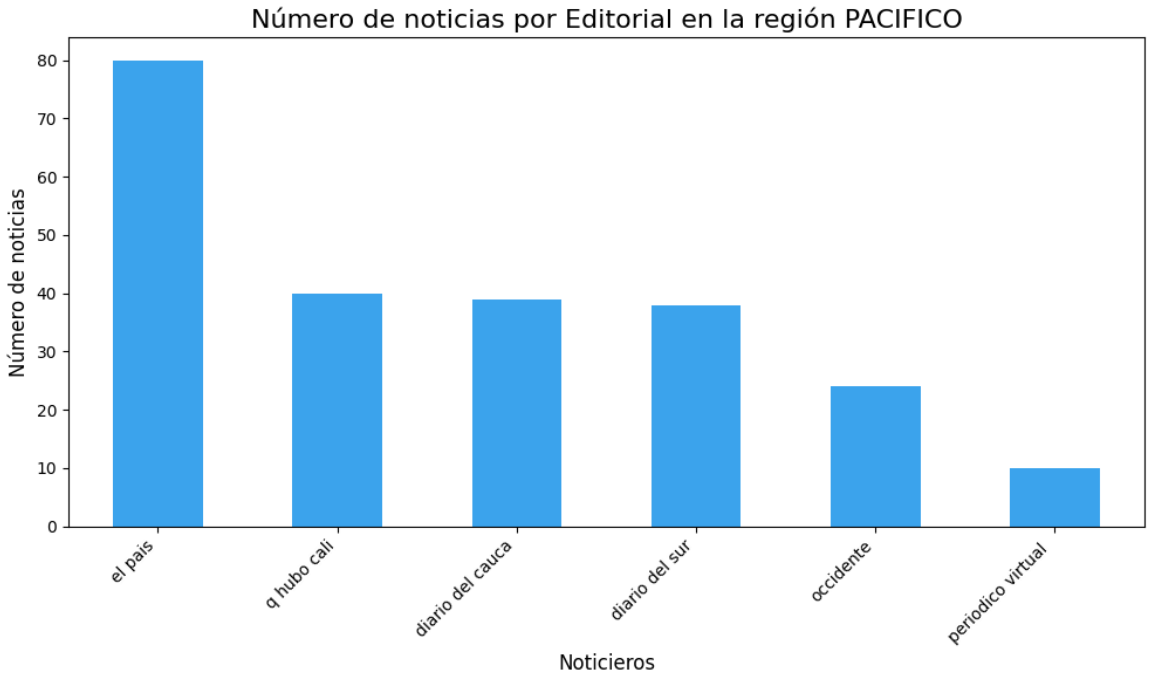

Figura 12. Numero de noticias por editorial en la región Pacifico

Para la región Pacífico, se observa en la Figura 12. una concentración de noticias en El País, que lidera la cobertura con un número claramente superior a los demás noticieros, con más de 70 publicaciones relacionadas con la reforma a la salud. Este medio se destaca enormemente en comparación con los otros medios en esta región. Los siguientes medios con mayor cobertura son Q Hubo Cali, Diario del Cauca, Diario del Sur y Occidente, con un número considerable de noticias, pero claramente inferiores al volumen reportado por El País. Finalmente, medios como Periódico Virtual tienen una participación mínima en cuanto a publicaciones, reflejando una menor implicación en la cobertura del tema.

Este gráfico demuestra que, en la región Pacífico, existe una fuerte concentración de la cobertura en El País, mientras que otros medios tienen una presencia moderada o limitada. Esto sugiere que, al igual que en otras regiones, los medios con mayor alcance tienen un papel predominante en la difusión de información sobre la reforma a la salud, mientras que los medios más pequeños tienen una influencia mucho menor.

7. ANALISIS DE RESULTADOS

7.1 SIMILITUD DE DOCUMENTOS

7.1.1 Modelo TF-IDF

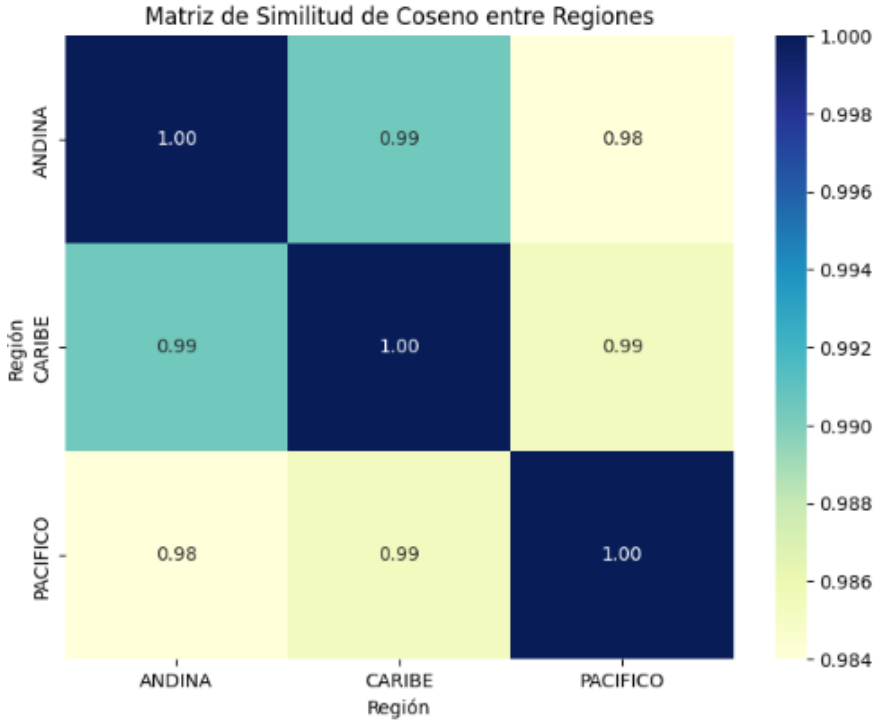

Figura 13. Resultados de la Matriz de Similitud de Coseno Modelo TF-IDF

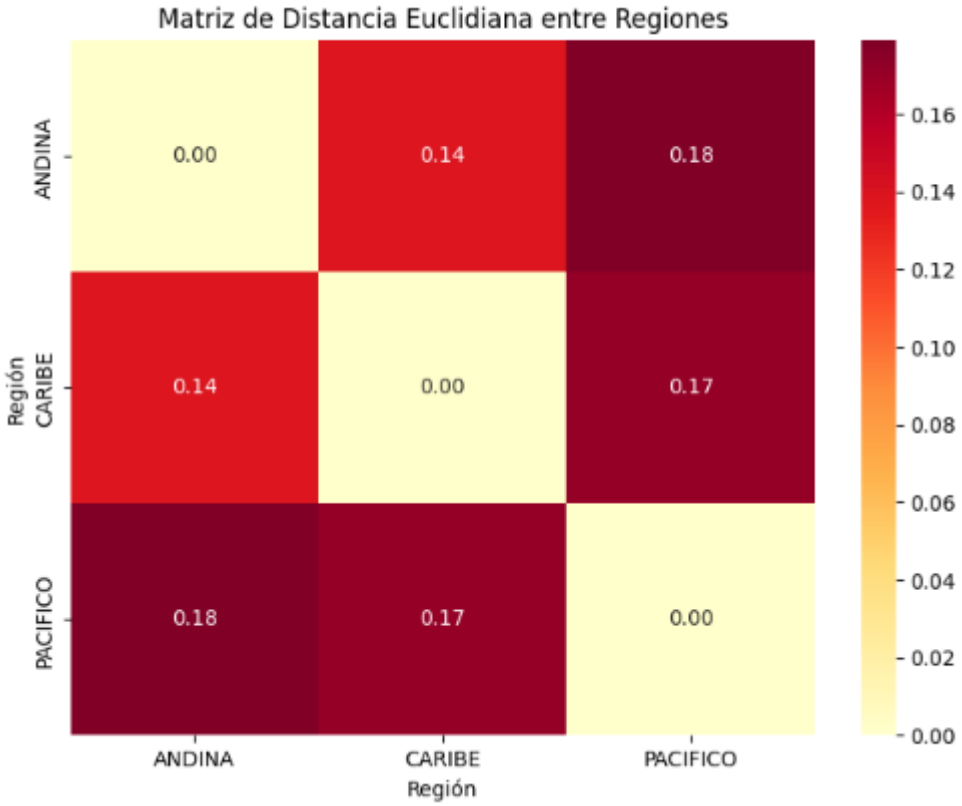

Figura 14. Resultados de la Matriz de Distancia Euclidiana Modelo TF-IDF

Los resultados que se observan en la Figura 13. muestran una alta similitud entre las regiones en cuanto a las noticias relacionadas con la reforma a la salud. Dentro de cada región, la similitud es total, con valores de 1.00 en las celdas diagonales, lo que indica coherencia interna en el tratamiento del tema. Al comparar entre regiones, se destaca una alta similitud entre Caribe y Pacífico (0.99), así como entre Andina y Caribe (0.99), lo que sugiere un enfoque o contenido muy similar en dichas regiones. La menor similitud relativa se presenta entre Andina y Pacífico (0.98), aunque este valor continúa siendo elevado, lo que indica que, en general, las noticias sobre la reforma a la salud presentan un tratamiento bastante homogéneo en todo el territorio analizado.

Por otro lado, las distancias entre las regiones tal como se muestran en la Figura 14. confirman estas observaciones. Las distancias más bajas se encuentran entre Caribe y Pacífico (0.17) y entre Caribe y Andina (0.14), lo que refuerza la idea de que las noticias en estas regiones son bastante semejantes. La distancia más alta se encuentra entre Andina y Pacífico (0.18), lo que indica que, aunque ambas regiones comparten un alto grado de similitud, existen algunas diferencias en el enfoque de la cobertura de la reforma a la salud, especialmente en términos de vocabulario o contexto. En resumen, las noticias sobre la reforma a la salud en las tres regiones muestran una gran coincidencia en términos de contenido, pero con algunas pequeñas variaciones, especialmente entre las regiones Andina y Pacífico, que podrían estar relacionadas con enfoques regionales específicos o diferencias en los temas tratados dentro

de la reforma.

7.1.2 Modelo Doc2Vec

Implementando este modelo se obtuvieron los siguientes resultados:

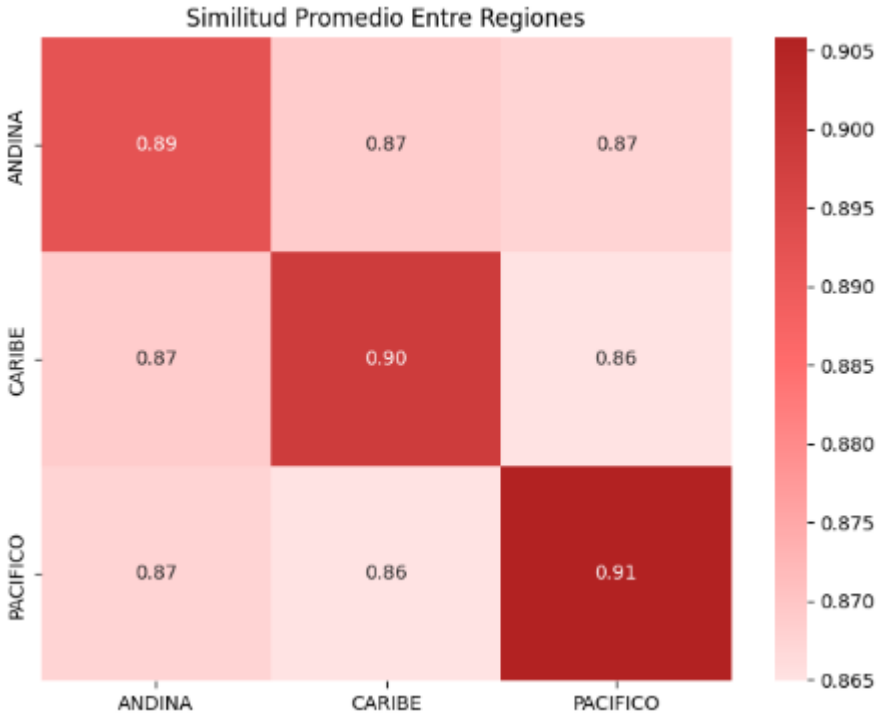

Figura 15. Resultados de la Matriz de Similitud Promedio Modelo Doc2Vec

De los resultados presentados en la Figura 15. la matriz de similitudes calculada mediante el modelo Doc2Vec aplicada a cada documento permite observar la métrica de similitud de coseno utilizada para comparar noticias sobre la reforma a la salud en Colombia. Esta revela un alto grado de cohesión semántica dentro de cada región. Los valores de similitud interna oscilan entre 0.89 (Andina), 0.90 (Caribe) y 0.91 (Pacífico), lo que indica que, si bien las noticias dentro de una misma región no son idénticas, comparten un enfoque temático y discursivo muy similar. Esto refleja una cobertura mediática homogénea en torno a la reforma. Por otro lado, las similitudes entre regiones, aunque ligeramente menores, siguen siendo significativas, con valores que van desde 0.86 (Andina-Caribe) hasta 0.87 (Andina-Pacifico). Esto sugiere que, a pesar de las diferencias geográficas, culturales y posiblemente contextuales, existen elementos comunes en la forma en que las noticias abordan la reforma a la salud, lo que podría indicar la presencia de narrativas compartidas o enfoques similares en la cobertura nacional.

7.1.3 Modelo MpNet aplicado a todas las particiones entre sí

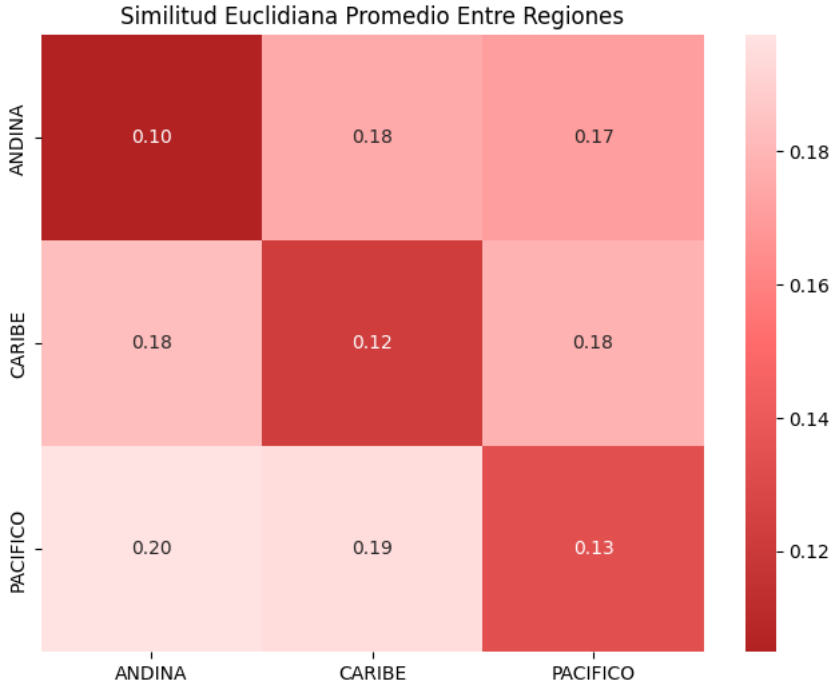

Figura 16. similitud Euclidiana Promedio Entre Regiones

La matriz presentada en la Figura 16. fue construida empleando el motor de búsqueda vectorial FAISS, utilizando la métrica de distancia euclidiana para calcular la similitud entre los vectores generados por el modelo paraphrase-multilingual-mpnet-base-v2, implementado a través de la librería fastembed. Esta matriz permite cuantificar el grado de similitud semántica entre fragmentos textuales agrupados por región. Los resultados obtenidos permiten observar que las regiones Andina, Caribe y Pacífico presentan una alta cohesión interna en el tratamiento mediático de la reforma a la salud en Colombia, con distancias bajas en la diagonal principal (Andina: 0.104, Caribe: 0.122, Pacífico: 0.133), lo que sugiere un enfoque semántico similar dentro de cada región. Sin embargo, las distancias entre regiones son considerablemente mayores, reflejando diferencias en el enfoque o prioridades informativas. En particular, las distancias Andina-Caribe (0.175), Andina-Pacífico (0.170) y Caribe-Pacífico (0.178) indican que, si bien existen elementos comunes en el discurso, cada región aborda el tema con matices distintivos. Cabe destacar que la región Andina, además de mostrar la menor distancia interna, presenta también las menores distancias con las otras regiones (0.175 con el Caribe y 0.170 con el Pacífico), lo que sugiere un enfoque más central o representativo en la cobertura noticiosa. En contraste, el Pacífico evidencia una mayor distancia tanto interna (0.133) como externa (0.170 con la Andina y 0.178 con el Caribe), lo cual podría interpretarse como un tratamiento más particular o diferenciado del tema. Estos hallazgos permiten concluir que, si bien existe una base semántica común en la forma en que los medios abordan la reforma a la salud, cada región introduce enfoques y énfasis específicos que posiblemente responden a contextos locales o a intereses mediáticos regionales.

7.1.4 Modelo MpNet aplicado de manera secuencial

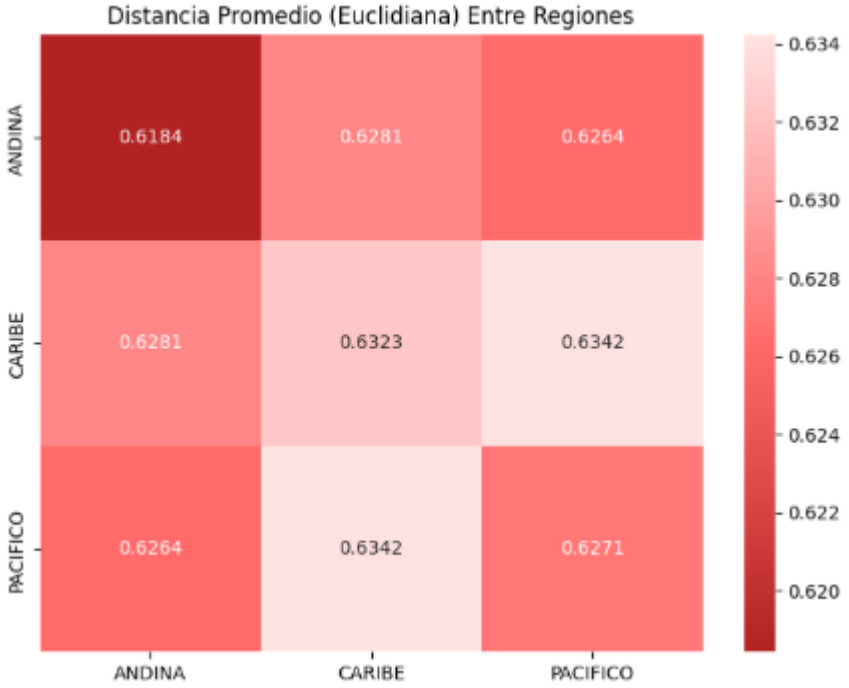

Figura 17. Distancia Promedio Euclidiana Entre Regiones

La matriz vista en la Figura 17. revela patrones en la cobertura noticiosa de la reforma a la salud en Colombia. Se observa que las distancias internas dentro de cada región (diagonal principal) son Andina: 0.621, Caribe: 0.617 y Pacífico: 0.614, lo que indica una cohesión moderada en el tratamiento mediático del tema dentro de cada zona. Estas distancias, aunque más altas que en el análisis anterior, siguen siendo relativamente bajas, sugiriendo que las noticias dentro de cada región comparten un enfoque semántico similar, aunque con una mayor variabilidad en comparación con el método previo.

Por otro lado, las distancias entre regiones son ligeramente mayores, pero aún cercanas a las distancias internas: Andina-Caribe: 0.622, Andina-Pacífico: 0.622 y Caribe-Pacífico: 0.620. Esto sugiere que, a pesar de las diferencias en el enfoque o prioridades en la cobertura, existen elementos comunes en el discurso noticioso entre las regiones. La región Pacífico muestra la menor distancia interna (0.614), lo que podría indicar una mayor consistencia en su cobertura, mientras que la región Andina tiene la mayor distancia interna (0.621), reflejando una mayor diversidad en el tratamiento del tema.

Estos resultados, obtenidos mediante una metodología secuencial, resaltan que, aunque cada región mantiene cierta coherencia interna en su cobertura, existen matices y variaciones que las diferencian.

7.2 ANÁLISIS DE SENTIMIENTOS

7.2.1 Modelo BETO finiteautomata/beto-sentiment-analysis

A continuación, se presentan los resultados obtenidos con el modelo BETO tras su afinamiento al corpus de noticias sobre la reforma a la salud en Colombia. Se muestran tanto la distribución global de las categorías de sentimiento como su desagregación porcentual y regional, con el fin de ilustrar cómo BETO capturó los matices emocionales presentes en la cobertura mediática.

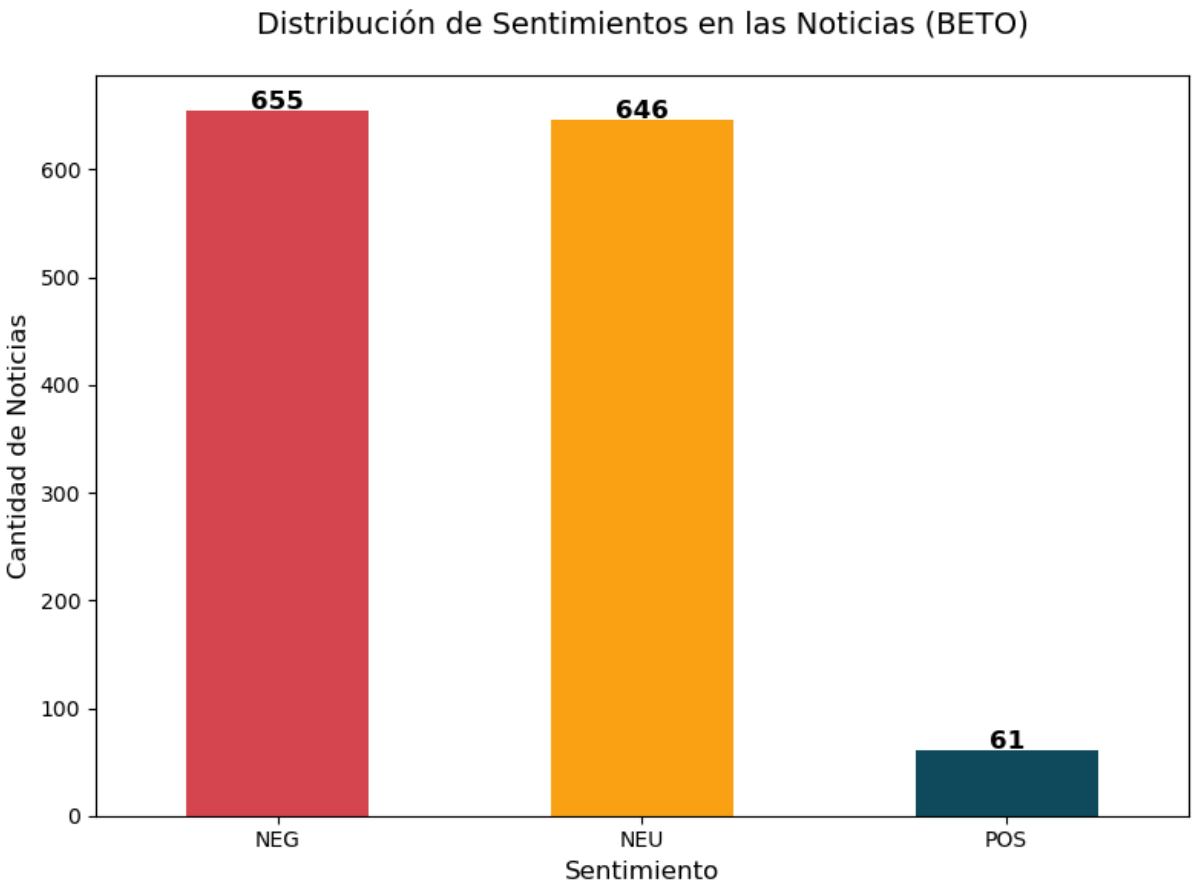

Figura 18. Distribución de Sentimiento en las Noticias BETO

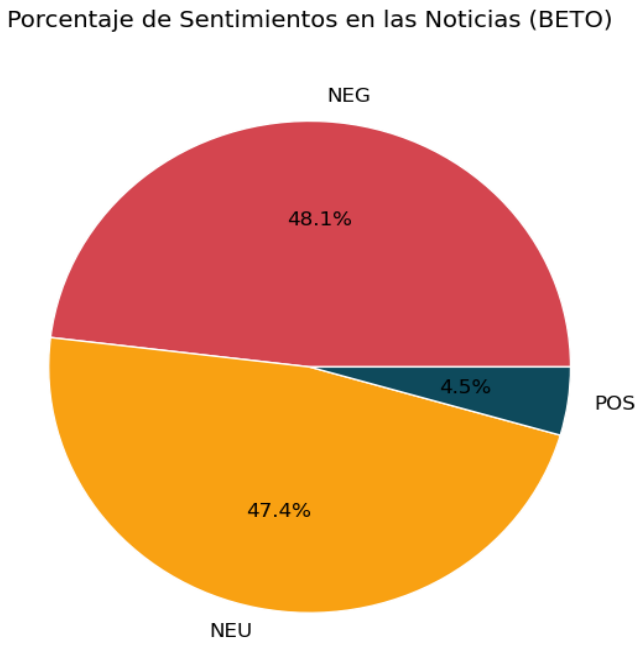

Figura 19. Porcentaje de Sentimiento en las Noticias BETO

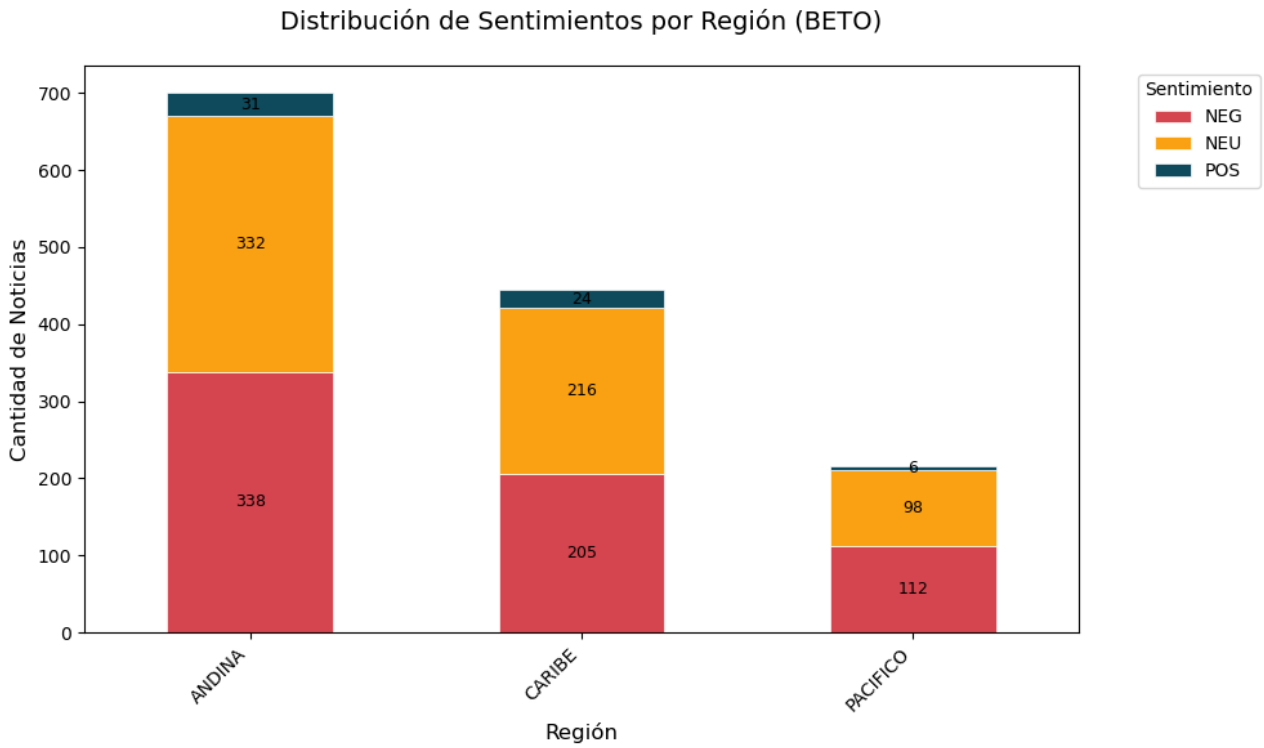

Figura 20. Distribución de sentimiento por región.

Las Figuras 18, 19 y 20 revelaron una distribución similar de sentimientos en las noticias. A nivel agregado (Figura 18), se contabilizaron 655 noticias con tono negativo (48,1 %), 646 neutrales (47,4 %) y únicamente 61 positivas (4,5 %). Esta configuración indicó que la cobertura mediática fue esencialmente informativa.

Al desagregar por región (Figura 19), en la zona Andina se registraron 340 noticias negativas (48,6 %), 330 neutrales (47,1 %) y 30 positivas (4,3 %). En la región Caribe, se identificaron 205 noticias negativas (45,6 %), 215 neutrales (47,8 %) y 30 positivas (6,7 %). Por su parte, en el Pacífico se contabilizaron 110 noticias negativas (52,4 %), 95 neutrales (45,2 %) y solo 5 positivas (2,4 %). Estas variaciones regionales sugirieron que, aunque el tono predominante fue neutral o negativo, el Caribe mostró un ligero incremento en la proporción de críticas y el Pacífico presentó la cobertura más pesimista, mientras que el Andina mantuvo un balance más cercano a la neutralidad.

En conjunto, los resultados evidenciaron que, pese a la falta de énfasis en los aspectos favorables de la reforma (solo 4,5 % de noticias positivas), la presencia de casi la mitad de las noticias con tono negativo (48,1 %) reflejó preocupaciones, críticas y problemáticas asociadas al proceso. Así, la narrativa mediática se orientó principalmente a describir hechos y cuestionar aspectos de la reforma, atendiendo de forma diferenciada las realidades y sensibilidades locales de cada región.

7.2.2 Modelo RoBERTa "*pysentimiento/robertuito-sentiment-analysis*".

A continuación, se presentan los resultados obtenidos con el modelo RoBERTa aplicado al mismo corpus de noticias sobre la reforma a la salud en Colombia. En las gráficas que siguen se muestran tanto la distribución global de las categorías de sentimiento como su comportamiento según las regiones Andina, Caribe y Pacífico, lo cual facilita comparar su sensibilidad y matices con los obtenidos.

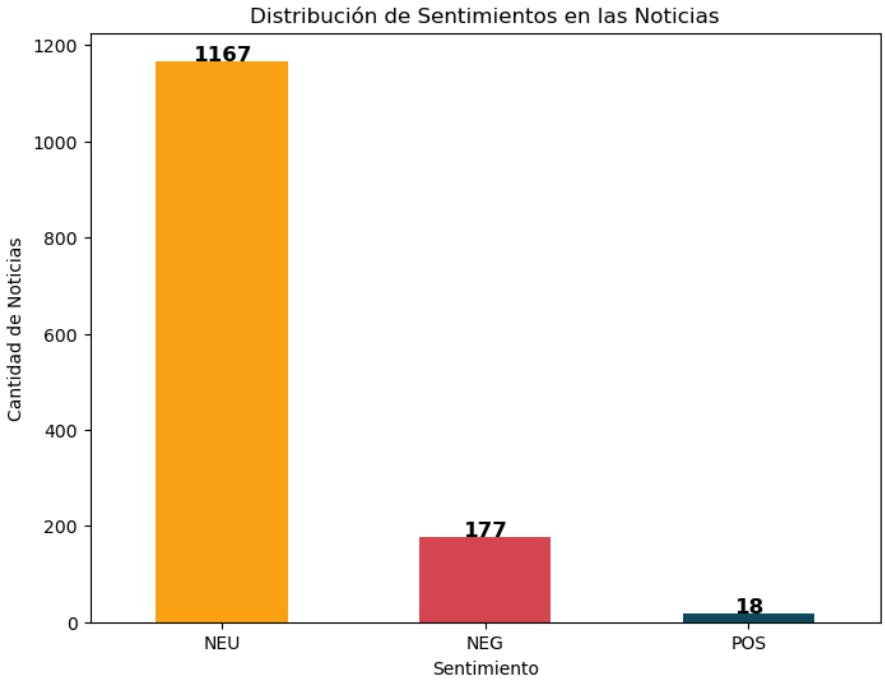

Figura 21. Distribución de Sentimiento en las Noticias RoBERTa

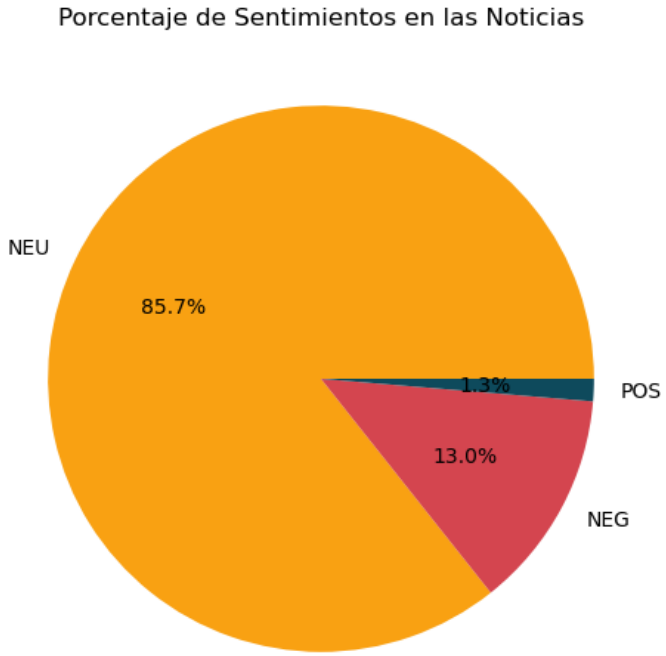

Figura 22. Porcentaje de Sentimiento en las Noticias RoBERTa

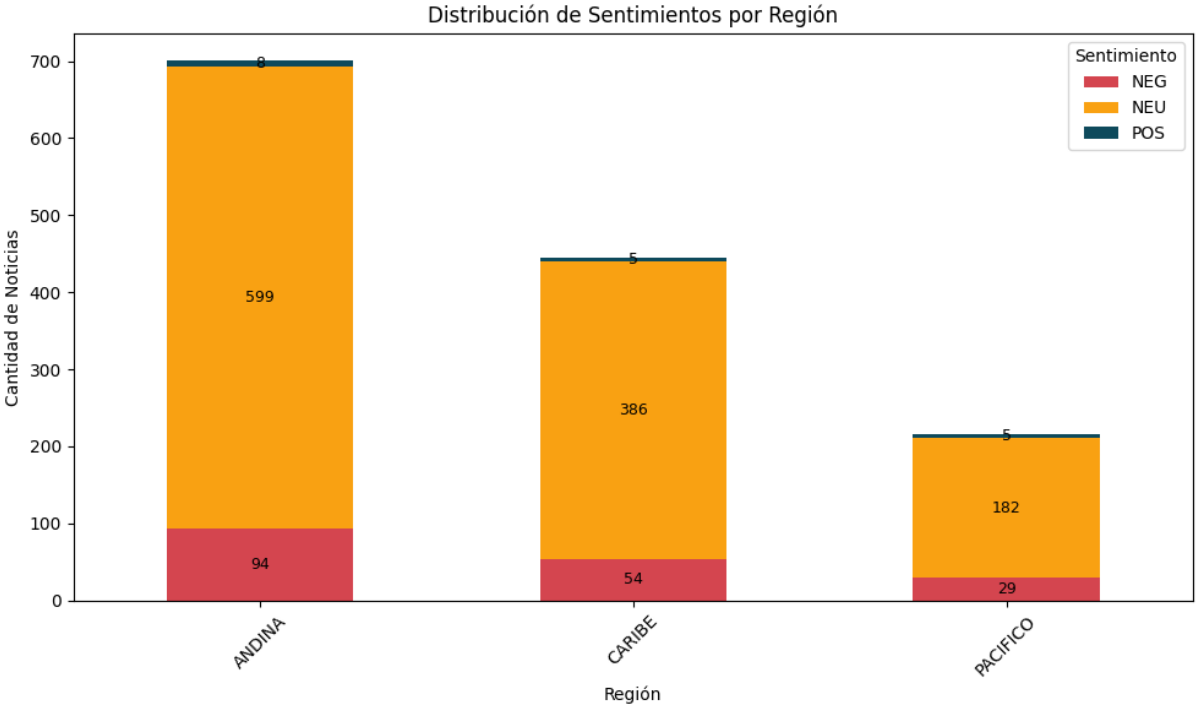

Figura 23. Distribución de sentimiento por región.

El análisis global revelo que el modelo RoBERTa clasifico el sentimiento de las 1352 noticias, de las cuales se contabilizaron 1167 (85,7 %) fueron clasificadas como neutrales, 177 (13,0 %) como negativas y tan solo 18 (1,3 %) como positivas. Estos resultados apuntaron a un claro sesgo hacia la cobertura informativa y descriptiva, en la que las valoraciones emocionales

quedaban relegadas, especialmente las de signo positivo.

Al examinar la distribución por regiones, se apreciaron ligeras diferencias que reflejaron dinámicas mediáticas locales. En la región Andina, la neutralidad alcanzó aproximadamente el 84 % de las noticias, las críticas un 13 % y las apreciaciones favorables un 3 %. En el Caribe, el 86 % de los contenidos registró un tono neutral, el 12 % uno negativo y el 2 % uno positivo. Por último, en el Pacífico sobresalió la neutralidad con un 87 %, las noticias negativas representaron el 12 % y las positivas fueron prácticamente inexistentes menos al 1 %.

La interpretación de estos hallazgos indicó que RoBERTa enfatizó en describir hechos concretos diagnósticos, declaraciones oficiales, cifras de impacto, siendo bastante neutral a la hora de clasificar el sentimiento. Por otra parte, la presencia baja de comentarios críticos (13 %) evidenció ciertos aspectos problemáticos de la política de salud, como la financiación, la cobertura en zonas rurales o los cuellos de botella administrativa. La casi nula proporción de noticias positivas (1,3 %) sugirió que los avances o beneficios percibidos, tales como mejoras en el acceso o la implementación de nuevas coberturas, no recibieron suficiente atención mediática.

7.3 EVALUACIÓN DE LOS MODELOS

7.3.1 Evaluación Modelo BETO

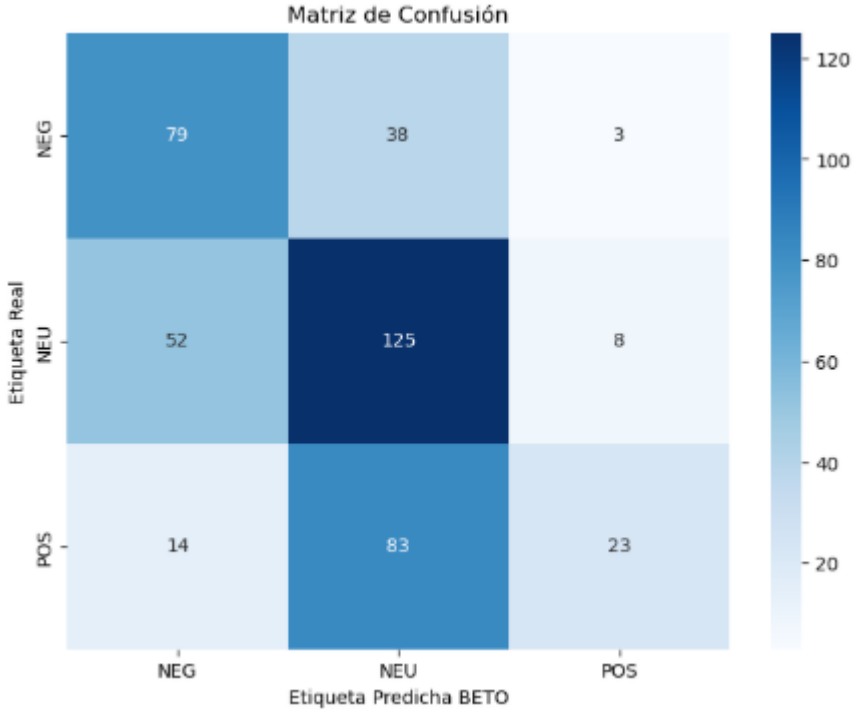

Figura 24. Matriz de confusión BETO

Clase	Precisión	Recall	F1-score	Soporte
NEG	0.54	0.66	0.60	120
NEU	0.51	0.58	0.58	185
POS	0.68	0.19	0.30	120
Exactitud	0.5341			
Macro Promedio	0.58	0.51	0.49	425
Prom. Ponderado	0.57	0.53	0.51	425

Tabla 2. Reporte de métricas del modelo BETO

Los resultados obtenidos con el modelo BETO sin fine-tuning reflejan un desempeño más limitado en la tarea de clasificación de sentimientos, con una exactitud del 53.41% y un F1-score de 0.5052, lo cual evidencia una menor capacidad del modelo para capturar adecuadamente las diferencias semánticas entre los distintos tonos. A nivel de clases, la mejor precisión se presenta en la categoría positiva (POS) con 0.68, aunque su recall es apenas 0.19, lo que indica que el modelo identifica muy pocas de las noticias positivas reales, fallando en su capacidad de detección. La clase negativa (NEG) tiene un recall del 0.66, lo que sugiere que el modelo detecta razonablemente bien este tipo de noticias, aunque su precisión es baja (0.54), lo cual implica que muchas clasificaciones negativas son incorrectas. Por su parte, la clase neutral (NEU) presenta un comportamiento similar, con una precisión de 0.51 y un recall de 0.68, revelando una alta proporción de falsos positivos desde otras clases, en especial desde la clase positiva.

La matriz de confusión refuerza esta interpretación: de las 120 noticias negativas, 79 fueron correctamente clasificadas, pero 38 fueron confundidas con neutrales, mostrando un solapamiento significativo entre ambas clases. En cuanto a las noticias neutrales, 125 de las 185 fueron correctamente identificadas, mientras que 52 se confundieron como negativas y 8 como positivas, lo que evidencia un patrón de sobre clasificación hacia lo negativo. El mayor error se da en la clase positiva, donde solo 23 de las 120 noticias reales fueron correctamente clasificadas, mientras que 83 fueron etiquetadas como neutrales. Este comportamiento sugiere que el modelo sin ajuste fino no logra distinguir adecuadamente los tonos favorables, lo cual afecta considerablemente la utilidad del sistema para identificar discursos positivos. En conjunto, estos resultados reflejan una tendencia del modelo a centrar sus predicciones en categorías más ambiguas como la neutral, y a minimizar la detección del tono positivo, lo que limita su eficacia para aplicaciones en análisis mediático sin una adaptación contextual específica.

7.3.2 Evaluación Modelo RoBERTa

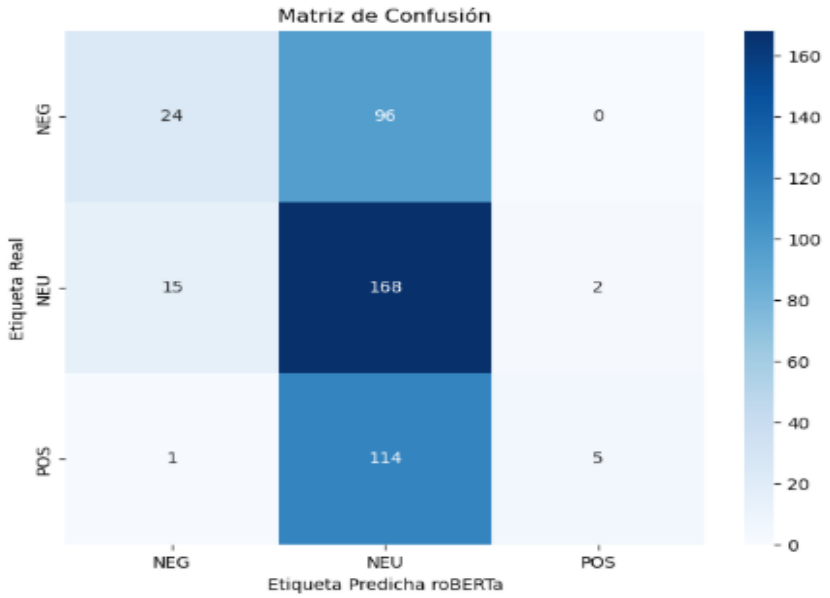

Figura 25. Matriz de confusión RoBERTa.

Clase	Precisión	Recall	F1-score	Soporte
NEG	0.60	0.20	0.30	120
NEU	0.44	0.91	0.60	185
POS	0.71	0.04	0.08	120
Exactitud	0.4635			
Macro Promedio	0.59	0.38	0.33	425
Prom. Ponderado	0.56	0.46	0.37	425

Tabla 3. Reporte de métricas del modelo RoBERTa

Los resultados obtenidos con el modelo RoBERTa sin fine-tuning evidencian un rendimiento limitado en la tarea de clasificación de sentimientos, con una exactitud de 46.35% y un F1-score global de 0.3667. Aunque la precisión general (0.5646) puede parecer moderada, esto se debe principalmente al desempeño sesgado hacia una sola clase. La clase neutral (NEU) es la que presenta mejor desempeño, con un recall de 0.91, lo que indica que el modelo identifica casi todos los textos realmente neutrales. Sin embargo, su precisión es baja (0.44), debido a que también clasifica erróneamente como neutrales una gran cantidad de textos negativos (96 de 120) y positivos (114 de 120). En contraste, la clase negativa (NEG) tiene un recall muy bajo (0.20), clasificando correctamente solo 24 de 120 textos, mientras que la clase positiva (POS) es la más afectada, con un recall de apenas 0.04, al identificar correctamente solo 5 textos positivos.

Estos resultados se ven reflejados de manera clara en la matriz de confusión, donde se observa una fuerte tendencia del modelo a clasificar textos de cualquier polaridad como neutrales.

Este sesgo hacia la clase NEU puede deberse a la falta de ajuste fino del modelo para el dominio mediático en español, lo cual limita su capacidad para diferenciar adecuadamente los extremos emocionales del lenguaje. El comportamiento observado sugiere que, sin un entrenamiento específico, RoBERTa tiende a interpretar los matices emocionales con ambigüedad, agrupando la mayoría de los casos en una zona intermedia, lo que compromete la utilidad del modelo en contextos donde la identificación precisa del tono (positivo o negativo) es esencial para el análisis de opinión pública o mediática.

7.4 FINE TUNING

En esta sección se describe el proceso de ajuste fino (fine-tuning) realizado sobre los modelos pre-entrenados en español BETO y RoBERTa, con el propósito de adaptarlos a la tarea específica de clasificación de sentimientos en un corpus de noticias relacionadas con la reforma a la salud en Colombia. Aunque estos modelos han sido entrenados con grandes volúmenes de texto en español, Fue necesario realizar el fine-tuning debido a las particularidades del lenguaje utilizado en el ámbito mediático y al carácter temático del corpus, que incluye expresiones técnicas, vocabulario institucional y matices discursivos específicos del debate sobre salud. Para este ajuste, se llevó a cabo un fine-tuning supervisado sobre un modelo de lenguaje pre-entrenado (base BERT), es decir, un entrenamiento adicional del modelo utilizando un conjunto de datos etiquetado donde el texto está asociado a una categoría emocional predefinida como neutro, positivo o negativo. Este enfoque permite ajustar los parámetros del modelo en función de una tarea especifica, optimizando directamente una función de perdida asociada a la clasificación.

Durante el proceso, se ajustaron las capas superiores del modelo (son las responsables de la representación contextual de alto nivel), mientras que las capas inferiores se mantuvieron parcialmente congeladas para preservar el conocimiento lingüístico general. Se empleó la técnica de early stopping que consiste en detener automáticamente el entrenamiento cuando el rendimiento en un conjunto de validación deja de mejorar para evitar sobreajuste. También se usó un learning rate reducido (de 2e-5) y un tamaño de batch de 16, optimizando así el entrenamiento para textos de longitud variable y estructura compleja. Este proceso de fine-tuning se llevó a cabo con el fin de mejorar las métricas de rendimiento en la clasificación, optimizando la precisión, sensibilidad y capacidad del modelo para interpretar correctamente los distintos tonos emocionales presentes en las noticias. A continuación, se presentan los análisis derivados de este proceso de entrenamiento.

7.4.1 Fine Tuning Modelo Beto

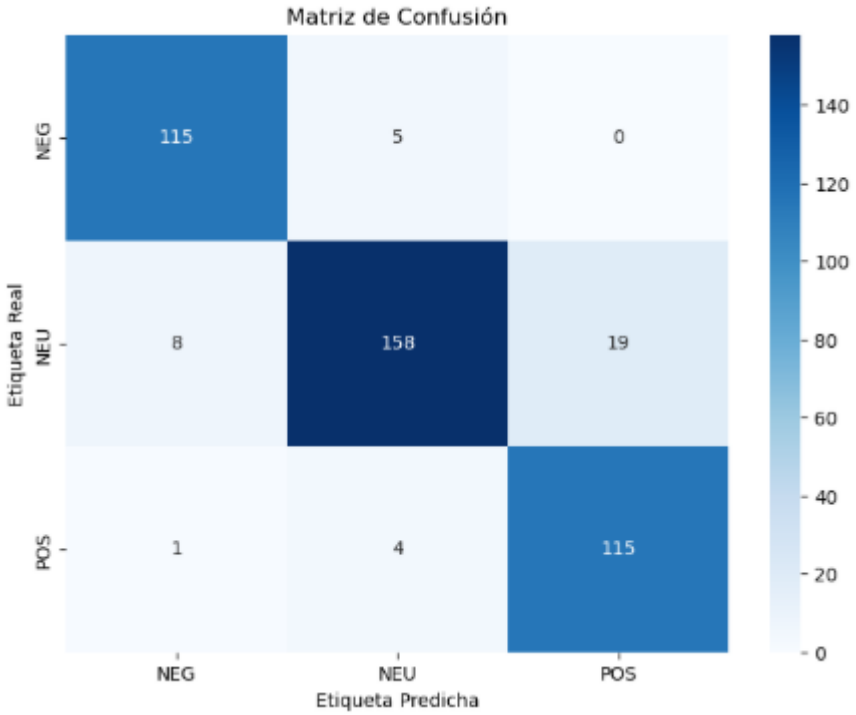

Figura 26. Matriz de confusión BETO – Fine-Tuning

Clase	Precisión	Recall	F1-score	Soporte
NEG	0.93	0.96	0.94	120
NEU	0.95	0.85	0.90	185
POS	0.86	0.96	0.91	120
Exactitud	0.9129			
Macro Promedio	0.91	0.92	0.92	425
Prom. Ponderado	0.92	0.91	0.91	425

Tabla 4. Reporte de métricas del modelo BETO – Fine-Tuning

El modelo BETO, tras ser ajustado mediante fine-tuning, presentó un excelente desempeño general en la tarea de clasificación de sentimientos. Las métricas globales evidencian una alta capacidad predictiva: la exactitud alcanzó un valor de 0.9129, mientras que el F1-score fue de 0.9126, con una precisión de 0.9160 y un recall de 0.9129. En términos de desempeño por clase, los sentimientos negativos (NEG) obtuvieron un F1-score de 0.94, con una precisión de 0.93 y un recall de 0.96, lo que indica que el modelo identifica eficazmente las noticias con tono negativo. La clase neutral (NEU) mostró una precisión de 0.95, un recall de 0.85 y un F1-score de 0.90, revelando una ligera dificultad en recuperar todos los casos neutros. Por su parte, los sentimientos positivos (POS) también se clasificaron con alto rendimiento, alcanzando una precisión de 0.86, un recall de 0.96 y un F1-score de 0.91.

La matriz de confusión refuerza este análisis mostrando un comportamiento equilibrado en la clasificación. De un total de 120 noticias negativas, 115 fueron correctamente clasificadas, mientras que solo 5 fueron confundidas como neutrales. En el caso de las 185 noticias neutrales, 158 fueron identificadas correctamente, pero 8 fueron clasificadas erróneamente como negativas y 19 como positivas, lo que explica la ligera disminución del recall en esta clase. Finalmente, de las 120 noticias positivas, 115 fueron clasificadas adecuadamente, mientras que 5 se confundieron como negativas o neutras. Estos resultados evidencian que el modelo BETO es particularmente robusto para identificar noticias negativas y positivas, con una pequeña área de mejora en la diferenciación de matices dentro de las noticias neutrales.

7.4.2 Fine Tuning Modelo Roberta

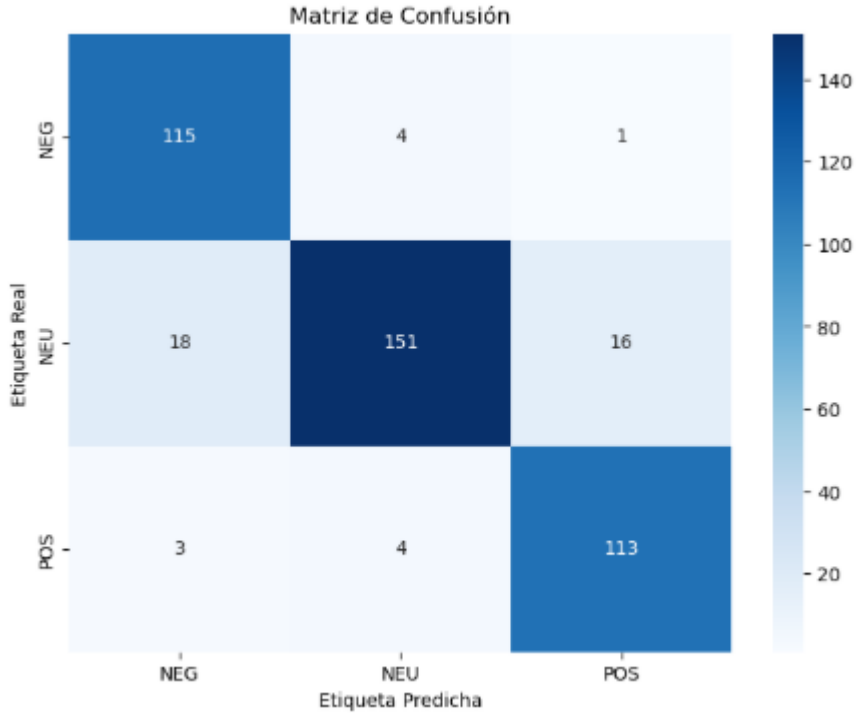

Figura 27. Matriz de confusión RoBERTa – Fine-Tuning

Métrica	NEG	NEU	POS	Promedio
Precisión	0.85	0.95	0.87	0.8976
Recall	0.96	0.82	0.94	0.8918
F1-score	0.90	0.88	0.90	0.8911
Soporte	120	185	120	425
Accuracy	0.8918			

Tabla 5. Reporte de métricas del modelo RoBERTa – Fine-Tuning

El modelo RoBERTa, ajustado mediante fine-tuning, mostró un desempeño sólido en la tarea de clasificación de sentimientos aplicados a noticias sobre la reforma a la salud, alcanzando una exactitud de 0.8918 y un F1-score promedio de 0.8911. En términos de desempeño por clase, el modelo logró identificar eficazmente las noticias negativas (NEG), con un F1-score de 0.90, una precisión de 0.85 y un recall de 0.96, lo que indica una alta capacidad para recuperar correctamente este tipo de textos, aunque con una leve tendencia a sobre clasificar otras categorías como negativas. La clase positiva (POS) también fue bien clasificada, con una precisión de 0.87, recall de 0.94 y F1-score de 0.90, mostrando un equilibrio adecuado entre precisión y sensibilidad.

Sin embargo, el mayor desafío del modelo se presentó en la identificación de las noticias neutrales (NEU), donde, a pesar de alcanzar una alta precisión (0.95), el recall se redujo a 0.82, evidenciando dificultades para recuperar todos los casos de esta categoría. La matriz de confusión confirma esta tendencia, revelando que varias noticias neutrales fueron clasificadas erróneamente como negativas o positivas. En conjunto, los resultados indican que RoBERTa es especialmente eficaz en la clasificación de noticias con carga emocional clara (positiva o negativa), aunque existe un margen de mejora en la distinción de contenidos neutrales, probablemente debido a su menor carga lingüística explícita de polaridad.

8. COMPARATIVA DE MODELOS ANTES Y DESPUÉS DEL FINE-TUNING

8.1 COMPARATIVA MODELO BETO

Clase/Métrica	Precisión	Recall	F1-score	Soporte
BETO sin fine-tuning				
NEG	0.54	0.66	0.60	120
NEU	0.51	0.58	0.58	185
POS	0.68	0.19	0.30	120
Exactitud	**0.5341**			
Precisión general	**0.5660**			
Recall general	**0.5341**			
F1-score general	**0.5052**			
BETO con fine-tuning				
NEG	0.93	0.96	0.94	120
NEU	0.95	0.85	0.90	185
POS	0.86	0.96	0.91	120
Exactitud	**0.9129**			
Precisión general	**0.9160**			
Recall general	**0.9129**			
F1-score general	**0.9126**			

Tabla 6. comparación de las métricas del modelo BETO

El modelo BETO sin fine-tuning presenta un rendimiento moderado, con una exactitud general del 53.41% y un F1-score promedio de 0.5052. Si bien muestra cierto equilibrio entre precisión y recall para las clases NEG (Negativo) y NEU (Neutral), tiene una marcada dificultad para identificar correctamente los textos positivos (POS), con un recall de apenas 0.19 y un F1-score de 0.30. Esto indica que, sin un ajuste específico al dominio, el modelo tiende a subestimar los sentimientos positivos, lo cual puede deberse a una falta de representación adecuada de este tipo de textos en los datos pre-entrenados.

En contraste, tras aplicar fine-tuning, BETO muestra una mejora sustancial en todas las métricas. La exactitud global se eleva al 91.29% y el F1-score general alcanza 0.9126, lo que indica una capacidad mucho mayor para clasificar correctamente los distintos tipos de

sentimientos. Las tres clases presentan valores altos de precisión y recall, con la clase POS mostrando una mejora significativa, pasando de un F1-score de 0.30 a 0.91. Esta mejora refleja la efectividad del fine-tuning para adaptar el modelo a las características lingüísticas y contextuales específicas del corpus analizado, permitiendo una clasificación más precisa, balanceada y coherente en el contexto de los textos mediáticos en español.

8.2 COMPARATIVA MODELO ROBERTA

Clase/Métrica	Precisión	Recall	F1-score	Soporte
RoBERTa sin fine-tuning				
NEG	0.60	0.20	0.30	120
NEU	0.44	0.91	0.60	185
POS	0.71	0.04	0.08	120
Exactitud	**0.4635**			
Precisión general	**0.3667**			
Recall general	**0.5646**			
F1-score general	**0.4635**			
RoBERTa con fine-tuning				
NEG	0.85	0.96	0.90	120
NEU	0.95	0.82	0.88	185
POS	0.87	0.94	0.90	120
Exactitud	**0.8918**			
Precisión general	**0.8911**			
Recall general	**0.8976**			
F1-score general	**0.8918**			

Tabla 7. Comparación de las métricas del modelo RoBERTa

Los resultados obtenidos por el modelo RoBERTa sin ajuste fino muestran un rendimiento limitado en la tarea de clasificación de sentimientos. Con una exactitud general del 46.35% y un F1-score de apenas 0.3667, el modelo tiene serias dificultades para distinguir adecuadamente entre las clases. Aunque la clase neutral (NEU) presenta un buen recall (0.91), lo que indica que el modelo detecta la mayoría de los textos con tono neutral, esto viene a costa de una baja precisión (0.44), lo que implica una alta tasa de falsos positivos. Por el contrario, las clases negativa (NEG) y positiva (POS) presentan un desempeño especialmente débil: NEG con un recall de 0.20 y POS con tan solo 0.04, evidenciando que el modelo casi no logra identificar correctamente los textos de esas categorías. Esto se traduce en un modelo

altamente desbalanceado, probablemente afectado por la falta de entrenamiento especializado en el dominio y en idioma español.

En contraste, tras el proceso de fine-tuning, RoBERTa mejora notablemente su desempeño. La exactitud se eleva hasta el 89.18% y el F1-score general alcanza 0.8911, lo que indica un modelo mucho más equilibrado y efectivo. Las tres clases muestran mejoras sustanciales: NEG alcanza un F1-score de 0.90, NEU mejora su precisión a 0.95 con un F1-score de 0.88, y POS también muestra un avance significativo con un F1-score de 0.90. Estas métricas reflejan un modelo afinado que no solo distingue correctamente los sentimientos expresados en los textos, sino que además lo hace de forma consistente entre las diferentes categorías. El ajuste fino ha permitido que RoBERTa se adapte adecuadamente a los matices del lenguaje y del contenido mediático en español, consolidándose como una herramienta sólida para el análisis de sentimientos en este contexto.

9. MODELO CHATGPT

Como ejercicio complementario al proceso principal de ajuste y evaluación de los modelos BETO y RoBERTa para la clasificación de sentimientos en textos relacionados con la reforma a la salud en Colombia, se implementó un modelo de clasificación basado en ChatGPT-4.0. Este análisis tiene como objetivo explorar el comportamiento de ChatGPT-4.0 frente a los modelos entrenados específicamente en español, con el fin de observar diferencias en la precisión, sensibilidad al contexto y capacidad de interpretación de matices lingüísticos en el dominio específico de la opinión pública y mediática. Aunque este ejercicio no forma parte del eje central de la investigación, permitirá ampliar la comprensión sobre las posibilidades actuales de los modelos de lenguaje en tareas de análisis de sentimientos y complementar los hallazgos obtenidos con los modelos entrenados en la etapa principal. A continuación, se presentará el desarrollo de dicho análisis.

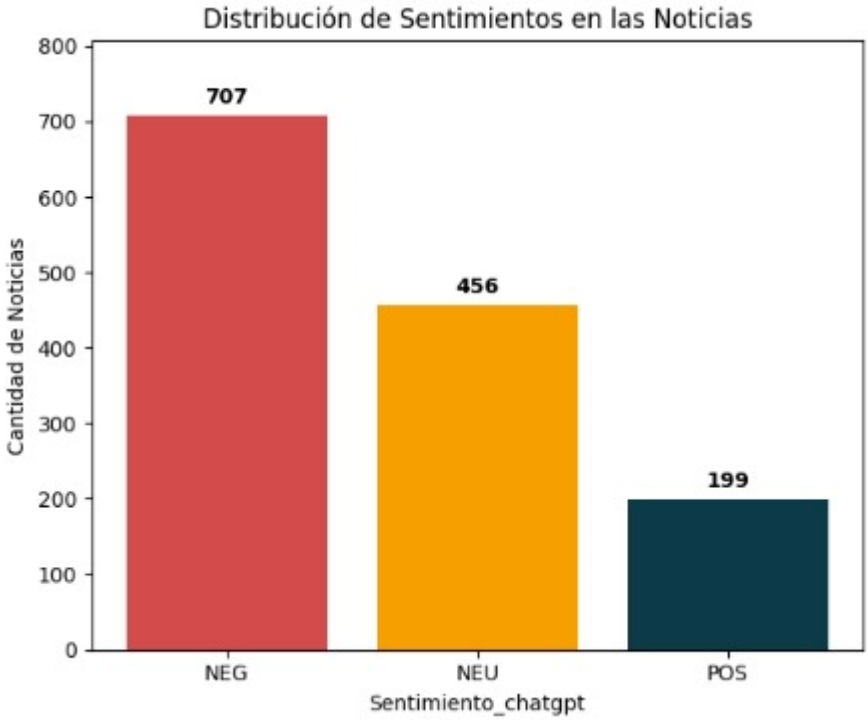

Figura 28. Distribución de sentimiento en las noticias Chatgpt

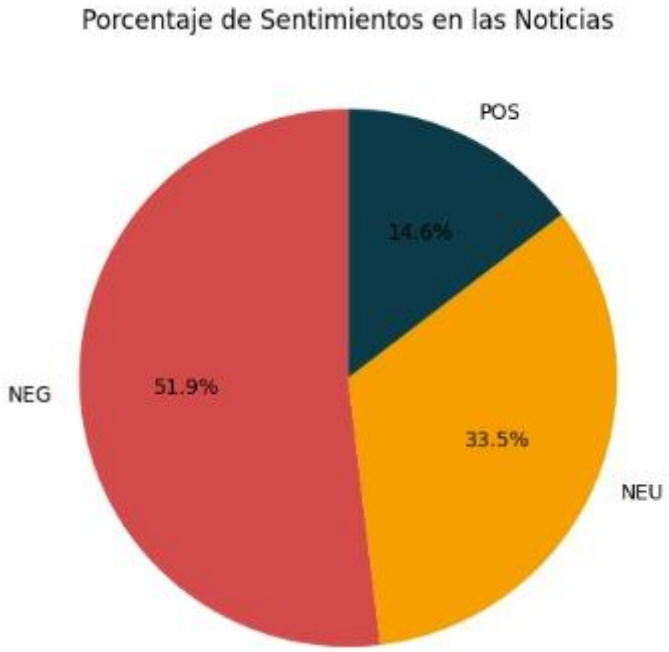

Figura 29. Porcentaje de sentimiento en las noticias

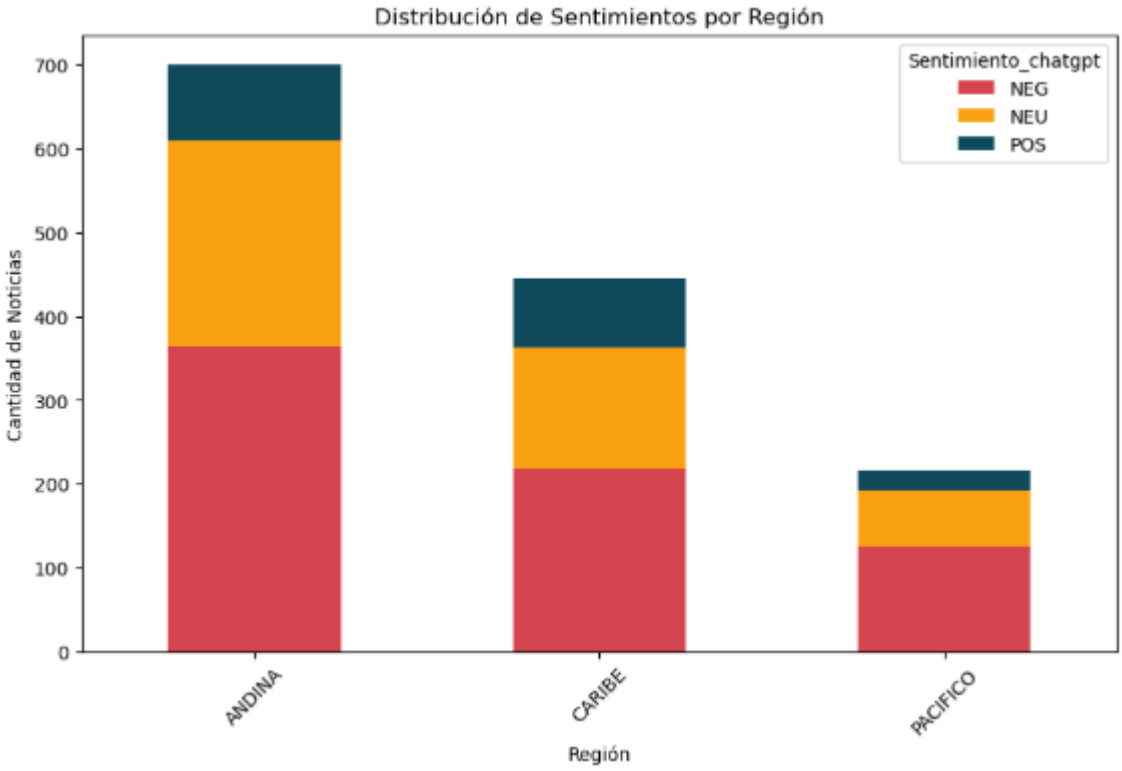

Figura 30. Distribución de sentimiento por regiones.

Las figuras que se presentan son la: 28, 29 y 30 que revelan una clara predominancia de sentimientos negativos (NEG) en la cobertura mediática, los cuales representan el 51.9% del total (707 noticias). Esta alta proporción evidencia que más de la mitad de las noticias analizadas presentan un enfoque crítico o problemático frente al tema tratado. Le siguen los

sentimientos neutrales (NEU), con un 33.5% (456 noticias), lo que sugiere una cobertura informativa relativamente objetiva en una parte importante del contenido. Por último, los sentimientos positivos (POS) son los menos frecuentes, con un 14.6% (199 noticias), lo que indica que los logros, beneficios o visiones favorables relacionados con el tema no han sido ampliamente resaltados en la narrativa mediática. En conjunto, esta distribución muestra una inclinación general hacia una percepción negativa del fenómeno analizado, lo cual puede tener implicaciones importantes sobre la opinión pública.

Región	NEG	NEU	POS
Andina	364	245	92
Caribe	217	145	83
Pacífico	126	66	24

Tabla 8. distribución de sentimientos por región, modelo Chatgpt

Al analizar la distribución de sentimientos por región, se identifican variaciones significativas que podrían estar relacionadas con particularidades locales o con los enfoques mediáticos específicos de cada zona. En la región Andina, que incluye ciudades como Bogotá y Medellín, se concentra la mayor cantidad de noticias negativas (364), lo que representa una proporción considerable del total en esta región. Le siguen las noticias neutrales (245) y, en menor medida, las positivas (92). Esta distribución sugiere una cobertura con tendencia crítica, aunque se destacan también algunos contenidos con matices más favorables al fenómeno analizado.

En la región Caribe, las noticias negativas también son predominantes (217), seguidas por las neutrales (145) y las positivas (83). Aunque la cantidad absoluta de noticias positivas es menor, su proporción en comparación con otras regiones es relativamente alta, lo cual podría indicar una narrativa mediática que, si bien crítica, otorga cierto espacio a la visibilidad de beneficios o aspectos positivos.

Por su parte, la región Pacífico se caracteriza por una marcada predominancia de noticias negativas (126), lo que evidencia una cobertura más crítica o preocupada. Las noticias neutrales (66) y las positivas (24) aparecen en menores proporciones, reflejando una percepción menos optimista del fenómeno tratado, posiblemente asociada con las dinámicas sociales, económicas o institucionales específicas de la región.

En conjunto, estos resultados muestran que, si bien existe una fuerte presencia de sentimientos negativos en todas las regiones analizadas, las diferencias en la proporción de noticias neutras y positivas reflejan enfoques narrativos diferenciados. La región Andina presenta el volumen más alto de noticias, tanto negativas como positivas, mientras que el Caribe se distingue por una relativa mayor presencia de contenido con tono positivo. En contraste, la región Pacífico proyecta una visión predominantemente crítica. Estas variaciones podrían estar influidas por la agenda mediática local, la cercanía al poder político, las condiciones regionales o el impacto diferencial del fenómeno abordado en cada contexto.

9.1 EVALUACIÓN MODELO CHATGPT-4O

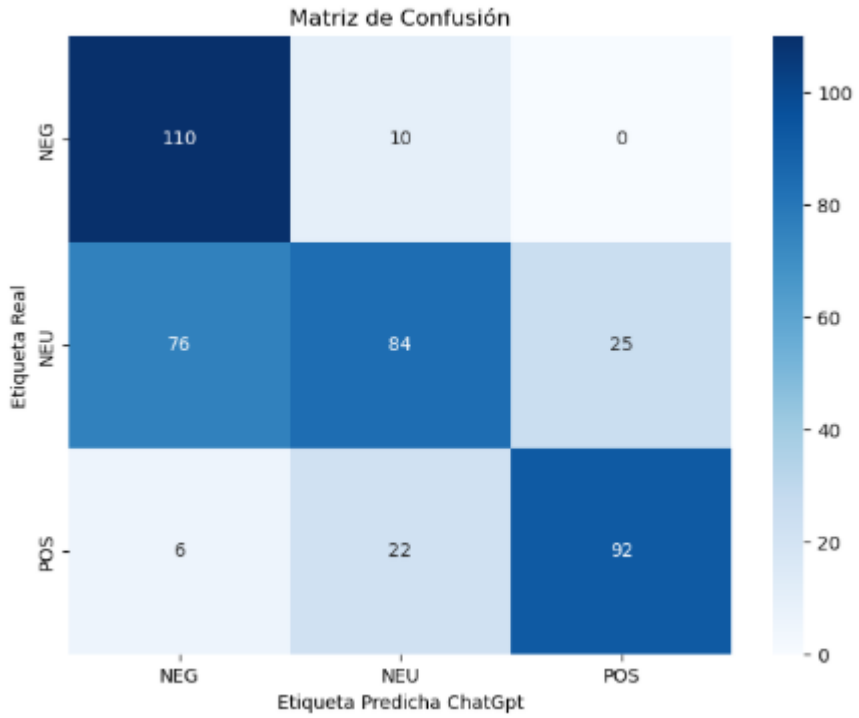

Figura 31. Matriz de confusión Chatgpt 4o

Clase	Precisión	Recall	F1-score	Soporte
NEG	0.57	0.92	0.71	120
NEU	0.72	0.45	0.56	185
POS	0.79	0.77	0.78	120
Exactitud (Accuracy)	0.6729			
Macro Promedio	0.69	0.71	0.68	425
Prom. Ponderado	0.70	0.67	0.66	425

Tabla 9. Reporte de métricas del modelo Chatgpt-4º

Los resultados obtenidos con el modelo de clasificación de sentimientos ChatGPT-4.0 muestran un desempeño general aceptable, con una exactitud del 67.29% y un F1-score de 0.6613. En términos de precisión, la clase positiva (POS) se destaca con un valor de 0.79 y un F1-score de 0.78, lo que indica que el modelo logra identificar de forma bastante precisa las noticias con tono favorable. La clase negativa (NEG) presenta una alta capacidad de detección, con un recall del 0.92, aunque con una precisión más baja (0.57), lo que sugiere que el modelo tiende a clasificar como negativas algunas noticias que no lo son, posiblemente por su sensibilidad ante términos con carga crítica. En contraste, la clase neutral (NEU) muestra mayores dificultades, con un recall del 0.45 y un F1-score de 0.56, lo que indica una menor capacidad del modelo para identificar correctamente los textos informativos o imparciales.

Estos resultados se ven reflejados en la matriz de confusión: el modelo identificó correctamente 110 de las 120 noticias negativas, mientras que confundió un número considerable de noticias neutrales, clasificando 76 como negativas y 25 como positivas. Asimismo, 92 de las 120 noticias positivas fueron correctamente identificadas, lo que reafirma el buen desempeño del modelo en esta categoría. El comportamiento observado sugiere que ChatGPT-4.0 tiende a sobredimensionar los sentimientos negativos, posiblemente debido a la forma en que fue entrenado y a la ausencia de un ajuste fino específico para el dominio mediático en español. Esto afecta particularmente la clasificación de tonos intermedios como el neutral, lo cual es clave para una interpretación más matizada de la opinión pública y mediática.

9.2 COMPARATIVA MODELO BETO Y ROBERTUITO VS CHATGPT-4O

Modelo	Precisión	Recall	F1-score	Exactitud
BETO (Ajustado)	0.9160	0.9129	0.9126	0.9129
RoBERTa (Ajustado)	0.8976	0.8918	0.8911	0.8918
ChatGPT-4º	0.6990	0.6729	0.6613	0.6729

Tabla 10. Comparación de las métricas entre los modelos ajustados

Los resultados presentados en la Tabla anterior permiten comparar el rendimiento de tres modelos de lenguaje para la clasificación de sentimientos en noticias en español: BETO y RoBERTa (ambos ajustados con fine-tuning), y ChatGPT-4.0. En términos generales, se observa un desempeño superior por parte de los modelos específicamente entrenados para el español, destacándose BETO, que alcanza la mayor exactitud (91.29%) y el mayor F1-score (0.9126), seguido de RoBERTa, con una exactitud de 89.18% y un F1-score de 0.8911. Estos valores indican una alta capacidad para clasificar correctamente los sentimientos en los textos evaluados, reflejando la efectividad del fine-tuning sobre un conjunto de datos etiquetado y balanceado.

En contraste, ChatGPT-4.0, a pesar de ser un modelo de propósito general y sin ajuste específico para este dominio, logra un desempeño aceptable con una exactitud del 67.29% y un F1-score de 0.6613. Sin embargo, su menor recall (0.6729) y precisión (0.6990) en comparación con los modelos entrenados indican una menor capacidad para identificar consistentemente las clases de sentimiento, especialmente en textos con tonos más matizados o neutros. Esta diferencia puede atribuirse a la falta de especialización del modelo de ChatGPT en el análisis de sentimientos en español, lo cual resalta la importancia del ajuste fino sobre corpus específicos y en idioma objetivo para tareas de clasificación supervisada.

10.ANÁLISIS DE RESULTADOS DE LOS MODELOS BETO Y ROBERTA

En este apartado se presentará el análisis de resultados obtenido a partir de los modelos BETO y RoBERTa, los cuales fueron adaptados mediante un proceso de fine-tuning para realizar la clasificación de sentimientos en noticias relacionadas con la reforma a la salud. Tras comprobar el buen desempeño de ambos modelos a través de métricas de evaluación vistas en el capítulo anterior, en esta sección se analizarán en detalle los resultados obtenidos, permitiendo identificar patrones de opinión y tendencias en el tratamiento mediático de la reforma.

10.1 ANÁLISIS GENERAL MODELO BETO

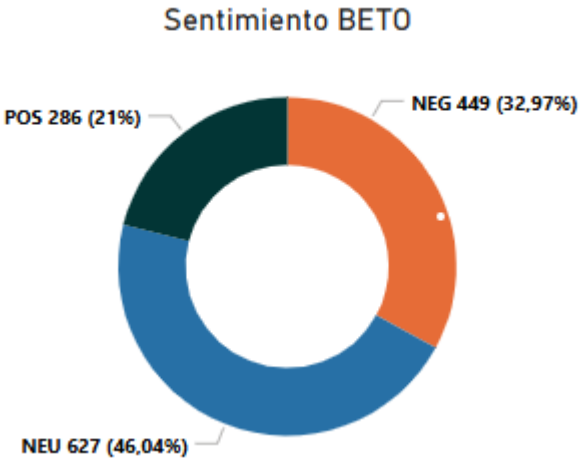

Figura 32. Distribución de sentimiento BETO

Como se logra observar en la Figura 32. El análisis de sentimiento clasificó los textos en tres categorías: positivo (POS), negativo (NEG) y neutro (NEU). De acuerdo con los resultados obtenidos, en el modelo BETO el sentimiento predominante es el neutro, con un 46,04% de las observaciones. Le siguen el sentimiento negativo con un 32,97% y finalmente el sentimiento positivo con un 21%.

Estos resultados permiten concluir que, según el modelo BETO, existe una mayor tendencia a expresiones neutrales respecto a la reforma a la salud. No obstante, el porcentaje de sentimientos negativos es considerablemente más alto que el de sentimientos positivos, lo que refleja una importante presencia de preocupaciones y críticas en la conversación pública. La menor proporción de opiniones positivas sugiere que el respaldo explícito a la reforma es limitado en los textos analizados.

Este patrón detectado por el modelo BETO será tomado en cuenta para el análisis comparativo posterior, en el cual se explorarán las diferencias de percepción entre las diferentes regiones

del país y los diversos medios de comunicación y plataformas digitales.

10.2 ANÁLISIS GENERAL MODELO ROBERTA

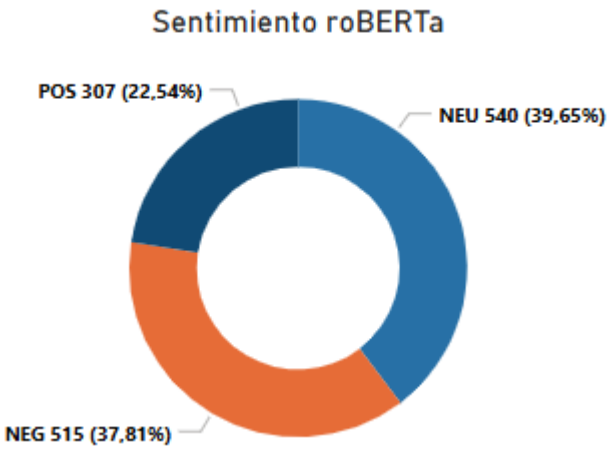

Figura 33. Distribución sentimiento RoBERTa

Según los resultados que se muestran en la Figura 33. El sentimiento del Modelo RoBERTa asociado a las noticias sobre la reforma a la salud evidencian que el 39,65% de las publicaciones se clasifican como neutras, mientras que el 37,81% muestran una carga negativa y el 22,54% presentan un sentimiento positivo. Esta distribución refleja que, si bien predomina una postura neutral, existe una importante proporción de opiniones negativas, lo cual indica un ambiente mediático crítico en torno a la reforma. En comparación con los resultados obtenidos anteriormente mediante el modelo BETO, se mantiene la primacía del sentimiento neutro; no obstante, en este caso la diferencia entre los sentimientos neutro y negativo es mucho menor, sugiriendo una mayor polarización del discurso. En ambos análisis el sentimiento positivo permanece como el menos representativo, aunque en este modelo su proporción es ligeramente más alta. Estos hallazgos permiten inferir una percepción pública que oscila entre la neutralidad y la crítica, limitando la presencia de apoyo explícito hacia la propuesta de reforma.

10.3 EVOLUCIÓN DEL SENTIMIENTO POR AÑO

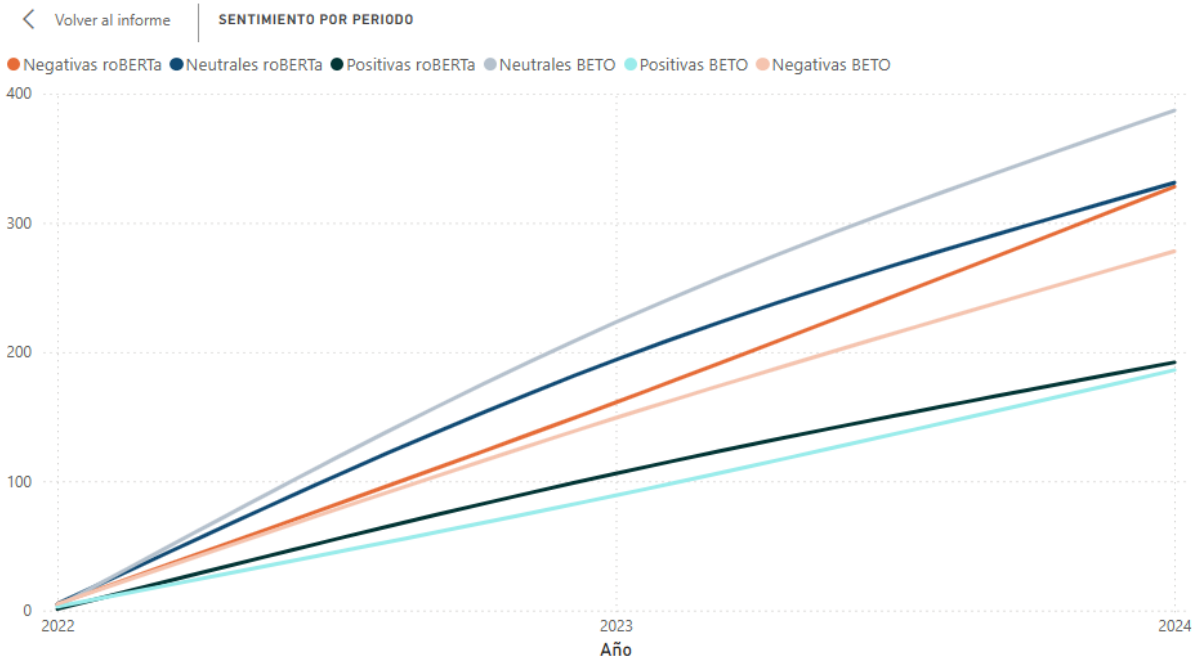

Figura 34. Evolución del sentimiento por Año

Al analizar la evolución del sentimiento por periodo, se evidencia un crecimiento general en todas las categorías entre 2022 y 2024; sin embargo, este crecimiento no es uniforme. En ambos modelos, BETO y RoBERTa, el sentimiento neutral inicialmente lidera en volumen, pero su curva muestra señales de desaceleración hacia 2024, indicando que el aumento de contenidos neutrales comienza a estabilizarse. En contraste, el sentimiento negativo, especialmente en el modelo RoBERTa, experimenta un crecimiento más acelerado, al punto de acercarse al volumen del sentimiento neutral. Este cruce de trayectorias sugiere que las percepciones críticas hacia la reforma a la salud se han intensificado en el periodo reciente. El sentimiento positivo, en cambio, mantiene un crecimiento mucho más moderado y constante en ambos modelos, consolidando su posición como el menos frecuente. En conjunto, la dinámica observada refleja que, si bien la neutralidad sigue siendo relevante, la crítica ha ganado fuerza relativa con el tiempo, configurando un panorama de opinión pública cada vez más polarizado.

10.4 ANALISIS POR REGIÓN

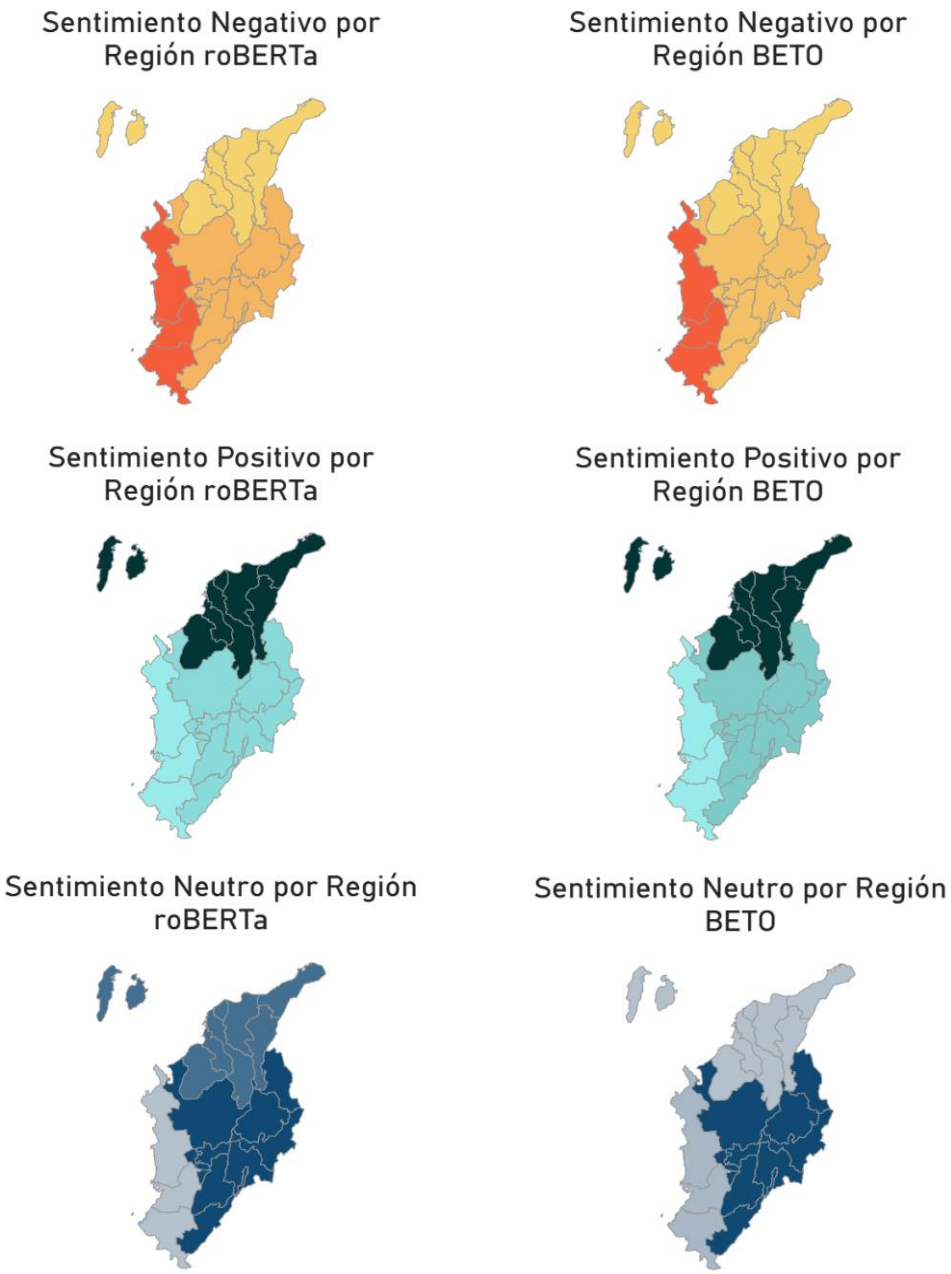

Figura 35. Distribución de sentimiento por región según del modelo

En la Figura 35 se observa el sentimiento por región basado en los modelos RoBERTa y BETO muestra una tendencia general a la predominancia de sentimientos neutros y negativos, con ligeras variaciones entre regiones. Según RoBERTa, la región Andina presenta el mayor número de sentimientos negativos (263 casos, 37,52%) y neutros (289 casos, 41,23%), seguida por el Caribe (157 negativos y 175 neutros) y el Pacífico (95 negativos y 76 neutros), con el Caribe liderando ligeramente en el porcentaje de positividad (25,39%). Por su parte, BETO confirma esta tendencia, aunque con algunas diferencias: el sentimiento negativo es ligeramente menor

en todas las regiones (32,97% en total), siendo el Pacífico el que muestra el porcentaje más alto de negatividad relativa (35,65%), mientras que Andina (47,22%) y Caribe (44,72%) mantienen altos niveles de neutralidad, y el Caribe nuevamente destaca en sentimientos positivos (23,15%). Visualmente, los mapas respaldan estos resultados, reflejando una mayor intensidad de sentimientos negativos y neutros especialmente en la región Andina y el suroccidente del país. En general, tanto RoBERTa como BETO coinciden en que la conversación regional se caracteriza más por la neutralidad y la negatividad, con una menor proporción de mensajes positivos, siendo la región Caribe la que presenta un matiz ligeramente más optimista en ambos modelos

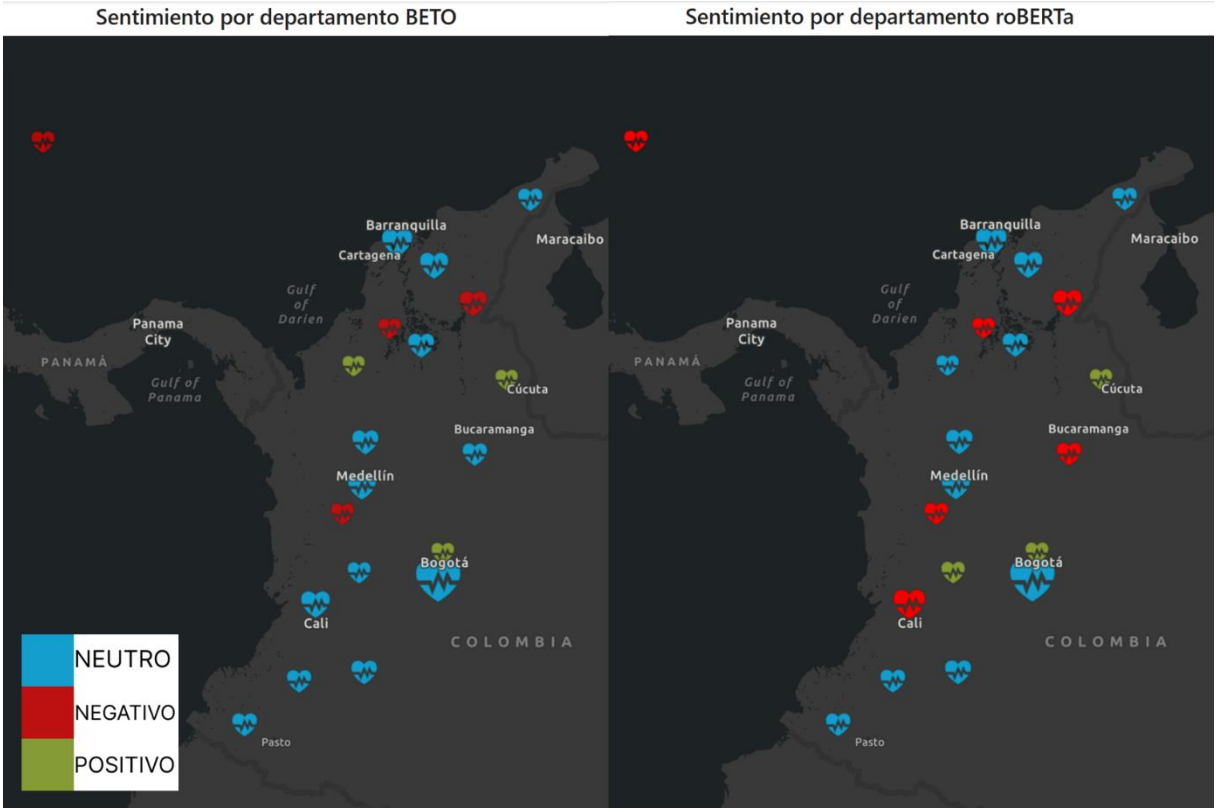

Figura 36. Distribución de sentimiento en los principales departamento

La Figura 36. muestra el sentimiento predicho por parte de los modelos BETO y RoBERTa mostraron coincidencias en clasificar la mayoría de las noticias como neutrales, aunque RoBERTa tendió a identificar más sentimientos negativos en comparación con BETO. Mientras BETO predijo emociones negativas principalmente en Cesar, Risaralda, San Andrés y Providencia, y Sucre, y positivas en Córdoba, Cundinamarca y Norte de Santander, RoBERTa detectó sentimientos negativos adicionales en Santander y Valle del Cauca, y positivos en Quindío. En los mapas, BETO mostró una distribución mayoritaria de sentimientos neutrales, mientras RoBERTa presentó más dispersión de emociones negativas, indicando que RoBERTa es más sensible a matices negativos. En general, ambos modelos son consistentes, pero RoBERTa resulta más crítico en la clasificación del sentimiento, lo que sugiere que BETO ofrece

una visión más neutral y RoBERTa una interpretación más severa del tono de las noticias.

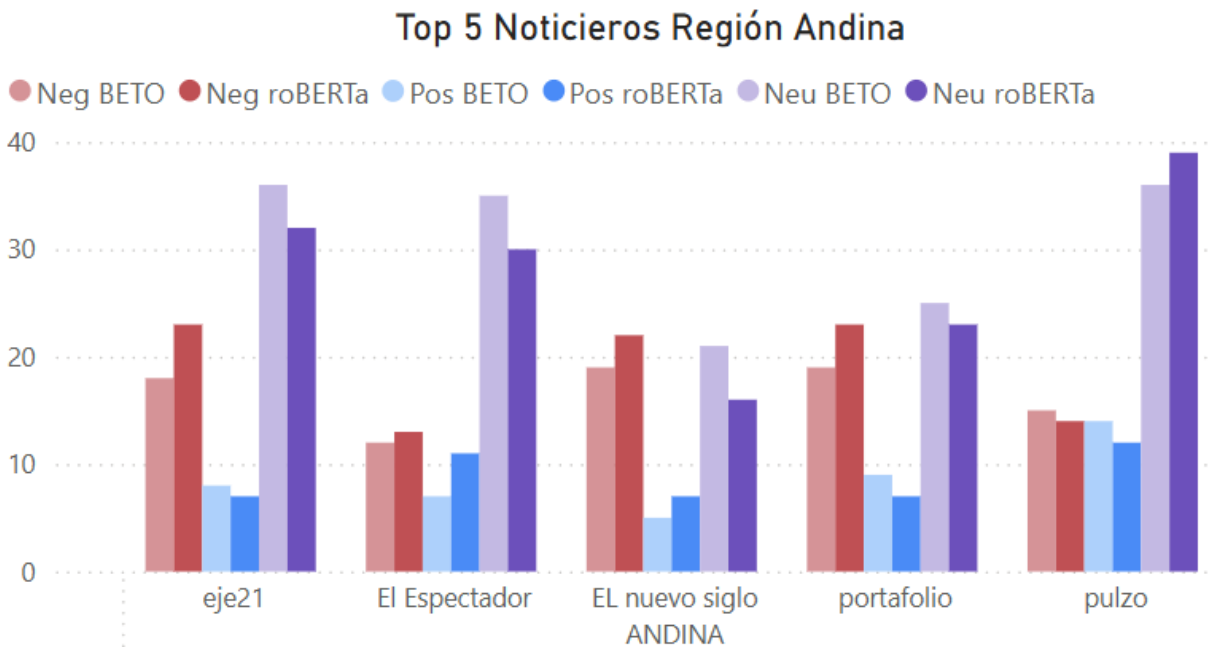

Figura 37. Top 5 de los noticieros en la Región Andina

En la Figura 37. se muestra el análisis de las distribuciones de sentimiento en la región Andina, el cual revela que la mayoría de las noticias se clasificaron como neutras, con valores que oscilaron entre 20 y 40 artículos. También se observó que RoBERTa asignó sistemáticamente un mayor número de etiquetas negativas que BETO, particularmente en portales como eje21 y Portafolio. Por su parte, BETO tendió a reclasificar parte del contenido que RoBERTa consideró negativo como neutral, lo cual sugiere que el umbral de decisión de BETO fue más conservador al identificar críticas.

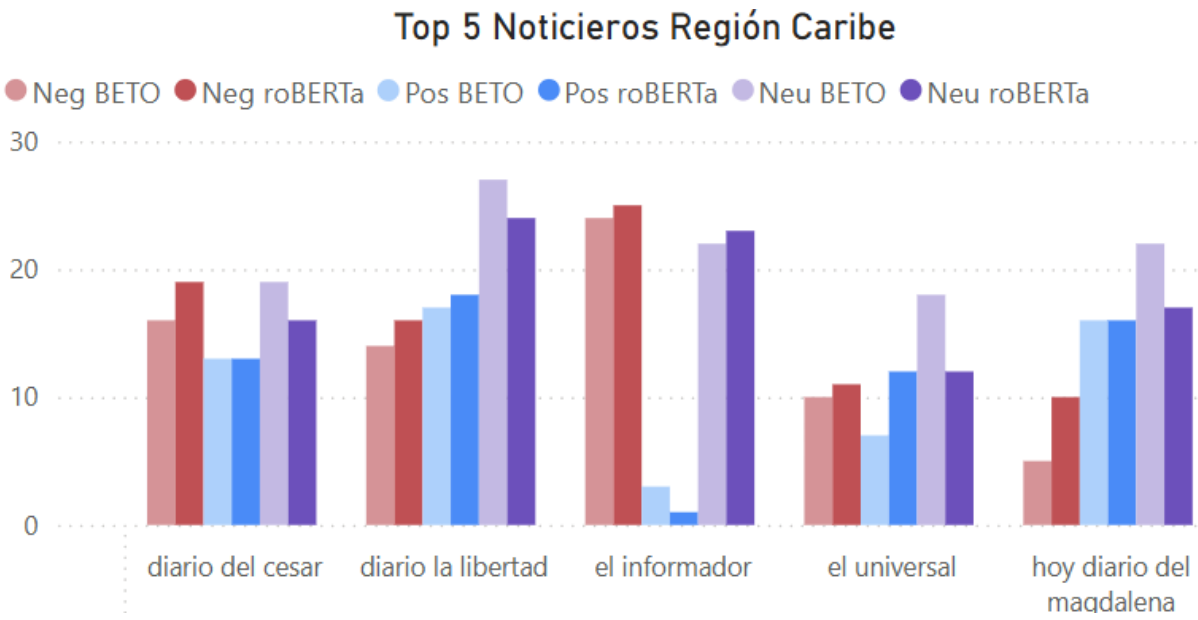

Figura 38. Top 5 de los noticieros en la Región Caribe

En la región Caribe Figura 38. los resultados mostraron una distribución más equilibrada entre neutralidad y negatividad. El portal El Informador presentó el pico más alto de noticias negativas en RoBERTa, mientras que BETO casi no detectó positivos en ese mismo medio. Se identificó también que Diario la Libertad concentró el porcentaje más alto de neutralidad, evidenciando diferencias sustanciales en la forma como cada modelo interpretó el tono de la cobertura regional.

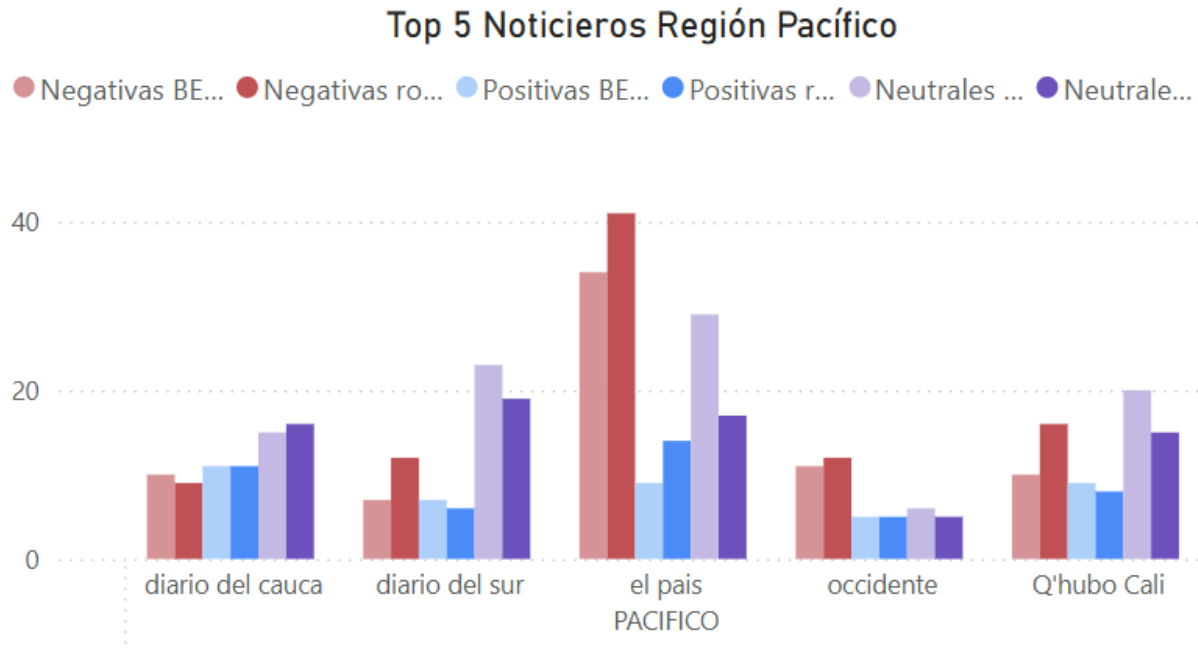

Figura 39. Top 5 de los noticieros en la Región Pacífico

Para la región Pacífico Figura 39. el volumen de contenido negativo fue especialmente

pronunciado en El País, con aproximadamente 40 noticias etiquetadas como negativas por RoBERTa y 35 por BETO. A pesar de ello, BETO registró un mayor número de neutrales en El País y Q'hubo Cali, contrastando con la posición más estricta de RoBERTa. El análisis de positivas confirmó que, en términos generales, ambos modelos fueron conservadores al asignar polaridad favorable, manteniendo los valores por debajo del 15 % del total de artículos en todos los portales.

Estos resultados repercutieron de manera directa en el entendimiento de la cobertura mediática de la reforma a la salud en Colombia. Al observar que la región Pacífico exhibió un marcado sesgo negativo con El País encabezando la crítica, que en la región Andina predominó la neutralidad en medios como Pulzo y Portafolio, y que en el Caribe se registró un balance más mixto, se puso de manifiesto la heterogeneidad de las dinámicas regionales. Estas diferencias reflejaron las realidades locales: en el Pacífico, las preocupaciones por la accesibilidad y la calidad del servicio generaron una narrativa más pesimista; en la Andina, la cercanía al epicentro político contribuyó a un tratamiento más moderado; y en el Caribe, la pluralidad de fuentes derivó en una cobertura más diversificada.

11.CONCLUSIONES

Como parte del desarrollo de este proyecto, se recolectaron sistemáticamente 1.401 documentos periodísticos mediante técnicas de web scraping, enfocados en noticias de las regiones Caribe, Andina y Pacífica relacionadas con la reforma al sistema de salud en Colombia. Para garantizar la calidad del corpus, se aplicó un proceso de preprocesamiento que incluyó la eliminación de publicidad y elementos irrelevantes, así como la remoción de stop words, tokenización y lematización.

Posteriormente, mediante un análisis estadístico basado en el rango intercuartílico (IQR), se identificaron y excluyeron documentos considerados atípicos, aquellos con más de 1.096,5 tokens, obteniendo así un corpus final más homogéneo compuesto por 1.362 documentos. Además, se realizó la etiquetación manual de un subconjunto representativo (360 documentos, el 26,43% del total), utilizando una rúbrica diseñada para clasificar tanto temáticamente como discursivamente los textos, sentando las bases metodológicas para las etapas de modelado y análisis posteriores.

Para la clasificación de sentimientos, se implementaron modelos de procesamiento de lenguaje natural (NPL) basados en arquitecturas encoder —BETO y RoBERTa—, así como el modelo decoder ChatGPT-4o. Inicialmente, sin ajuste fino, BETO y RoBERTa mostraron un desempeño limitado, con exactitudes de 53.41% y 46.35% respectivamente, evidenciando dificultades para captar tonos positivos y negativos en un dominio temático especializado como el de la reforma a la salud.

El proceso de fine-tuning resultó determinante: BETO alcanzó una accuracy de 91.29% y un F1-score de 91.26%, mientras que RoBERTa obtuvo una accuracy de 89.18% y un F1-score de 89.11%. BETO mostró un mejor equilibrio entre precisión y recall en todas las clases, especialmente en sentimientos positivos y neutrales, mientras que RoBERTa destacó en identificar emociones negativas, aunque con más errores en textos neutrales. Por otra parte, el análisis realizado con ChatGPT-4o, empleado sin ajuste específico, obtuvo una exactitud aceptable (67.29%) pero con tendencia a clasificar excesivamente los textos como negativos.

A partir de las predicciones, se diseñó un tablero interactivo en Power BI que permitió visualizar las tendencias por región. El análisis regional reveló diferencias importantes: en la región Andina predominó la neutralidad, posiblemente asociada a una cobertura más institucional; en el Caribe, aunque la neutralidad fue alta, se registró el mayor porcentaje de opiniones positivas; mientras que en el Pacífico destacó una cobertura marcadamente negativa, especialmente en medios como El País, lo cual sugiere un mayor malestar frente a las condiciones locales del sistema de salud.

El análisis temporal evidenció un aumento progresivo en el volumen de noticias entre 2022 y 2024, acompañado de una intensificación de los sentimientos negativos hacia el final del periodo, reflejando un ambiente mediático más crítico y polarizado.

Finalmente, mediante técnicas de análisis de similitud como TF-IDF, Doc2Vec y MpNet, se encontró una alta similitud interna en el tratamiento de la noticia dentro de cada región, aunque con ligeras variaciones en los matices de enfoque entre Caribe y Pacífico. Esto sugiere que, aunque existe un discurso relativamente homogéneo a nivel nacional, las diferencias socioculturales y geográficas también moldean la narrativa mediática.

En conjunto, los resultados confirman que el ajuste fino de modelos es crucial para alcanzar altos niveles de desempeño en tareas de análisis de sentimientos en contextos especializados y que BETO, en particular, emerge como una herramienta especialmente adecuada para el análisis mediático en español. Además, los resultados evidencian una narrativa mediática dominada por la neutralidad matizada con rasgos negativos y positivos debido a las diferencias regionales.

12.TRABAJOS FUTUROS

Este proyecto deja abiertas varias puertas para continuar explorando el análisis de medios en Colombia desde una perspectiva regional y basada en datos. La experiencia adquirida durante el desarrollo del estudio, tanto en la recolección y limpieza del corpus como en el uso de modelos de lenguaje ajustados al contexto, ofrece una base útil para futuras investigaciones y aplicaciones prácticas. En este sentido, una de las principales proyecciones consiste en ampliar el enfoque actual centrado en medios digitales hacia un análisis multicanal que incluya también la televisión y la radio, con el fin de contrastar las narrativas, estilos discursivos y enfoques editoriales presentes en diferentes formatos mediáticos; este contraste permitiría una comprensión más integral de cómo se construyen y difunden los discursos sobre las reformas sociales en el país. Como proyección inmediata, este trabajo tiene previsto convertirse en un artículo científico para ser sometido a una revista académica de alto impacto, lo que permitirá divulgar los hallazgos obtenidos, así como validar el enfoque metodológico y analítico ante la comunidad investigadora. La sistematicidad en la construcción del corpus, el riguroso proceso de etiquetado y el uso de modelos lingüísticos finamente ajustados en español constituyen aportes metodológicos significativos que pueden ser de utilidad para estudios futuros en análisis de medios, políticas públicas y procesamiento de lenguaje natural. Esta publicación también servirá como insumo para debates académicos sobre la representación mediática de reformas sociales en América Latina y el papel del lenguaje en la construcción de percepciones colectivas.

Por otra parte, se identifican varias líneas de investigación complementarias que podrían desarrollarse a partir de este estudio. Una posibilidad es extender el análisis a otras regiones del país o a otros temas de política pública, como la reforma pensional o educativa, para contrastar estilos discursivos y patrones de sentimiento. Asimismo, se sugiere aplicar esta metodología en estudios longitudinales con actualizaciones periódicas del corpus, lo que permitiría observar cómo evoluciona la narrativa mediática frente a los cambios en la agenda pública. Finalmente, este tipo de análisis puede tener una utilidad práctica para equipos de comunicación estratégica, partidos políticos o entidades gubernamentales que cuenten con analistas de información o percepción pública, ya que proporciona una herramienta robusta para monitorear, comprender y anticipar las dinámicas mediáticas en torno a temas sensibles y de alto impacto social.

Finalmente, una línea de mejora identificada a partir de este estudio consiste en ampliar progresivamente la base de datos etiquetada utilizada para el entrenamiento y validación de los modelos. Si bien el corpus total recolectado fue amplio (1.362 documentos), la muestra etiquetada manualmente incluyó 360 textos, una cantidad que permitió poner en marcha el proceso de modelado y evaluar la viabilidad del enfoque propuesto. Sin embargo, una cobertura más amplia contribuiría a mejorar la capacidad de generalización de los modelos,

reducir posibles sesgos y capturar de manera más representativa la diversidad temática y discursiva presente entre regiones y medios. En futuras investigaciones, se sugiere incorporar estrategias como el aprendizaje activo, el aumento de datos o la validación cruzada, que permitan escalar el proceso de etiquetado sin comprometer su calidad. Esta ampliación fortalecería la solidez del enfoque metodológico y habilitaría aplicaciones más robustas en el monitoreo automatizado de medios, el análisis de coyuntura y los estudios comparativos en diferentes contextos regionales o temáticos.

13.REFERENCIAS

[1] W. Atencia, J. Bustillo y J. Rambal, «Analizador de tweets asociados a la política y polarización Colombiana (Proyecto De Grado),» 2020. [En línea]. Available: https://manglar.uninorte.edu.co/bitstream/handle/10584/9280.

[2] F. Bonilla-Escobar y H. A. García-Perdomo, «Reforma a la salud en Colombia: oportunidades, tensiones y desafíos,» *Revista de Salud Pública,* vol. 25, pp. 1-3, 2023.

[3] M. E. Wills, «La salud en disputa: debates sobre el modelo de atención y las reformas en Colombia,» *Revista Colombiana de Sociología,* vol. 45, pp. 121-140, 2022.

[4] A. M. Martínez, «Medios y regiones: una relación desigual,» Fundación para la Libertad de Prensa (FLIP), Bogotá, 2017.

[5] C. A. Castro y J. Jiménez, «Medios de comunicación y opinión pública en Colombia: aproximaciones desde la ciencia política,» *Revista de Estudios Sociales,* vol. 77, pp. 15-30, 2021.

[6] CNMH, «Una verdad que circula: medios de comunicación y conflicto armado,» Centro Nacional de Memoria Histórica, Bogotá, 2015.

[7] M. I. Jordan, D. M. Blei y A. Y. Ng, «Latent Dirichlet Allocation,» *Journal of Machine Learning Research,* vol. 3, pp. 993-1022, 2003.

[8] J. Camacho-Collados y M. T. Pilehvar, «From Word to Sense Embeddings: A Survey on Vector Representations of Meaning,» *Journal of Artificial Intelligence Research,* vol. 63, p. 743–788, 2018.

[9] Steven Bird, Ewan Klein, Edward Loper, Natural Language Processing with Python, O'Reilly Media, 2009.

[10] M. Honnibal, I. Montani, Natural Language Understanding with Bloom Embeddings, Convolutional Neural Networks and Incremental Parsing, 2017.

[11] R. Rehurek, P. Sojka, Software Framework for Topic Modelling with Large Corpora, Malta: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, 2010.

[12] C. D. Manning, P. Raghavan, H. Schütze, Introduction to Information Retrieval, Cambridge University Press, 2008.

[13] F. Chollet, Deep Learning with Python, Manning Publications, 2017.

[14] Ronen Feldman ; James Sanger ., The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data, cambridge: Prensa de la Universidad de Cambridge, 2006.

[15] B. Liu, Sentiment Analysis: Mining Opinions, Sentiments, and Emotions, cambridge University Press, 2015.

[16] D. jurafsky; J. H. Martin, Speech and Language Processing, Prentice Hall, 2020.

[17] J. Marías, Léxico y semántica, Barcelona: mheducation.es, 2018.

[18] J. Moré, Cómo interpretar y analizar la informacion textual., Barcelona: Universidad Oberta de catalunya, 2019.

[19] C. Sammut, Encyclopedia of Machine Learning, Boston : School of Computer Science and Engineering, 2011.

[20] T. Mikolov y Q. Le, «Distributed Representations of Sentences and Documents,» *arXiv,* vol. 2, pp. 1405-1453, 2014.

[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gómez, Ł. Kaiser e I. Polosukhin,, Attention Is All You Need, California EE.UU: Advances in Neural Information Processing Systems , 2017.

[22] V. Tripathy, «Arquitectura Transformer en detalle,,» 2022. [En línea]. Available: https://aprendemachinelearning.com/arquitectura-transformer. [.

[23] T. Wolf et al, «Transformers: State-of-the-Art Natural Language Processing,» *Proceedings of EMNLP 2020,* 2020.

[24] H. Face, «Hugging Face,» [En línea]. Available: https://huggingface.co. [Último acceso: 16 06 2024].

[25] J. Devlin, M.-W. Chang, K. Lee y K. Toutanova, «BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,» *Proc. 2019 Conf. North American Chapter of the Association for Computational Linguistics (NAACL),* 2019.

[26] J. Devlin, M. Chang, K. Lee, y K. Toutanova, ERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, orth American Chapter of the Association for Computational Linguistics, 2019.

[27] orth American Chapter of the Association for Computational Linguistics, XLNet: Generalized Autoregressive Pretraining for Language Understanding, https://arxiv.org/abs/1906.08237, 2019.

[28] H. Face, sentence-transformers/paraphrase-multilingual-mpnet-base-v2 Model Card, Hugging Face Models Repository: https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2, 2021.

[29] J. D. M. &. J. H. Johnson, «Billion-scale similarity search with GPUs,» *IEEE Transactions on Big Data,* vol. 7, nº 3, p. 535–547, 2019.

[30] H. Face, «finiteautomata/beto-sentiment-analysis,» 2021. [En línea]. Available: https://huggingface.co/finiteautomata/beto-sentiment-analysis. [Último acceso: 2025].

[31] C. Cardellino, L. Rodríguez, M. Armentano y M. M. Pereyra, «Spanish Pre-Trained BERT Model and Evaluation Data,» 2019. [En línea]. Available: https://arxiv.org/pdf/1810.04805. [Último acceso: 2025].

[32] J. M. Pérez, «“RoBERTuito: a pre-trained language model for social media text in Spanish,,» *Proc. of the 13th Language Resources and Evaluation Conference (LREC), Marseille ,* 2022.

[33] J. Devlin, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Proc. of NAACL-HLT, 2019.

[34] Y. Liu, RoBERTa: A Robustly Optimized BERT Pretraining Approach, rXiv preprint arXi, 2019.

[35] J. &. R. S. Howard, «Universal language model fine-tuning for text classification,» *Association for Computational Linguistics.,* p. 328–339, 2018.

[36] A. S. N. P. N. Vaswani, «Attention Is All You Need.,» de *NeurIPS (Conference on Neural Information Processing Systems).*, 2017.

[37] T. M. B. R. N. Brown, «Language Models are Few-Shot Learners.,» *OpenAI,* 2020.

[38] S. Rahbar, «GPT-4 Omni — Under the Hood.,» *Medium,* 2024.

[39] G. M. Olguín, Métricas de similaridad, mexico: RITI Journal, 2019.

[40] M. a. H. Schütze, Foundations of Statistical Natural Language Processing, Cambridge: Cambridge, 1999.

[41] J. a. J. H. Martin, Speech and Language Processing, USA : Prentice Hall, 2022.

[42] A. Díaz, J. González, J. Peinado y M. Tellez, «Overview of TASS 2020: Introducing Emotion Detection,» *Proceedings of TASS 2020 at SEPLN,* 2020.

[43] M. d. S. y. P. social, «proyecto de Ley de Reforma al Sistema de Salud,» Ministerio de Salud y Protección Social, 2023. [En línea].

[44] C. d. l. R. d. Colombia, «Informe de ponencia para primer debate del Proyecto de Ley 339 de 2023 Cámara,» *Gaceta del Congreso,* 2023.

[45] J. T. y. L. Franco, «Revisión crítica del sistema de salud colombiano: avances y desafíos,» *Revista de Salud Pública,* 2022.

[46] D. N. d. P. (DNP), «Estudio de inequidades en el acceso a servicios de salud,» *Gobierno Nacional colombiano,* 2021.

[47] A. G. y. F. Ruiz, «Reforma a la salud en Colombia: una mirada técnica a sus riesgos e implicaciones,» *Economía & Desarrollo,* 2023.

[48] D. L. y. P. Martínez, «Controversias y tensiones en la reforma a la salud: una lectura política del debate público,» *Estudios Sociales Contemporáneos,* 2023.

[49] G. A. J. MARTÍNEZ, Proyecto de ley "Por medio del cual se transforma el Sistema de salud", bogota: Ministro de Salud y Protección Social, 2024.

[50] F. D. y. M. C. Andreu Casas, CONFLICTOS Y COBERTURA MEDIÁTICA, Dialnet, 2018.

[51] D. K. y. A.-h. P. D. Park, Agendas on Nursing in South Korea Media: Natural Language Processing and Network Analysis of News From 2005 to 2022, Journal of Medical Internet Research, 2024.

[52] A. Ruelens, Analyzing user-generated content using natural language processing: a case study of public satisfaction with healthcare systems, Journal of Computational Social Science, 2022.

[53] Y. C. L. Y. y. J. W. M. Chu, Language interpretation in travel guidance platform: Text mining and sentiment analysis of TripAdvisor reviews, Frontiers in Psychology, 2022.

[54] A. Flores Pastor, Análisis de similitud semántica de tweets sobre sostenibilidad utilizando modelos BERT pre-entrenados y técnicas de generación de embeddings, Universitat Rovira i Virgili., 2023.

[55] N. G. y. V. Elangovan, COMPARISON OF DOCUMENT SIMILARITY, International Journal of Artificial Intelligence and Applications, 2023.

[56] W. Atencia, J. Bustillo y J. Rambal, «Analizador de tweets asociados a la política y polarización Colombiana (Proyecto De Grado),» 2020. [En línea]. Available: https://manglar.uninorte.edu.co /bitstream/handle/10584/9280. [Último acceso: 2024].

[57] R. d. l. española, Artist, *Análisis de Scimago Journal & Country Rank.* [Art]. Universidad Metropolitana del Ecuador, 2019.

[58] S. Bird, E. Klein, E. Loper, Natural Language Processing with Python, O'Reilly Media, 2009.

[59] R. N y G. I, «Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks,» *arXiv,* vol. 1, pp. 1908-10084, 2019.

[60] x. Wang y J. Chang, «FastText and Word2Vec: Comparative Analysis,» *Journal of Machine Learning Research (JMLR),* vol. 22, pp. 1-23, 2021.

[61] J. Cañete, G. Chaperon, R. Fuentes y J. Pérez, «Spanish Pretrained BERT Model and Evaluation Data,» *Proceedings of the PML4DC at ICLR,* 2020.

[62] OpenAI, «GPT-4 Technical Report,» *arxiv,* vol. 6, pp. 1-100, 2024.

[63] OpenAI, «OpenAI developer platform,» OpenAI, 2023. [En línea]. Available: https://platform.openai.com/docs/overview. [Último acceso: 25 1 2025].

[64] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 1810.04805: arXiv preprint arXiv, 2018.

[65] Luis Lira; Bolívar Quiroga, «Instituto Latinoamericano y del Caribe de Planificación,» Marzo 2009. [En línea]. Available: https://repositorio.cepal.org/server/api/core/bitstreams/c6aa2d07-862d-465e-b3d1-a1f8259e367a/content. [Último acceso: Junio 2024].

[66] M. Chu , Y. Chen , L. Yang y J. Wang, «Language interpretation in travel guidance platform: Text mining and sentiment analysis of TripAdvisor reviews,» *Frontiers in Psychology,* pp. 1-8, 2022.

[67] diccionariomarketing, «diccionariomarketing,» Rush Compta SASU, 21 10 2023. [En línea]. Available: https://diccionariomarketing.es/definition/cobertura-mediatica/. [Último acceso: Junio 2024].

[68] N. Gahman y V. Elangovan, «A COMPARISON OF DOCUMENT SIMILARITY,» *International Journal of Artificial Intelligence and Applications (IJAIA),* vol. 14, nº 2, pp. 41-50, 2023.

14.ANEXOS

1. web scraping y preprocesamiento de los datos https://github.com/bryanshm/Proyecto_Aplicado/blob/workspace_bryan/extraccion_y_limpieza.ipynb
2. Análisis estadístico https://github.com/bryanshm/Proyecto_Aplicado/blob/workspace_bryan/Analisis_estadistico.ipynb
3. Modelo.TF-IDF https://github.com/bryanshm/Proyecto_Aplicado/blob/main/Modelo_TF_IDF.ipynb
4. Modelo Doc2Vec https://github.com/bryanshm/Proyecto_Aplicado/blob/workspace_bryan/analisis_modelo_Doc2Vec.ipynb
5. Modelo sentence-transformers/paraphrase-multilingual-mpnet-base-v2 https://github.com/bryanshm/Proyecto_Aplicado/blob/workspace_bryan/Analisis_comparativo_modelos_avanzados.ipynb
6. Modelo finiteautomata/beto-sentiment-analysis y ChatGPT. https://github.com/bryanshm/Proyecto_Aplicado/blob/workspace_bryan/analisis_de_sentimientos.ipynb