Modelo de Procesamiento de Lenguaje Natural para Análisis de Sentimientos y Extracción de Entidades en Reseñas de Aplicaciones Móviles en Colombia

Ana Gabriela Hernández Peña y Carlos Mateo Vera Grimaldo

Trabajo de Grado para Optar el Título de Ingeniero de Sistemas

Director

Laura Viviana Galvis Carreño

PhD. en Ingeniería Eléctrica y Computación

Universidad Industrial de Santander

Facultad de Ingenierías Físicomecánicas

Escuela de Ingeniería de Sistemas

Bucaramanga

2026

Dedicatoria

A mis padres, por ser raíz y refugio; en realidad, el triunfo es de ustedes.

A mi adorado gato Tommy, porque con su presencia me enseñó que, aun en los días más grises, habita la posibilidad de la luz.

Ana Gabriela Hernández Peña

Agradecimientos

A mis amados padres, Yolanda Peña y Oscar Hernández, por ser ejemplo de perseverancia; por enseñarme que incluso en la adversidad florece la fuerza, y por permanecer cuando todo en mí tambaleaba. Este triunfo es, en esencia, un reflejo del amor y la fortaleza que sembraron en mí.

A mi compañero de investigación y de vida, Mateo Vera, por caminar a mi lado en cada reto; por su paciencia infinita y su voz alentadora, y por hacer del amor una forma de crecer juntos.

A mi directora, Laura Viviana Galvis, por ser guía durante el proceso y por su disposición constante y valioso acompañamiento.

A la profesora Sonia Cristina Gamboa, por ser mi mentora durante mi paso por la universidad; mujer que me inspira con su entrega, su sabiduría y su calidez humana; por creer en mí y por dejar una huella imborrable en mi camino.

A la Universidad Industrial de Santander, por ser mi laboratorio de ideas; el escenario donde crecí, me equivoqué y volví a intentar, y donde aprendí a transformar la curiosidad en conocimiento.

Ana Gabriela Hernández Peña

Dedicatoria

A mi mamá, por ser la mujer más fuerte y valiente del mundo. Todo lo que soy es gracias a ella y para ella.

A mi papá, por trabajar todos los días para sacarme adelante. Este logro, sin duda, es uno de los frutos de todo su esfuerzo.

A Milu y Tara, por ser mis más fieles compañeras y lo que más amo en este mundo.

Carlos Mateo Vera Grimaldo

Agradecimientos

A mis papás, por ser el motor que me impulsa a perseguir mis sueños y por tanto amor que me han brindado.

A mis abuelos, por haber sido mi hogar en esta aventura.

A mis perros, Milu, Tara, Queso y Jack, porque son en lo primero que pienso cuando me levanto y en lo último cuando voy a dormir.

A mi hermano, Nicolás, por llenarme de risas y enojos, porque sin esto la vida se sentiría un poco aburrida.

A mi compañera de trabajo y de vida, Ana, por ser esa luz que ilumina mi sendero todos los días.

A don Oscar y a doña Yolanda, por acogerme en su familia y hacerme sentir parte de ella.

A nuestra directora, Laura Galvis, por haber creído en nosotros y acompañarnos durante este proceso.

A la profesora Sonia Gamboa, por creer en mis capacidades.

A mis amigos, compañeros de estudio y de trabajo, por todos los momentos compartidos, las risas y todas las enseñanzas que han aportado para mi formación como profesional.

A Cristiano Ronaldo, por ser figura de perseverancia y mi inspiración desde pequeño.

Carlos Mateo Vera Grimaldo

Tabla de Contenido

Lista de Tablas

Lista de Figuras

Lista de Apéndices

Resumen

Título: Modelo de Procesamiento de Lenguaje Natural para Análisis de Sentimientos y Extracción de Entidades en Reseñas de Aplicaciones Móviles en Colombia*

Autor: Ana Gabriela Hernández Peña, Carlos Mateo Vera Grimaldo**

Palabras clave: Análisis de sentimientos, procesamiento de lenguaje natural, análisis de sentimientos por aspecto, reseñas de aplicaciones móviles, inteligencia artificial, español colombiano

Descripción: El crecimiento acelerado de las aplicaciones móviles ha generado un alto volumen de reseñas de usuarios en plataformas digitales, las cuales constituyen una fuente clave para comprender la percepción y la experiencia de uso. En el contexto colombiano, estas reseñas presentan particularidades lingüísticas que dificultan su análisis automático mediante enfoques tradicionales. Aunque el análisis de sentimientos (SA, del inglés *Sentiment Analysis*) permite clasificar opiniones de forma general, resulta insuficiente para identificar con precisión los elementos específicos de las aplicaciones que influyen en la satisfacción del usuario.

En este trabajo se propone el diseño e implementación de un modelo de análisis de sentimientos a nivel de aspecto (ABSA, del inglés *Aspect-Based Sentiment Analysis*), con un enfoque centrado en evidencia textual, orientado al procesamiento de reseñas de aplicaciones móviles en español de Colombia. El modelo emplea un esquema de anotación basado en cuatro dimensiones: fragmento textual de evidencia, categoría, aspecto y sentimiento, lo que permite identificar los fragmentos textuales relevantes, los componentes de la aplicación a los que hacen referencia y la polaridad asociada.

En un contexto caracterizado por la escasez de datos etiquetados, este trabajo aborda el análisis mediante estrategias metodológicas adecuadas para escenarios de baja disponibilidad de recursos, lo que permite generar información estructurada sobre la percepción de los usuarios. A partir de estos resultados, se desarrolla un prototipo web que facilita su visualización e interpretación. De este modo, la propuesta contribuye al fortalecimiento del procesamiento de lenguaje natural (NLP, del inglés *Natural Language Processing*) en español colombiano y ofrece una herramienta práctica para el análisis de reseñas de aplicaciones móviles.

---

* Trabajo de Grado

** Facultad de Ingenierías Físicomecánicas. Escuela de Ingeniería de Sistemas. Director: Laura Viviana Galvis Carreño. PhD. en Ingeniería Eléctrica y Computación

Abstract

Title: Natural Language Processing Model for Sentiment Analysis and Entity Extraction in Mobile Application Reviews in Colombia*

Author: Ana Gabriela Hernández Peña, Carlos Mateo Vera Grimaldo**

Key words: Sentiment analysis, natural language processing, sentiment analysis by aspect, mobile app reviews, artificial intelligence, Colombian Spanish

Description: The rapid growth of mobile applications has led to a high volume of user reviews on digital platforms, which constitute a key source for understanding user perception and experience. In the Colombian context, these reviews exhibit linguistic particularities that hinder their automatic analysis using traditional approaches. Although sentiment analysis (SA) allows for the general classification of opinions, it is insufficient to accurately identify the specific elements of applications that influence user satisfaction.

This work proposes the design and implementation of an aspect-based sentiment analysis (ABSA) model, with a text-evidence-centered approach, aimed at processing mobile application reviews in Colombian Spanish. The model employs an annotation scheme based on four dimensions: evidence *span*, category, aspect, and sentiment, enabling the identification of relevant textual fragments, the application components they refer to, and the associated polarity.

In a context characterized by the scarcity of labeled data, this work addresses the analysis through methodological strategies suitable for low-resource scenarios, enabling the generation of structured information about user perception. Based on these results, a web-based prototype is developed to facilitate their visualization and interpretation. In this way, the proposal contributes to strengthening Natural Language Processing (NLP) in Colombian Spanish and offers a practical tool for the analysis of mobile application reviews.

* Degree Work

** Faculty of Physical and Mechanical Engineering. School of Systems Engineering. Director: Laura Viviana Galvis Carreño. PhD in Electrical and Computer Engineering

Introducción

En los últimos años, el mundo se ha orientado de manera creciente hacia la experiencia del consumidor; en este contexto, comprender cómo se percibe un producto o servicio resulta fundamental para las organizaciones que buscan mejorar la calidad de su oferta (Chris et al., 2024; Gunathilaka y De Silva, 2022; Guzman y Maalej, 2014). Una herramienta ampliamente utilizada para este fin es el análisis de sentimientos, entendido como una tarea de inferencia que, apoyada en la inteligencia artificial (AI, del inglés *Artificial Intelligence*), permite identificar, extraer y cuantificar automáticamente las valoraciones expresadas en las opiniones de los usuarios (Bose et al., 2019).

En particular, este trabajo se enfoca en reseñas de aplicaciones móviles obtenidas de la tienda *Google Play*, las cuales constituyen una fuente valiosa para analizar la percepción de los usuarios (Guzman y Maalej, 2014). La elección de este dominio se sustenta en la alta penetración de los dispositivos móviles en Colombia, donde el 97,7 % de los usuarios de Internet mayores de 16 años accede a la red a través de un *smartphone* (Branch, 2025), consolidando a las aplicaciones móviles como uno de los principales canales de interacción digital.

Si bien el análisis de sentimientos tradicional ha permitido clasificar opiniones en polaridades generales como positivas, negativas o neutras, este enfoque resulta limitado cuando se busca comprender en detalle los factores específicos que influyen en la percepción del usuario (Pontiki et al., 2014). En respuesta a esta limitación, surge el ABSA, el cual permite identificar no solo la polaridad de una opinión, sino también los aspectos concretos del producto o servicio a los que esta se refiere, junto con los fragmentos textuales que sustentan dicha valoración (Schouten y Frasincar, 2015).

No obstante, en el contexto colombiano, la aplicación de técnicas de ABSA enfrenta un reto importante: la escasez de conjuntos de datos anotados en español de Colombia. Esta limitación dificulta el desarrollo de modelos capaces de capturar adecuadamente las particularidades lingüísticas y semánticas del contexto local.

Con el fin de abordar esta problemática, en el presente trabajo se propone el desarrollo de un modelo de ABSA basado en Transformadores (del inglés, *Transformers*), fundamentado en un esquema de etiquetado propio diseñado para adaptarse a la naturaleza de las reseñas analizadas. Este esquema permite anotar cada instancia mediante la identificación de un fragmento de evidencia textual, junto con la asignación de una categoría funcional, un aspecto específico y su respectiva polaridad, ofreciendo una representación más estructurada de la información.

De manera complementaria, se construyó un conjunto de datos propio en español colombiano, compuesto por reseñas provenientes de diferentes aplicaciones y dominios, el cual fue anotado siguiendo el esquema propuesto. Este recurso no solo permitió el entrenamiento y evaluación del modelo desarrollado, sino que también constituye un aporte relevante para la investigación en NLP en contextos de recursos limitados.

Como parte de la validación del enfoque planteado, se desarrolló un prototipo web que permite la interacción con el modelo de inferencia, facilitando la visualización de los resultados de clasificación, incluyendo los fragmentos de evidencia, las categorías, los aspectos y la polaridad asociada a cada reseña. Este componente aporta una dimensión aplicada al proyecto, al transformar los resultados del modelo en una herramienta accesible para el análisis de la percepción de los usuarios.

En coherencia con lo anterior, la pregunta de investigación que orienta el desarrollo de este trabajo es:

¿Cómo integrar diferentes herramientas de NLP para la clasificación de reseñas de usuario según su polaridad y entidades abordando la limitación de datos debidamente etiquetados, en busca de mejorar la experiencia y calidad de prestaciones de aplicaciones móviles en el contexto colombiano?

1. Objetivos

1.1 Objetivo general

Desarrollar un modelo que integre el reconocimiento de entidades y el análisis de sentimientos para clasificar reseñas de aplicaciones móviles en español, mediante herramientas de procesamiento de lenguaje natural, con el fin de extraer información relevante sobre la experiencia del usuario y generar conocimiento aplicable al contexto colombiano.

1.2 Objetivos específicos

Recolectar y preprocesar un corpus de reseñas de aplicaciones móviles en español del contexto colombiano para el entrenamiento y la validación del modelo.

Implementar un modelo de clasificación de reseñas que integre técnicas de procesamiento de lenguaje natural en escenarios de datos limitados.

Evaluar el rendimiento del modelo de procesamiento y análisis de sentimientos por aspecto y polaridad utilizando el corpus de reseñas creado.

Implementar un prototipo de ambiente web que permita visualizar los resultados del modelo de clasificación de reseñas para aplicaciones móviles específicas.

2. Marco referencial

2.1 Marco teórico

El presente marco reúne los fundamentos teóricos, conceptuales y analíticos que sustentan la investigación, integrando los principales enfoques relacionados con la experiencia de usuario, el procesamiento de lenguaje natural, el análisis de sentimientos, el análisis basado en aspectos y las

técnicas avanzadas de aprendizaje profundo. Esta organización permite establecer una progresión conceptual desde la motivación del problema hasta los enfoques metodológicos que lo hacen abordable en la práctica.

2.1.1 Experiencia de Usuario

La experiencia de usuario (UX, del inglés *User Experience*) constituye un concepto central en el diseño y evaluación de productos digitales y servicios interactivos. Se entiende como el conjunto de percepciones, emociones y respuestas que una persona experimenta al interactuar con un producto o sistema, o incluso al anticipar su uso (Organización Internacional de Normalización (ISO, 2019)). Estas percepciones están influenciadas por factores como la usabilidad, la utilidad, el rendimiento, la estética, la accesibilidad y el contexto de uso. En entornos digitales contemporáneos, los usuarios esperan interacciones fluidas, eficientes e intuitivas; por ello, fallas en el funcionamiento, tiempos de respuesta elevados o diseños poco claros suelen generar frustración y abandono de las aplicaciones.

En Colombia, donde los dispositivos móviles representan el principal medio de acceso a internet y las aplicaciones móviles cumplen un rol fundamental en ámbitos como la comunicación, los servicios, la educación y el entretenimiento, la UX se convierte en un factor crítico para la satisfacción y fidelización de los usuarios (Branch, 2025; Hassenzahl y Tractinsky, 2006). En este contexto, las reseñas publicadas en plataformas como *Google Play Store* constituyen una fuente de información valiosa para comprender la experiencia real de los usuarios (Guzman y Maalej, 2014).

Aunque las calificaciones numéricas ofrecen una medida cuantitativa general, el contenido textual de las reseñas proporciona información más detallada. En muchos casos, existe una discrepancia entre la puntuación otorgada y el sentimiento expresado en el texto, lo que evidencia la necesidad de realizar análisis más profundos que permitan interpretar adecuadamente la UX (Šmíd y Král, 2025). No obstante, en el contexto colombiano persisten desafíos relevantes, como la escasez de estudios y conjuntos de datos locales (Šmíd y Král, 2025), la presencia de

variaciones lingüísticas, el uso de expresiones coloquiales y la formulación de críticas implícitas, factores que dificultan la interpretación automática de la UX y resaltan la importancia de enfoques computacionales adaptados al dominio y al contexto cultural.

2.1.2 Procesamiento de Lenguaje Natural

A partir de la relevancia del contenido textual como fuente principal de información sobre la UX, el NLP se constituye como el conjunto de técnicas computacionales orientadas a permitir que las máquinas comprendan, interpreten y generen lenguaje humano (Jurafsky y Martin, 2026). En el análisis de reseñas de aplicaciones móviles, el NLP resulta fundamental para transformar grandes volúmenes de texto no estructurado en información organizada y utilizable (Feldman y Sanger, 2006), permitiendo identificar patrones lingüísticos, clasificar opiniones y extraer significados relevantes más allá de la simple frecuencia de palabras (Liu, 2012).

El NLP actúa, por tanto, como el marco general que posibilita el análisis automático de opiniones de usuarios, sirviendo como base para tareas más específicas orientadas a la interpretación de la percepción y las valoraciones expresadas en el texto (Šmíd y Král, 2025).

2.1.3 Análisis de Sentimientos

Dentro del NLP, el análisis de sentimientos es una de las tareas más utilizadas, ya que busca determinar la polaridad de un texto, generalmente clasificada como positiva, negativa o (Liu, 2012). Esta técnica permite estimar la percepción general de los usuarios sobre un producto o servicio, ofreciendo una visión agregada del sentimiento expresado en una reseña.

No obstante, presenta importantes limitaciones, dado que identifica únicamente la polaridad global del texto, sin precisar a qué componente específico del objeto evaluado se refiere dicha valoración. En el contexto de las aplicaciones móviles, donde una misma reseña puede incluir múltiples opiniones sobre diferentes funcionalidades o características, este enfoque resulta insuficiente para comprender de manera detallada los factores que influyen en la UX (Chris et al., 2024).

2.1.4 Análisis de Sentimientos Basado en Aspectos

Con el propósito de superar esta limitación, surge el ABSA, un enfoque más granular dentro del NLP. A diferencia del análisis tradicional, el ABSA permite identificar aspectos específicos del objeto de estudio y determinar la polaridad asociada a cada uno de ellos, posibilitando una representación estructurada y detallada de las opiniones (Pontiki et al., 2014; Šmíd y Král, 2025).

Este enfoque resulta especialmente útil en el análisis de reseñas de aplicaciones móviles, ya que una misma opinión puede expresar valoraciones distintas sobre diferentes componentes del sistema. Por ejemplo, un usuario puede manifestar una percepción positiva respecto a la facilidad de uso de una aplicación, mientras expresa una opinión negativa sobre su rendimiento o seguridad. El ABSA permite capturar estas diferencias, ofreciendo una visión más precisa de la UX y facilitando la identificación de aspectos específicos de mejora.

Desde una perspectiva metodológica, el ABSA no constituye una única tarea, sino un conjunto de subtareas que pueden abordarse de manera independiente o conjunta, conforme a la formulación del problema y los recursos disponibles (Zhang et al., 2018). Estas subtareas abarcan desde la extracción de aspectos y opiniones hasta formulaciones compuestas que integran múltiples componentes en una sola estructura predictiva. La Tabla 1 presenta las principales subtareas identificadas en la literatura (Hua et al., 2024).

Tabla 1
Principales subtareas del ABSA

Subtarea	Descripción
Extracción de Términos de Aspecto (AE, del inglés *Aspect Term Extraction*)	Identifica explícitamente en el texto los términos que hacen referencia a atributos o componentes del objeto evaluado.

Continúa en la siguiente página

Tabla 1 – *continuación*

Subtarea	**Descripción**
Extracción de Términos de Opinión (OE, del inglés *Opinion Term Extraction*)	Detecta las expresiones lingüísticas que manifiestan una opinión o juicio de valor asociado a un aspecto.
Clasificación de Sentimiento a Nivel de Aspecto (ASC, del inglés *Aspect Sentiment Classification*)	Determina la polaridad (positiva, negativa o neutra) asociada a un aspecto previamente identificado.
Detección de Categorías de Aspecto (ACD, del inglés *Aspect Category Detection*)	Identifica la categoría general a la que pertenece un aspecto mencionado en el texto, usualmente dentro de un conjunto predefinido.
Análisis de Sentimiento por Categoría de Aspecto (ACSA, del inglés *Aspect Category Sentiment Analysis*)	Asigna una polaridad a cada categoría de aspecto detectada en la reseña.
Extracción de Pares Aspecto-Opinión (AOPE, del inglés *Aspect-Opinion Pair Extraction*)	Identifica simultáneamente los aspectos y las expresiones de opinión asociadas, estableciendo la relación entre ambos elementos.
Co-Extracción de Aspecto-Polaridad (APCE, del inglés *Aspect Polarity Co-Extraction*)	Extrae conjuntamente los aspectos y la polaridad correspondiente sin separar explícitamente las subtareas.
Extracción de Tripletas Aspecto-Sentimiento (ASTE, del inglés *Aspect Sentiment Triplet Extraction*)	Identifica de manera conjunta el aspecto, el término de opinión y la polaridad asociada, formando una tripleta estructurada.
Extracción o Predicción de Cuádruplas Aspecto-Sentimiento (ASQE, del inglés *Aspect Sentiment Quadruple Extraction*)	Extiende la formulación anterior incorporando una categoría de aspecto adicional, generando una estructura compuesta por aspecto, categoría, opinión y polaridad.
ABSA de Extremo a Extremo (E2E ABSA, del inglés *End-to-End Aspect-Based Sentiment Analysis*)	Modelo unificado que resuelve múltiples subtareas de ABSA de manera conjunta mediante un único sistema, capturando dependencias entre ellas.

Nota. Subtareas fundamentales y compuestas comúnmente abordadas en el ABSA, según la literatura reciente.

2.1.5 Aprendizaje Profundo

El aprendizaje profundo (DL, del inglés *Deep Learning*) es una subárea de la inteligencia artificial y del aprendizaje automático que se caracteriza por el uso de redes neuronales con múltiples capas para la extracción y representación jerárquica de patrones complejos en los datos (Goodfellow et al., 2016). A diferencia de los modelos tradicionales, que dependen de características diseñadas manualmente, el aprendizaje profundo permite aprender representaciones automáticamente a partir de grandes volúmenes de información, mejorando la capacidad de generalización y el desempeño en tareas no lineales (LeCun y Hinton, 2015).

En el ámbito del NLP, el aprendizaje profundo ha supuesto un avance significativo al permitir el desarrollo de modelos capaces de capturar relaciones semánticas, sintácticas y contextuales con mayor precisión (Young et al., 2018). Arquitecturas como las redes neuronales recurrentes (RNN, del inglés *Recurrent Neural Networks*), las redes de memoria a largo plazo (LSTM, del inglés *Long Short-Term Memory*) y, más recientemente, los modelos basados en *Transformers* han demostrado resultados sobresalientes en tareas como la clasificación de textos, el análisis de sentimientos y la extracción de información (Vaswani et al., 2017; Wang et al., 2016).

En particular, los modelos de lenguaje preentrenados basados en *Transformers* han mostrado una alta efectividad en tareas granulares como el ABSA y la identificación de *spans* relevantes, incluso en contextos con expresiones ambiguas, irónicas o coloquiales (Devlin et al., 2019; Hua et al., 2024). Además, técnicas como el aprendizaje por transferencia (del inglés, *transfer learning*) y el ajuste fino (del inglés, *fine-tuning*) permiten reutilizar conocimiento adquirido en dominios con abundancia de datos y adaptarlo a escenarios con recursos limitados, lo que resulta especialmente relevante para el análisis de reseñas en español colombiano (Devlin et al., 2019; Ruder et al., 2019).

2.1.6 Aumento de datos

La técnica de aumento de datos (del inglés, *data augmentation*) consiste en generar nuevas muestras a partir de datos existentes, preservando sus características semánticas relevantes con el objetivo de incrementar la diversidad de los conjuntos de datos, mejorar la generalización de modelos de DL y mitigar problemas asociados al desbalance de clases y la escasez de datos anotados (Sennrich et al., 2016; Wei y Zou, 2019).

En el contexto del NLP, el aumento de datos presenta desafíos particulares, dado que transformaciones simples pueden alterar el significado del texto. Por esta razón, se han propuesto diversas estrategias como sustitución léxica, inserción o eliminación de palabras, traducción automática inversa (del inglés, *back-translation*), las cuales buscan mantener la coherencia semántica de las muestras generadas (Feng et al., 2021).

2.2 Estado del arte

El ABSA se ha consolidado como una de las líneas de investigación más relevantes dentro del NLP, debido a su capacidad para ofrecer interpretaciones estructuradas de las opiniones expresadas en textos (Hoang et al., 2019). La literatura ha mostrado una evolución progresiva desde enfoques basados en reglas y modelos clásicos hacia arquitecturas de DL, particularmente aquellas fundamentadas en *Transformers*. De manera paralela, han surgido nuevas formulaciones del problema que integran tareas de extracción, clasificación y alineación de opiniones, dando lugar a enfoques más integrales como los modelos generativos y los métodos basados en *spans*. En esta sección se revisan las principales tendencias y aportes relevantes para el presente trabajo.

2.2.1 Modelos basados en transformadores

La introducción de la arquitectura *Transformer* representó un punto de inflexión en el desarrollo de sistemas de NLP, al permitir modelar dependencias contextuales de largo alcance mediante

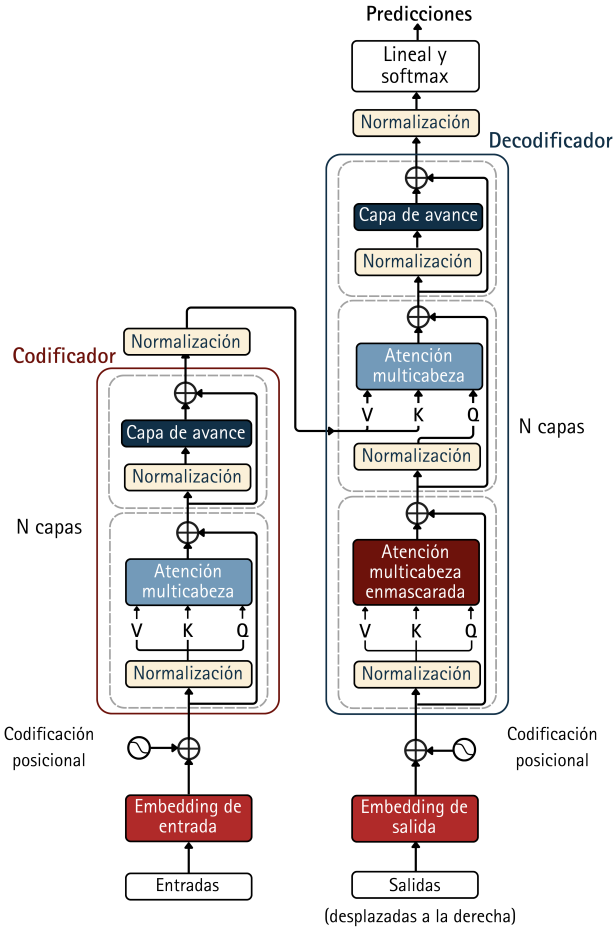
mecanismos de auto-atención (del inglés, *self-attention*). Un *Transformer* estándar combina capas de *encoder* y *decoder*, cada una con bloques de atención multi-cabeza (del inglés, *multi-head*). En particular, el mecanismo de atención empleado corresponde a la denominada atención de producto punto escalado (del inglés, *Scaled Dot-Product*), la cual se define como:

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

donde Q, K y V representan las matrices de consultas, claves y valores, y d_k es la dimensión de las claves. Este mecanismo permite que cada *token* asigne pesos de relevancia a todos los demás *tokens* de la secuencia, capturando relaciones contextuales de manera explícita (Vaswani et al., 2017). Una representación visual de la arquitectura general de un *Transformer* se puede observar en la Figura 1.

Figura 1

Diagrama de la arquitectura estándar del Transformer



Nota. Adaptado de Vaswani et al., 2017. La figura presenta la arquitectura *Transformer* estándar.

Los modelos preentrenados basados en la arquitectura *Transformer*, como BERT y sus variantes, han establecido el estado del arte en múltiples tareas de NLP, incluyendo clasificación de texto y análisis de sentimientos (Lin et al., 2022). En particular, BETO constituye una adaptación de BERT entrenada exclusivamente sobre corpus en español, lo que le permite capturar con mayor precisión las regularidades lingüísticas propias de este idioma (Cañete et al., 2020).

A nivel arquitectónico, BETO implementa únicamente el bloque de *encoder* del *Transformer*, el cual se compone de una pila de N capas idénticas. Cada una de estas capas integra un

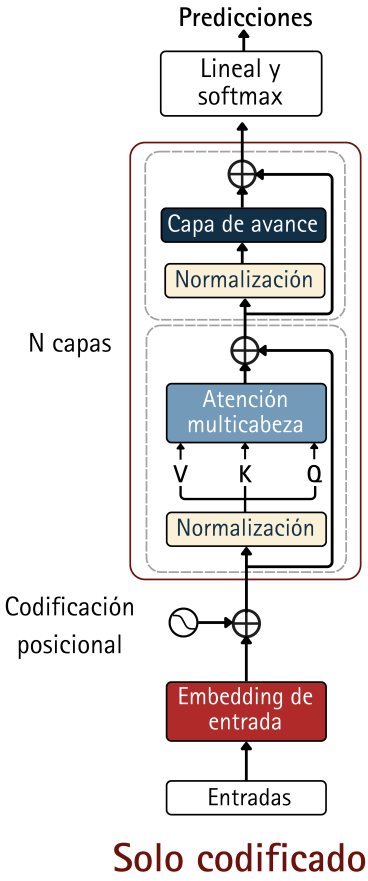
mecanismo de atención multi-cabeza, seguido de una red neuronal con retroalimentación hacia adelante (del inglés, *feed-forward*), ambos acompañados por operaciones de normalización y conexiones residuales (Noronha et al., 2025). Este diseño permite modelar dependencias contextuales de largo alcance, ya que cada *token* puede atender dinámicamente a todos los demás dentro de la secuencia.

La representación de entrada se construye a partir de incrustaciones (del inglés, *embeddings*) *tokenizados*, a los cuales se les incorpora información de posición mediante codificación posicional (del inglés, *positional encoding*), permitiendo al modelo conservar el orden secuencial de las palabras. Durante el preentrenamiento, BETO utiliza el objetivo de modelado de lenguaje enmascarado (del inglés, *masked language modeling*), en el cual ciertos *tokens* son ocultados y el modelo aprende a predecirlos a partir de su contexto (de la Torre, 2025).

En conjunto, esta arquitectura posibilita la construcción de representaciones contextualizadas altamente informativas, que resultan especialmente adecuadas para tareas donde es necesario identificar relaciones semánticas complejas dentro del texto (Li et al., 2019b). La Figura 2 presenta una vista esquemática del bloque codificador utilizado por BETO.

Figura 2

Diagrama de la arquitectura BERT y/o BETO



Nota. Adaptado de Vaswani et al., 2017. La figura presenta la arquitectura de *Transformer* BERT y/o BETO.

En el contexto de ABSA, los modelos basados en *Transformers* se han consolidado como el enfoque predominante para abordar tareas como ATE, ASC, ACD (véase Tabla 1) y formulaciones conjuntas. Su capacidad para modelar dependencias de largo alcance y construir representaciones contextualizadas permite identificar con mayor precisión la relación entre aspectos y expresiones de opinión, incluso en presencia de ambigüedad o información implícita.

En el caso del español, diversos estudios han mostrado que los modelos monolingües pre-entrenados ofrecen ventajas frente a alternativas multilingües. En particular, BETO ha alcanzado

resultados competitivos en múltiples tareas de NLP, lo que lo posiciona como una alternativa adecuada para escenarios que involucran variaciones lingüísticas, expresiones coloquiales y fenómenos característicos del español, como los presentes en reseñas de aplicaciones móviles.

En este contexto, el aprendizaje por transferencia mediante *fine-tuning* se establece como una estrategia efectiva para adaptar modelos preentrenados a tareas específicas de ABSA (Mosin et al., 2023), permitiendo obtener un desempeño robusto incluso cuando la disponibilidad de datos anotados es limitada.

2.2.2 Formulaciones conjuntas en ABSA

En los últimos años, las formulaciones conjuntas en ABSA han cobrado mayor relevancia con la adopción de modelos de DL, los cuales permiten integrar en un único marco las distintas subtareas de este campo (véase Tabla 1). Estos enfoques buscan modelar de manera unificada las dependencias entre dichas subtareas, superando las limitaciones de los métodos por etapas (Baishya y Baruah, 2022; Xu et al., 2018).

Con el avance de las arquitecturas basadas en DL, surgieron propuestas que incorporan mecanismos de atención e interacción entre subtareas, permitiendo que la información fluya entre la detección de aspectos y el análisis de sentimiento. Este tipo de estrategias ha demostrado ser efectivo para capturar relaciones contextuales más complejas y mejorar la coherencia de las predicciones (Hua et al., 2024; Li et al., 2019a).

Más recientemente, las formulaciones conjuntas han evolucionado hacia esquemas más estructurados orientados a la identificación directa de relaciones entre aspectos, términos de opinión y polaridad. Estos enfoques modelan explícitamente dichas interacciones mediante representaciones enriquecidas, como estructuras basadas en grafos o el uso de información posicional, lo que facilita capturar dependencias no locales dentro del texto (Peng et al., 2020; Xu et al., 2020).

En conjunto, estas aproximaciones han demostrado que el modelado unificado basado en DL no solo mejora el desempeño frente a enfoques secuenciales, sino que también reduce la

propagación de errores entre subtareas y permite representar de manera más precisa la estructura semántica de las opiniones (He et al., 2019).

2.2.3 Modelos basados en spans

Una línea de investigación relevante dentro del ABSA corresponde a los enfoques basados en *spans*. A diferencia de los modelos que operan únicamente a nivel de *token* o mediante clasificación global, los enfoques a nivel de *span* consideran fragmentos completos como unidades semánticas, permitiendo modelar explícitamente aspectos y expresiones de opinión que abarcan múltiples palabras (Zhou et al., 2019).

Este tipo de formulación se relaciona con tareas compuestas como ASTE (ver Tabla 1), en la cual el objetivo es identificar conjuntamente los componentes estructurales de una opinión. En estos escenarios, el modelado a nivel de *span* facilita la representación directa de los segmentos textuales asociados a cada elemento predicho.

Los enfoques a nivel de *span* han demostrado ventajas en contextos donde los aspectos y las opiniones no se limitan a una sola palabra o donde las dependencias entre ellos son implícitas o de largo alcance. Al considerar segmentos completos, estos modelos capturan interacciones semánticas de mayor nivel y permiten una correspondencia más clara entre las predicciones y el texto original (Xu et al., 2021).

2.2.4 Generación de datos sintéticos en ABSA

En los últimos años, la generación de datos sintéticos ha emergido como una estrategia de *data augmentation* prometedora para abordar la escasez de datos anotados en ABSA (Zhong et al., 2024). Este enfoque resulta particularmente relevante en contextos de bajos recursos, donde la disponibilidad de datos etiquetados es limitada o inexistente (Goyal y Mahmoud, 2024).

Con la aparición de modelos de lenguaje a gran escala (LLMs del inglés, *large language models*), se ha facilitado la generación automática de muestras sintéticas que simulan reseñas reales

junto con sus respectivas anotaciones. Estos modelos, basados en arquitecturas *Transformers*, permiten generar texto coherente y contextualmente rico, lo que los convierte en herramientas útiles para la expansión de conjuntos de entrenamiento (Hellwig et al., 2024; Liu et al., 2025).

Diversos trabajos han demostrado que los datos sintéticos generados mediante LLMs pueden mejorar el desempeño de modelos basados en DL en tareas de ABSA, especialmente cuando se combinan con datos reales en esquemas de entrenamiento híbridos. Además, estas estrategias permiten controlar características específicas de los datos generados, como la distribución de polaridades o la presencia de ciertos aspectos, lo que contribuye a reducir sesgos en los conjuntos de entrenamiento (Nadăş et al., 2025).

No obstante, el uso de datos sintéticos también plantea desafíos importantes, como la posible introducción de ruido, inconsistencias semánticas o patrones artificiales que no reflejan completamente el lenguaje natural. Por ello, investigaciones recientes han explorado técnicas de filtrado, validación automática y aprendizaje contrastivo para mejorar la calidad de los datos generados (Hellwig et al., 2024; Liu et al., 2025; Nadăş et al., 2025).

En conjunto, la generación de datos sintéticos mediante LLMs representa una línea de investigación emergente que ofrece nuevas oportunidades para fortalecer el desempeño de modelos de ABSA en escenarios de recursos limitados (Hellwig et al., 2024), particularmente en idiomas distintos al inglés y dominios específicos (Ziv et al., 2025).

2.2.5 Limitaciones y oportunidades en Español y dominios de recursos limitados

A pesar de los avances significativos, la mayoría de los modelos y evaluaciones en ABSA se han desarrollado y validado principalmente en inglés(Šmíd y Král, 2025) y en dominios específicos como reseñas de restaurantes o productos electrónicos. Esto genera una brecha importante para idiomas distintos y para dominios como las reseñas de aplicaciones móviles en español de Colombia.

En particular, el español colombiano presenta desafíos adicionales asociados a variaciones dialectales, uso de expresiones coloquiales, abreviaturas, errores ortográficos y construcciones

implícitas (Chris et al., 2024), lo que dificulta la transferencia directa de modelos entrenados en otros contextos. Asimismo, la escasez de conjuntos de datos debidamente etiquetados limita el entrenamiento supervisado de modelos complejos(Šmíd y Král, 2025).

En este contexto, los enfoques basados en aprendizaje por transferencia, modelos preentrenados monolingües, formulaciones basadas en *spans* y estrategias de generación de datos sintéticos representan oportunidades relevantes para desarrollar sistemas más robustos, interpretables y adaptados al dominio, alineándose con los objetivos del presente trabajo.

3. Metodología

A continuación, se presenta la metodología empleada para el desarrollo del presente trabajo de investigación. Esta se encuentra organizada en fases secuenciales, cada una orientada al cumplimiento de un objetivo específico.

3.1 Recolección y construcción del corpus de evaluación

Esta fase contempla la recolección de reseñas de aplicaciones móviles escritas por usuarios en el contexto colombiano, así como el procesamiento y organización de la información para garantizar su calidad y representatividad. Adicionalmente, se incluye la generación de datos sintéticos mediante LLMs, con el fin de enriquecer el conjunto de datos disponible y asegurar la cobertura de distintos escenarios de opinión.

3.1.1 Identificación y selección de fuentes de reseñas de aplicaciones móviles

Se realizó una selección de tipos de aplicaciones relevantes para el análisis de la UX. Los tipos elegidos fueron finanzas, compras y redes sociales, debido a su alta penetración en el uso cotidiano de los usuarios colombianos.

A partir de estos tipos, se realizó una preselección inicial de algunas aplicaciones por cada uno, considerando su popularidad y relevancia en el mercado colombiano. Posteriormente,

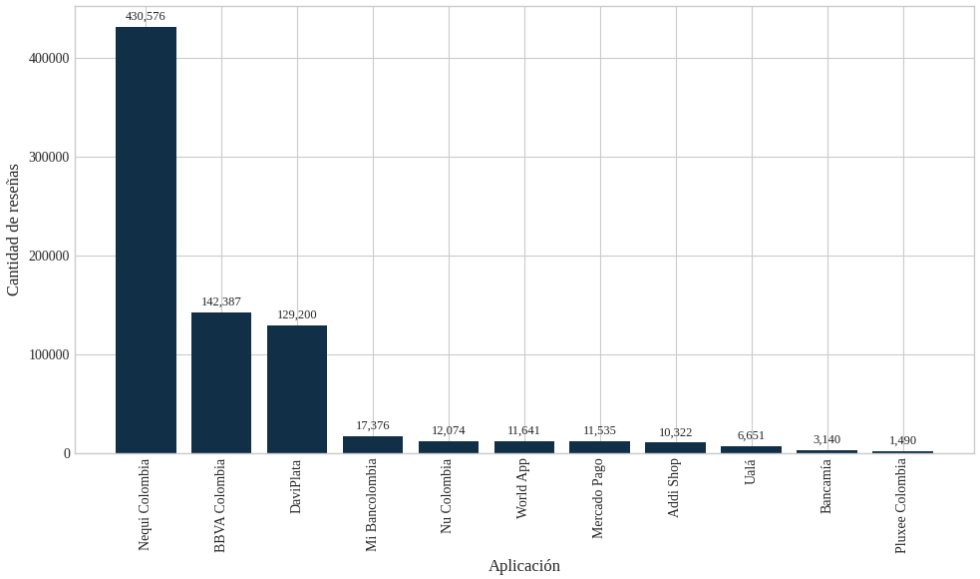
mediante el uso de librerías de *Python*, específicamente *Google Play Scraper* y *Pandas*, se extrajo información relacionada con la disponibilidad de reseñas en idioma español para cada aplicación.

Con base en el número de reseñas disponibles, se seleccionaron las tres aplicaciones con mayor volumen de datos por cada tipo de aplicación. En particular, se eligieron: *Nequi*, *BBVA* y *Daviplata* para el tipo de aplicaciones de finanzas; *Rappi*, *Mercado Libre* y *Temu* para el tipo de aplicaciones de compras; y *WhatsApp*, *YouTube* e *Instagram* para el tipo de aplicaciones de redes sociales.

Finalmente, los resultados del proceso de selección se presentan mediante gráficos comparativos que muestran la cantidad de reseñas en español disponibles por aplicación y por tipo (véanse Figuras 3, 4 y 5).

Figura 3

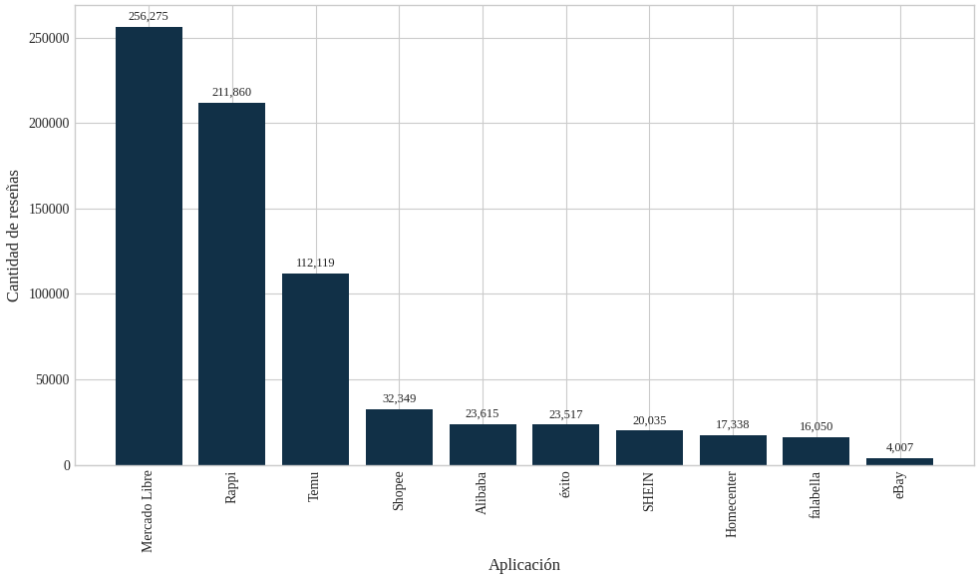
Cantidad de reseñas por aplicación - Tipo: Finanzas



Nota. La figura muestra la cantidad de reseñas disponibles por aplicación, correspondiente al tipo finanzas.

Figura 4

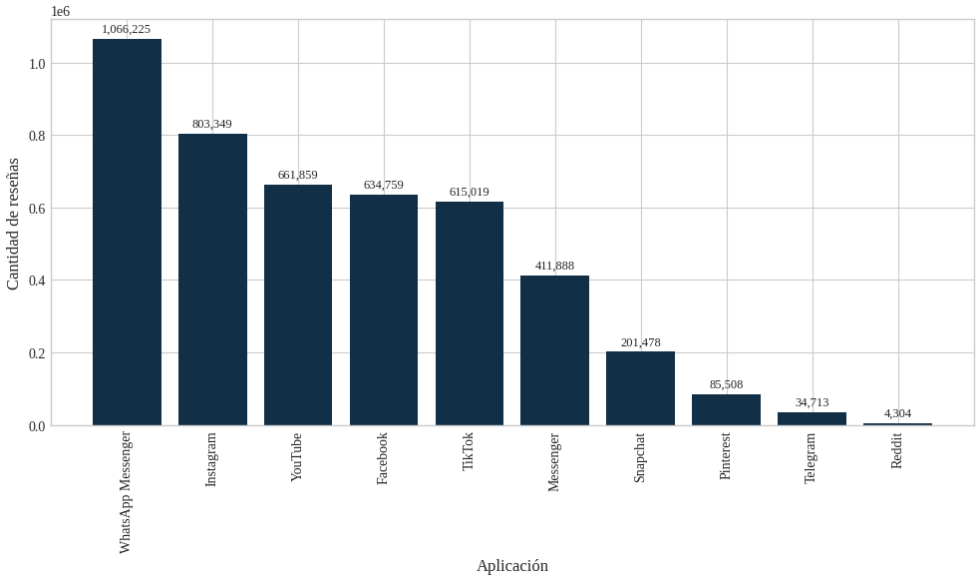
Cantidad de reseñas por aplicación - Tipo: Compras



Nota. La figura muestra la cantidad de reseñas disponibles por aplicación, correspondiente al tipo compras.

Figura 5

Cantidad de reseñas por aplicación - Tipo: Redes sociales



Nota. La figura muestra la cantidad de reseñas disponibles por aplicación, correspondiente al tipo redes sociales.

3.1.2 Recolección y almacenamiento de datos en un repositorio estructurado

Una vez definidas las aplicaciones objeto de estudio, se procedió a la extracción sistemática de las reseñas mediante el uso de la librería *Google Play Scraper*. El proceso de recolección se realizó de manera independiente para cada aplicación seleccionada.

Posteriormente, los conjuntos de datos obtenidos fueron procesados y concatenados, con el fin de conformar un único conjunto de datos general en formato JSON, compuesto por **3.770.108** filas, correspondientes al total de reseñas recopiladas. En la Figura 6, se muestran los campos que conforman cada reseña.

Figura 6
Ejemplo de reseña recopilada en formato JSON mediante la librería Google Play Scraper


```
{
  {"review_id":"0d53b8c7-135a-4df0-aa39-8a55bebcdb41",
  "review_text":"Excelente herramienta de trabajo",
  "app_id":"com.instagram.android",
  "category":"social"
  ,"source":"google_play",
  "extracted_at":"2025-12-15T04:08:54.338"}
}
```


Nota. La figura presenta un ejemplo de una reseña extraída mediante la librería *Google Play Scraper*, que se encuentra en el conjunto de datos de partida.

3.1.3 Preprocesamiento lingüístico de los textos

Con el objetivo de garantizar la calidad, consistencia y homogeneidad del corpus de análisis, se llevó a cabo un proceso integral de preprocesamiento, limpieza y filtrado lingüístico sobre las reseñas recolectadas. En este sentido, se seleccionó exclusivamente el campo *review_text* (véase Figura 6), correspondiente al contenido de la reseña, dado que constituye la fuente principal de información semántica necesaria para el desarrollo del estudio. Los demás atributos disponibles,

tales como identificadores, metadatos de la aplicación, categoría, fuente y fecha de extracción, fueron descartados al no aportar valor significativo para los objetivos planteados.

Inicialmente, se eliminaron las reseñas vacías. Posteriormente, todo el texto fue convertido a minúsculas. A continuación, se eliminaron elementos ajenos al contenido opinativo, tales como enlaces, menciones y referencias externas. Asimismo, se suprimieron los números y los caracteres no alfabéticos, conservando únicamente letras del alfabeto español.

Como parte del proceso de normalización, se eliminaron las tildes. Adicionalmente, se depuraron los espacios en blanco redundantes y se eliminaron palabras repetidas de forma consecutiva. Todas estas operaciones fueron implementadas mediante algoritmos en *Python*, utilizando librerías especializadas en NLP, específicamente *SpaCy*, *Stanza* y *NLTK*.

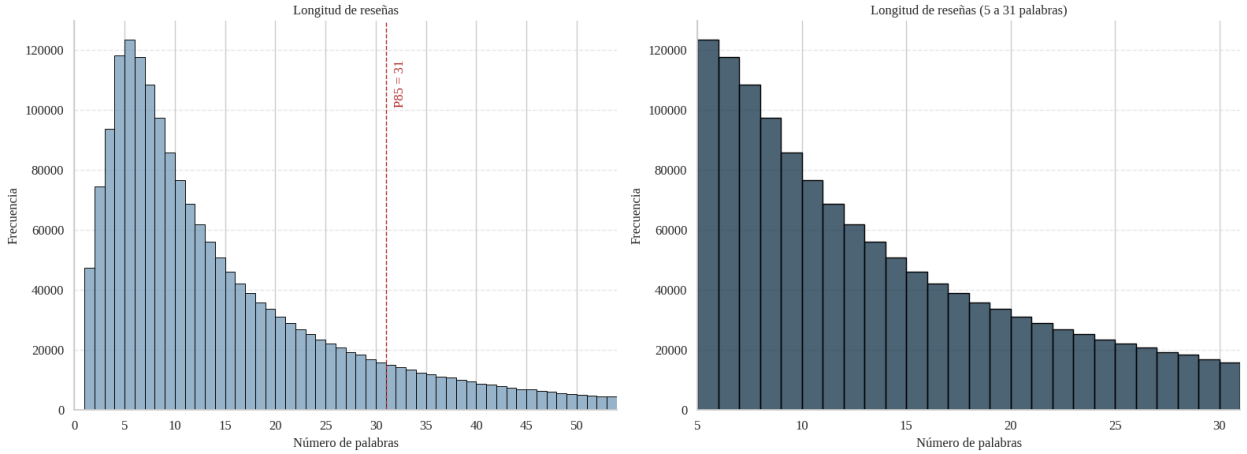
Finalmente, se aplicó un criterio de filtrado por longitud de las reseñas, calculando el número de palabras por texto y descartando aquellas con menos de 5 palabras, por considerarse insuficientes para expresar una opinión clara, así como aquellas con más de 31 palabras, dado que suelen contener información dispersa y poco adecuada para el ABSA. Para este límite superior se tomó como referencia el percentil 85 del conjunto de datos.

Como resultado del preprocesamiento y filtrado, el conjunto de datos final quedó depurado y listo para el análisis, conformando un dataset de **1.307.211** reseñas.

En la Figura 7 se presentan las distribuciones de reseñas por número de palabras antes y después de los criterios de filtrado definidos.

Figura 7

Distribución de reseñas por número de palabras antes y después del preprocesamiento lingüístico



Nota. La figura muestra las distribuciones de las reseñas por cantidad de palabras antes y después del preprocesamiento lingüístico.

3.1.4 Definición de reglas para el etiquetado de reseñas

El proceso de anotación se estructuró mediante un esquema de etiquetado basado en cuádruplas. Cada instancia anotada está compuesta por: (i) la *categoría*, que clasifica el tipo de evaluación realizada; (ii) el *aspecto*, que identifica la dimensión específica de la aplicación a la que hace referencia el comentario; (iii) la *polaridad*, que indica la orientación del sentimiento expresado; y (iv) el *span de evidencia*, que corresponde al segmento textual exacto dentro de la reseña que sustenta explícitamente la anotación.

A continuación, se presentan las tablas que describen formalmente cada una de estas dimensiones consideradas en el esquema de anotación.

3.1.4.1 Selección de categorias. Se definió un conjunto de categorías que delimitan las principales dimensiones de la UX consideradas en el estudio. El conjunto está compuesto por seis: funcionalidad, usabilidad, costos, seguridad, rendimiento y general (véase Tabla 2). Estas categorías fueron seleccionadas con base en modelos clásicos de calidad en ingeniería de software,

como FURPS+ y el estándar ISO 25002, los cuales identifican atributos como funcionalidad, usabilidad, rendimiento y seguridad como dimensiones fundamentales en la evaluación de sistemas software (ISO, 2024; Puspita et al., 2024).

Tabla 2
Categorías definidas para el esquema de anotación

Categoría	Descripción
Costos	Abarca comentarios asociados a precios, cobros, suscripciones, comisiones y cualquier elemento económico percibido por el usuario.
Funcionalidad	Agrupa opiniones relacionadas con las características y servicios ofrecidos por la aplicación, incluyendo la disponibilidad y el correcto funcionamiento de sus funciones principales.
General	Corresponde a una categoría miscelánea destinada a reseñas que expresan una valoración global de la aplicación, sin referencia explícita a un aspecto específico.
Rendimiento	Contempla opiniones relacionadas con la velocidad, estabilidad, consumo de recursos y la presencia de errores o fallos técnicos.
Seguridad	Incluye referencias a la protección de datos, privacidad, autenticación y confianza en el manejo de la información personal.
Usabilidad	Se refiere a la facilidad de uso, la interfaz gráfica y la experiencia de interacción del usuario con la aplicación.

Nota. Descripción de las categorías consideradas para el análisis de sentimientos a nivel de aspecto.

3.1.4.2 Selección de aspectos. Adicionalmente, se establecieron criterios operativos para cada aspecto, definidos mediante un enfoque heurístico e iterativo, en el que dichos criterios fueron ajustados progresivamente a medida que se revisaba el corpus, con el fin de estandarizar su identificación (véase Tabla 3).

Tabla 3
Aspectos definidos para el esquema de anotación

Aspecto	Descripción
Anuncios	Abarca comentarios relacionados con la presencia, frecuencia y gestión de anuncios dentro de la aplicación, incluyendo publicidad mostrada al usuario.
Aplicación (General)	Agrupa comentarios generales sobre la aplicación que no encajan claramente en otros aspectos, incluyendo funcionalidades globales, comparaciones con otras aplicaciones y valoraciones amplias del servicio.
Atención y soporte	Agrupa comentarios relacionados con el servicio de atención al cliente y el soporte técnico ofrecido por la aplicación, incluyendo tiempos de respuesta, calidad del servicio, canales de contacto y resolución de problemas.
Autenticación e ingreso	Incluye todos los comentarios relacionados con los procesos de autenticación y acceso a la aplicación, tales como inicio de sesión, registro, verificación mediante códigos, biometría, contraseñas y mecanismos similares, así como problemas asociados a estos procesos.
Consumo de recursos	Se refiere al uso de recursos del dispositivo por parte de la aplicación, tales como consumo de batería, memoria, datos móviles y desempeño en dispositivos con capacidades limitadas.
Contenido	Incluye opiniones sobre el contenido presentado por la aplicación, tales como productos, publicaciones, mensajes, videos, imágenes, chats u otros elementos informativos o interactivos ofrecidos al usuario.
Envíos	Incluye opiniones relacionadas con el proceso de envío físico de productos, tales como tiempos de entrega, estado de los pedidos, desempeño de repartidores y cumplimiento de entregas.

Continúa en la siguiente página

Tabla 3 – *continuación*

Aspecto	**Descripción**
Estabilidad	Se refiere al estado técnico general de la aplicación, incluyendo fallos, bloqueos, cierres inesperados, errores, problemas de conexión, lentitud, tiempos de carga y cualquier otro comportamiento que afecte la correcta ejecución de la app.
Flujo de uso	Abarca comentarios relacionados con la secuencia de pasos, organización de procesos y estructura de navegación de la aplicación, así como la facilidad o dificultad para completar tareas desde una perspectiva procedural.
Gestión de cuenta	Abarca opiniones sobre la administración de la cuenta del usuario dentro de la aplicación, incluyendo perfil, datos personales, contactos, configuración de cuenta, vinculación con otros dispositivos y manejo general de la información asociada a la cuenta.
Interfaz	Corresponde a opiniones sobre los aspectos visuales y estéticos de la aplicación, tales como diseño gráfico, colores, botones, menús y presentación general de la interfaz.
Notificaciones	Incluye referencias a los sistemas de avisos y alertas de la aplicación, tales como notificaciones push, correos electrónicos u otros mecanismos de comunicación automática con el usuario.
Promociones y reembolsos	Comprende referencias a promociones, cupones, reembolsos, regalos, cashback y promesas comerciales, así como el cumplimiento o incumplimiento de estos beneficios.
Transferencias	Comprende todos los comentarios relacionados con el envío, recepción y movimiento de dinero u otros activos, incluyendo transferencias, pagos, convenios con entidades financieras y operaciones monetarias realizadas a través de la aplicación.

Nota. Descripción de los aspectos considerados en el esquema de anotación para el análisis de sentimientos basado en aspectos.

3.1.4.3 Clasificación de las polaridades. Finalmente, se definieron las polaridades consideradas (véase Tabla 4).

Tabla 4
Polaridades definidas para el esquema de anotación

Polaridad	Descripción
Positiva	Expresa satisfacción, aprobación o valoración favorable respecto a un aspecto específico de la aplicación. Incluye comentarios que resaltan un buen funcionamiento, facilidad de uso, confianza, rapidez o beneficios percibidos.
Negativa	Indica insatisfacción, quejas o valoración desfavorable asociada a un aspecto de la aplicación. Incluye referencias a errores, fallos, dificultades, cobros indebidos, problemas de seguridad o cualquier experiencia percibida como perjudicial por el usuario.
Neutra	Corresponde a opiniones descriptivas, informativas o ambiguas que no expresan claramente una valoración positiva ni negativa. Incluye consultas y recomendaciones.

Nota. Descripción de las polaridades consideradas para el análisis de sentimientos a nivel de aspecto.

Es importante señalar que no todas las categorías definidas en el esquema de anotación se relacionan con todos los aspectos. En lugar de permitir combinaciones arbitrarias, se estableció un conjunto restringido de relaciones categoría-aspecto, con el fin de reflejar de manera más precisa la naturaleza semántica de las reseñas y evitar asignaciones conceptualmente inconsistentes.

De esta forma, cada categoría solo puede asociarse con aquellos aspectos que resultan coherentes con su definición funcional y con los tipos de opiniones que típicamente expresan los usuarios (véase Tabla 5).

Tabla 5
Combinaciones válidas entre categorías y aspectos

Aspecto	Categorías permitidos
Anuncios	Funcionalidad, Usabilidad
Aplicación (General)	Funcionalidad, Seguridad, General
Atención y soporte	Funcionalidad, Usabilidad, Seguridad
Autenticación e ingreso	Funcionalidad, Usabilidad
Consumo de recursos	Rendimiento
Contenido	Funcionalidad, Usabilidad, Costos, Seguridad
Envíos	Funcionalidad, Costos, Seguridad
Estabilidad	Rendimiento
Flujo de uso	Usabilidad
Gestión de cuenta	Funcionalidad, Usabilidad, Seguridad
Interfaz	Usabilidad
Notificaciones	Funcionalidad, Usabilidad, Seguridad
Promociones y reembolsos	Funcionalidad, Usabilidad
Transferencias	Funcionalidad, Usabilidad, Costos, Seguridad

Nota. La tabla presenta las combinaciones válidas entre categorías y aspectos definidas en el protocolo de anotación.

3.1.4.4 Fragmentos centrados en evidencia textual.

En este trabajo se introduce el concepto de fragmento de evidencia o *evidence span* como unidad fundamental de anotación. Un *evidence span* se define como el fragmento de texto contiguo más corto dentro de una reseña que justifica la asignación de una tripleta semántica compuesta por aspecto, categoría y polaridad. A diferencia de enfoques tradicionales en los que la unidad de anotación corresponde a la reseña completa, en este esquema la unidad atómica de etiquetado es el propio *evidence span*. En consecuencia, una misma reseña puede contener múltiples anotaciones, donde cada una corresponde a un *evidence span* distinto. Finalmente, cada *evidence span* se representa mediante las posiciones de inicio y fin dentro de la secuencia de palabras de la reseña, lo que permite delimitar de manera precisa el fragmento textual que sustenta cada anotación.

En el contexto de este esquema de anotación, se introduce el concepto de núcleo semántico, entendido como el conjunto mínimo de palabras dentro del *evidence span* cuya presencia es indispensable para preservar el significado completo de la opinión expresada. El núcleo semántico representa la porción esencial del enunciado sobre la cual se construye la interpretación de la valoración, preferencia o juicio emitido por el autor de la reseña.

A diferencia de una definición centrada exclusivamente en los términos valorativos, el núcleo semántico puede incluir tanto expresiones de valoración como otros elementos lingüísticos necesarios para que dicha valoración conserve sentido. Su identificación se basa en un criterio de indispensabilidad semántica: si el fragmento seleccionado se elimina del enunciado, el texto restante pierde la información fundamental que permite comprender la opinión expresada o interpretar correctamente la relación entre sus componentes.

Por ejemplo, en el enunciado “me gustan los productos y los precios”, la expresión “me gustan” constituye parte del núcleo semántico, ya que su eliminación deja únicamente la secuencia “los productos y los precios”, la cual ya no expresa ninguna valoración. De manera similar, en “me gustan los productos pero no los precios”, la eliminación de “me gustan” impide interpretar adecuadamente la preferencia manifestada por el autor. Asimismo, en la oración “el domicilio es bueno y barato”, el segmento “el domicilio” también forma parte del núcleo semántico, puesto que sin él la expresión “es bueno y barato” pierde la referencia al aspecto evaluado y, por tanto, deja de transmitir una opinión completa.

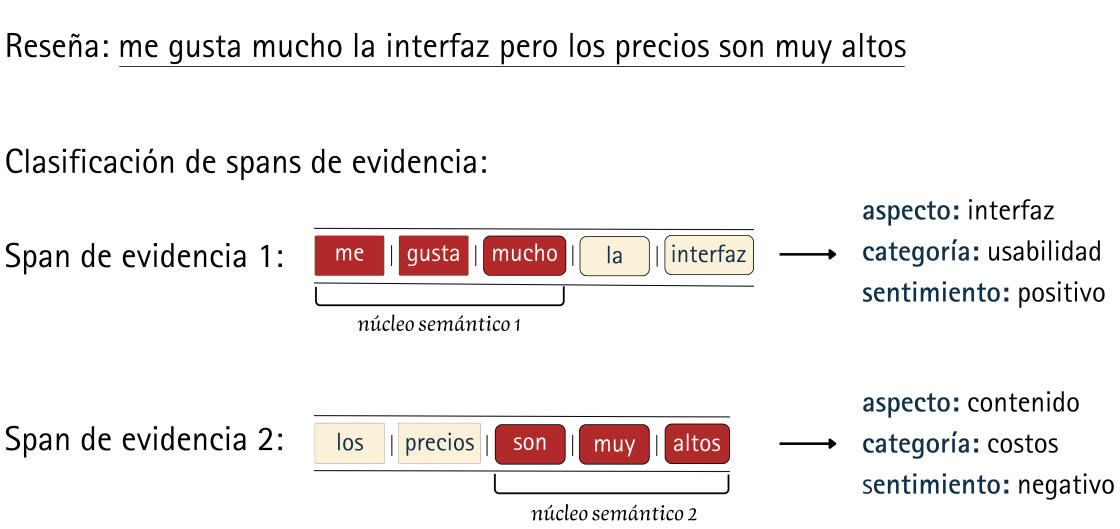
En muchos casos, el núcleo semántico aparece junto al término que referencia el aspecto evaluado dentro de un mismo fragmento textual, lo que permite delimitar un *evidence span* contiguo que contiene toda la información necesaria para justificar la tripleta semántica. No obstante, también pueden presentarse situaciones en las que un mismo núcleo semántico sustenta la evaluación de múltiples aspectos dentro de la misma oración. En tales casos, algunos fragmentos que mencionan aspectos pueden carecer por sí mismos de contenido evaluativo explícito. Debido a la regla de anotación adoptada en este trabajo (según la cual un *evidence span* debe corresponder al fragmento

contiguo más corto que justifique la asignación de la tripleta), se establece que estos aspectos compartan el mismo *evidence span* asociado al núcleo semántico que expresa la valoración.

A partir de esta distinción entre núcleo semántico y *evidence span*, en el proceso de anotación pueden presentarse diferentes configuraciones dependiendo de cómo se distribuyan las opiniones dentro de la reseña. A continuación se describen los dos casos principales considerados en la investigación.

3.1.4.4.1 Núcleo semántico independiente. Ocurre cuando distintos fragmentos de una misma reseña sustentan opiniones sobre aspectos diferentes. En este caso, cada opinión se asocia a un *evidence span* distinto, claramente delimitado dentro del texto (véase Figura 8).

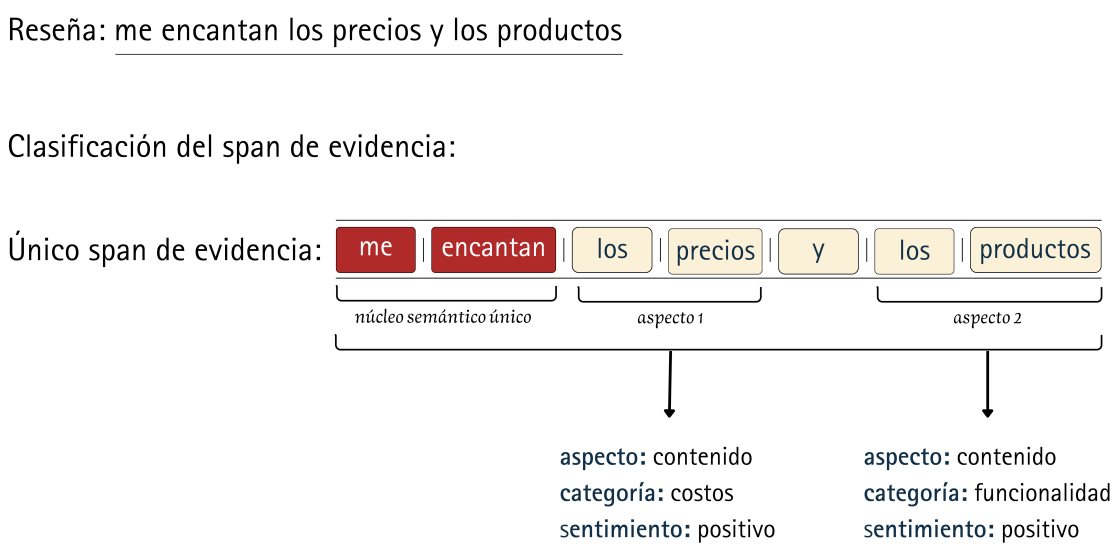
Figura 8
Ejemplo de núcleos semánticos independientes



Nota. La figura presenta la clasificación de *evidence spans* en el caso de una reseña con núcleos semánticos independientes.

3.1.4.4.2 Núcleo semántico compartido. Se presenta cuando un mismo fragmento textual justifica simultáneamente la valoración de múltiples aspectos. En esta situación, el mismo *evidence span* se asocia a más de una tripleta semántica (véase Figura 9).

Figura 9
Ejemplo de núcleo semántico compartido



Nota. La figura presenta la clasificación de *evidence spans* en el caso de una reseña con núcleo semántico compartido.

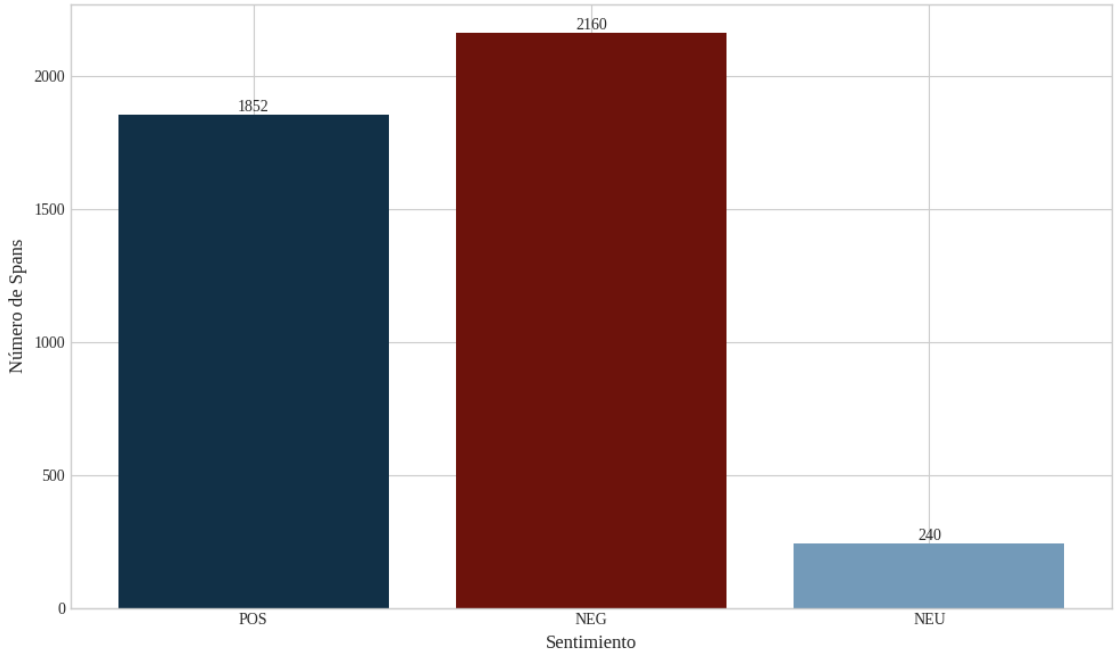
3.1.5 Creación del conjunto de datos de evaluación a partir del corpus procesado

3.1.5.1 Selección de reseñas y etiquetado manual. El proceso de anotación se llevó a cabo de manera manual, revisando cada reseña individualmente y registrando las etiquetas correspondientes en un archivo en formato JSON con los campos de anotación previamente definidos (véase Sección 3.1.4).

Una vez consolidado el conjunto de prueba y completado el proceso de anotación manual, se procedió a analizar la distribución de las etiquetas con el fin de evaluar el equilibrio del corpus. Esto permitió identificar un claro desbalance en la distribución de las etiquetas, especialmente en aquellas de carácter neutro, las cuales estaban notablemente subrepresentadas (véase Figura 10).

Figura 10

Frecuencia de sentimientos según los spans de evidencia



Nota. La figura presenta la frecuencia de aparición de cada sentimiento según los *spans* de evidencia.

3.1.5.2 Generación de data sintética mediante el uso de LLMs. Para abordar este desafío, se aplicó la técnica de *data augmentation* mediante generación de datos sintéticos. En este caso, se utilizó el LLM *ChatGPT-5.0* para generar reseñas sintéticas que pretenden imitar las reseñas originales en estilo, longitud y estructura. Cada una de estas reseñas ya incluía las etiquetas asignadas, siguiendo de manera consistente la guía de anotación definida previamente (véase Sección 3.1.4).

La generación de datos sintéticos fue un proceso supervisado, realizado en lotes de 10 reseñas para garantizar su calidad y coherencia, utilizando la plantilla de *prompt* construida específicamente para este fin (véase Apéndice C). Este procedimiento se diseñó siguiendo metodologías y estudios previos sobre generación de datos sintéticos y balanceo de conjuntos de datos en tareas de análisis de opinión y clasificación de textos (Hellwig et al., 2024), garantizando un conjunto de

datos robusto y confiable para el entrenamiento del modelo.

Con el fin de emular de manera realista las particularidades del lenguaje presente en las reseñas auténticas, algunas reseñas generadas fueron revisadas y modificadas manualmente, incorporando variaciones léxicas y sintácticas, así como expresiones coloquiales propias del español colombiano y ruido controlado. Este proceso se llevó a cabo de manera iterativa, de modo que no todas las reseñas generadas por el LLM fueron incorporadas al conjunto de datos. Por el contrario, cada instancia fue evaluada bajo criterios de aceptación definidos, lo que implicó el descarte de un número considerable de reseñas que no cumplían con los estándares. Para una descripción detallada de estos criterios, véase el Apéndice B.

3.2 Implementación del modelo de clasificación de reseñas con técnicas de NLP

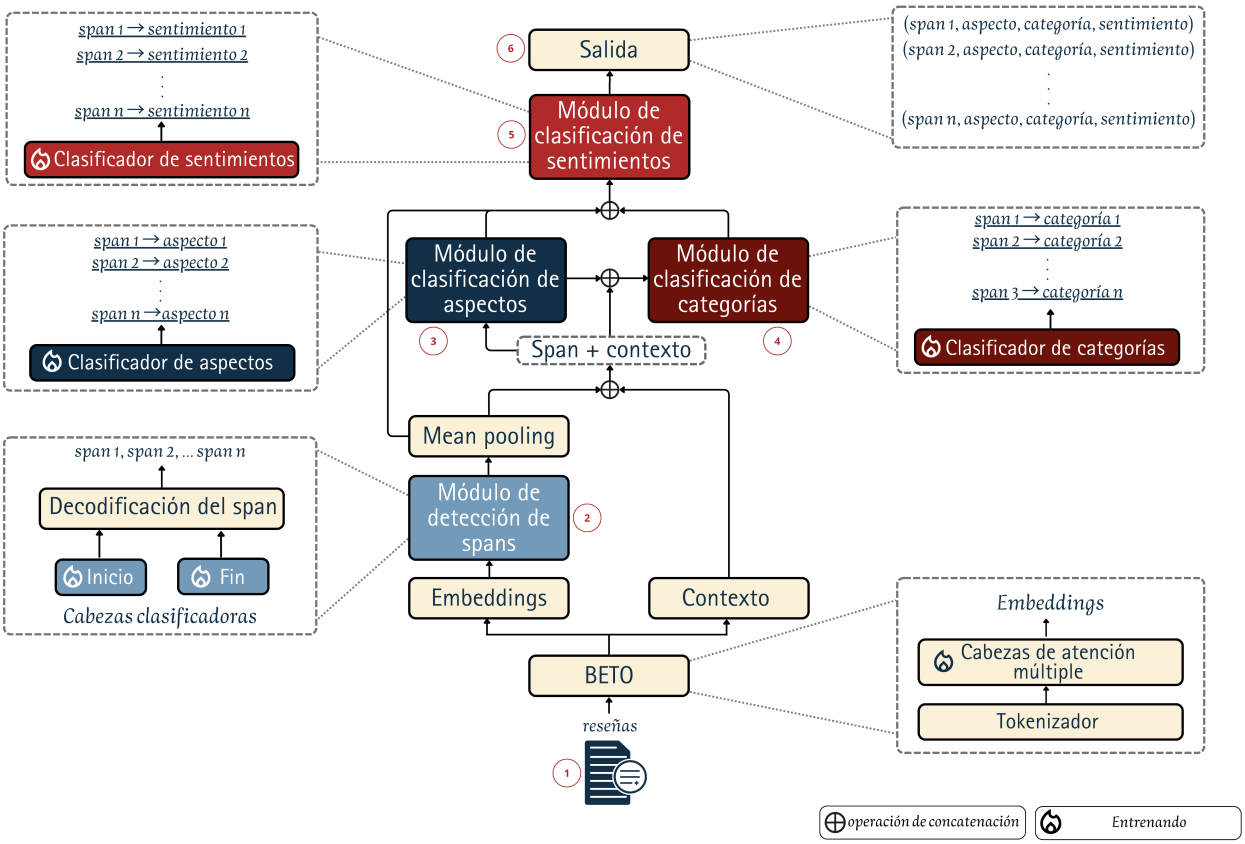
Se desarrolló un enfoque de ABSA centrado en evidencia textual para el análisis automático de reseñas de aplicaciones móviles en español. La propuesta incorpora explícitamente la identificación de los *evidence spans*. De esta manera, el sistema identifica de forma conjunta tanto las evidencias textuales como las tripletas semánticas asociadas a cada una de ellas, compuestas por aspecto, categoría y sentimiento.

La Figura 11 presenta una visión global del modelo propuesto, ilustrando el flujo completo de procesamiento desde la entrada de la reseña ① hasta la generación de las predicciones finales ⑥. En ella se observa, en primer lugar, la codificación contextual del texto mediante el *Transformer* BETO, a partir de la cual se obtienen *embeddings* contextualizados de cada *token*, a través de su mecanismo de *self-attention*. Sobre estas representaciones, el modelo se estructura en dos componentes principales: por un lado, el módulo de detección de *evidence spans* ②, que identifica posibles fragmentos relevantes mediante clasificadores de inicio y fin, seguido de un proceso de selección y decodificación. Por otro lado, el módulo de clasificación jerárquica, que opera sobre cada *evidence span* detectado: inicialmente se construye una representación agregada mediante agrupamiento promedio (del inglés, *mean pooling*) concatenada con el *token* especial CLS, el cual

actúa como el contexto global de la reseña. A partir de esta representación, el modelo emplea un conjunto de cabezas de clasificación específicas para cada subtarea, implementadas como capas lineales seguidas de funciones de activación acordes a la naturaleza del problema (*sigmoid* para tareas de clasificación multietiqueta y *softmax* para clasificación multiclase). Sobre esta base, se realizan de forma secuencial y condicionada las predicciones de aspecto ③, categoría ④ y polaridad ⑤. Este enfoque basado en evidencia permite vincular explícitamente las predicciones semánticas con su justificación textual, lo que facilita tanto la interpretación de los resultados como una evaluación más rigurosa basada en la coincidencia entre *evidence spans* y etiquetas semánticas.

Figura 11

Modelo ABSA propuesto



Nota. La figura presenta la arquitectura general del modelo ABSA propuesto, destacando los módulos para la detección y clasificación de fragmentos de evidencia en reseñas.

3.2.1 Diseño de la arquitectura experimental del modelo

Para el modelo, las entradas corresponden a un conjunto de reseñas textuales anotadas que conforman el corpus de entrenamiento. Formalmente, el corpus se define como

$$\mathcal{D} = \{(r_i, L_i)\}_{i=1}^{N},$$

donde r_i representa una reseña textual, L_i el conjunto de anotaciones asociadas a dicha reseña y N corresponde al número total de reseñas del conjunto de datos. Cada anotación describe una opinión presente en el texto mediante un *evidence span* y un conjunto de etiquetas semánticas. En particular, para cada reseña se define su conjunto de anotaciones

$$L_i = \{\ell_{i,1}, \ell_{i,2}, \ldots, \ell_{i,k}\},$$

donde cada anotación $\ell_{i,k}$ se representa como una cuádrupla

$$\ell_{i,k} = (e_{i,k}, a_{i,k}, c_{i,k}, s_{i,k}),$$

siendo $e_{i,k}$ el fragmento de evidencia asociado a la anotación k-ésima de la reseña i, $a_{i,k} \in \mathcal{A}$ un aspecto perteneciente al conjunto de aspectos $\mathcal{A}$ definidos en la Tabla 3, $c_{i,k} \in C$ una categoría perteneciente al conjunto de categorías C descritas en la Tabla 2, y $s_{i,k} \in \mathcal{S}$ la polaridad asociada a la opinión expresada, que corresponde al conjunto de polaridades $\mathcal{S}$ presentado en la Tabla 4. De esta manera, el modelo recibe como entrada una reseña completa y aprende a identificar simultáneamente los fragmentos de texto relevantes y las etiquetas semánticas que describen la opinión contenida en ellos.

3.2.1.1 Tokenización y alineación de spans de evidencia. Para cada reseña r_i, se aplica el tokenizador del modelo BETO, llamado *WordPiece Tokenizer*, el cual segmenta el texto en una secuencia de *tokens* y proporciona, para cada *token*, su correspondencia con el texto original mediante pares de *offsets* de caracteres. Formalmente, para cada *token t* se obtiene un par

$$\text{offset}_{i,t} = \left(o_{i,t}^{(0)}, o_{i,t}^{(1)}\right),$$

donde $o_{i,t}^{(0)}$ y $o_{i,t}^{(1)}$ indican, respectivamente, los índices de inicio y fin del fragmento del texto original al que corresponde dicho *token*.

Por otra parte, los *evidence spans* del conjunto de datos están etiquetados originalmente mediante índices de palabras dentro de la reseña, es decir, como pares

$$(w_{i,k}^{\text{start}}, w_{i,k}^{\text{end}}).$$

Para poder utilizarlos durante el entrenamiento, dichos índices se transforman a *offsets* de caracteres

$$e_{i,k} = \left(ch_{i,k}^{(0)}, ch_{i,k}^{(1)}\right),$$

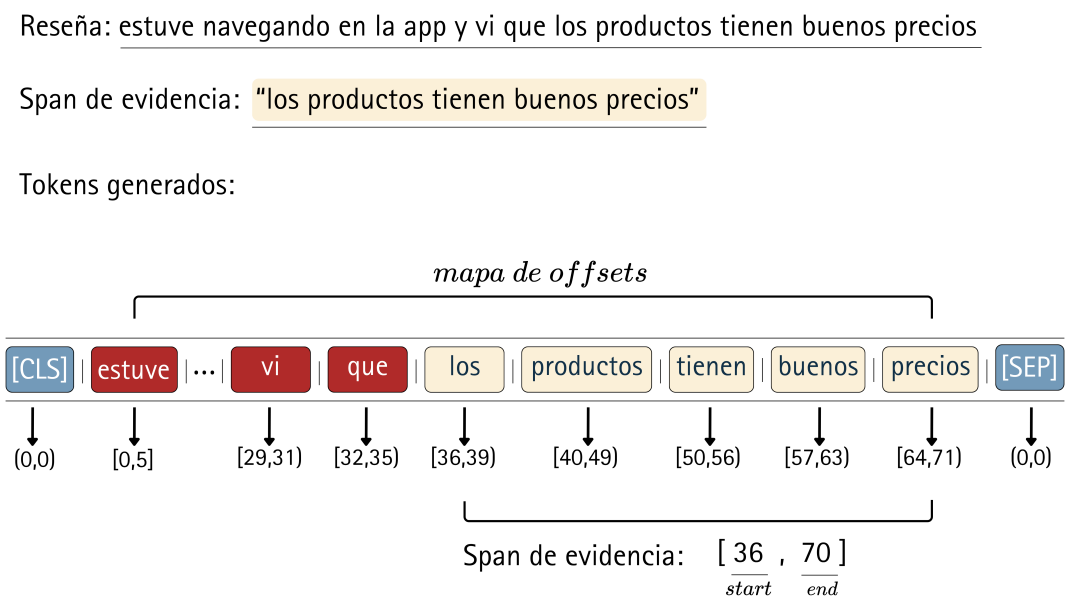
mediante el recorrido del texto original y la localización exacta del fragmento correspondiente. De esta manera, cada evidencia queda representada como un intervalo de caracteres dentro de la reseña.

Finalmente, utilizando el mapa de *offsets* proporcionado por el tokenizador, estos intervalos de caracteres se alinean con la secuencia de *tokens* generada por BETO. Este proceso permite construir las etiquetas a nivel de *token* que se utilizan durante el entrenamiento del modelo para aprender a localizar los fragmentos de evidencia dentro de cada reseña.

En la Figura 12 se ilustra el proceso de construcción del mapa de *offsets*. Para cada *token*, se asignan sus posiciones de inicio y fin mediante intervalos cerrados por la izquierda y abiertos

por la derecha. Esta convención se adopta debido a que cada *token* incluye un carácter adicional correspondiente al espacio que separa las palabras en la reseña original. En consecuencia, el *span* de evidencia resultante se representa mediante intervalos cerrados en ambos extremos, considerando como límite superior la posición del último carácter significativo del *token* final que lo compone, excluyendo el espacio asociado.

Figura 12
Mapeo de caracteres



Nota. La figura presenta una simplificación del proceso de mapeo de anotación de nivel palabra a nivel *token*, en el que las palabras y los *tokens* coinciden.

3.2.1.2 Construcción de las entradas del modelo. Durante el entrenamiento, las reseñas se procesan en lotes y se representan mediante tensores de entrada. En particular, la secuencia de *tokens* se codifica como

$$\mathbf{X} \in \mathbb{Z}^{B \times L},$$

donde B corresponde al tamaño del lote (número de reseñas procesadas simultáneamente, véase Tabla 6) y L a la longitud máxima de secuencia. Dado que las reseñas pueden tener longitudes variables, se fija un valor máximo L (véase Tabla 6) y se aplica un proceso de normalización

de longitud: aquellas reseñas con más de L *tokens* son truncadas, conservando únicamente los primeros L elementos, mientras que las más cortas se completan mediante *tokens* de relleno (del inglés, *padding*) hasta alcanzar dicha longitud. De esta forma, cada fila de $\mathbf{X}$ representa una reseña con tamaño uniforme.

Adicionalmente, se define una máscara de atención

$$\mathbf{M} \in \{0, 1\}^{B \times L},$$

que indica qué posiciones corresponden a *tokens* reales (1) y cuáles a posiciones de relleno (0). Esta máscara permite al modelo ignorar el *padding* durante el cálculo de las representaciones contextualizadas, asegurando que la atención se compute únicamente sobre los *tokens* válidos de cada reseña.

3.2.1.3 Codificación contextual de la reseña. El primer paso en el procesamiento de la reseña consiste en obtener representaciones contextualizadas de cada *token* mediante el *encoder* de BETO. En este tipo de modelos, los denominados estados ocultos (del inglés, *hidden states*) corresponden a vectores latentes que se construyen para cada *token* a lo largo de las capas del *encoder*. Este proceso se apoya fundamentalmente en el mecanismo de *self-attention*, el cual permite que cada vector incorpore información proveniente del resto de la secuencia. De esta manera, el modelo logra capturar dependencias semánticas bidireccionales al ponderar la relevancia de cada palabra respecto a las demás dentro de la misma oración. Como resultado, los estados ocultos permiten al modelo entender el sentido completo de la reseña.

Formalmente, sea $E(\cdot)$ el *encoder* contextual del modelo. Dada una secuencia de entrada $\mathbf{X}$, este produce una secuencia de estados ocultos:

$$H = E(\mathbf{X}) \in \mathbb{R}^{B \times L \times d},$$

donde d denota la dimensionalidad del espacio latente del modelo ($d = 768$ para BETO), y cada vector $h_{b,t} \in H$ representa la codificación contextual del *token* t en la reseña b.

Adicionalmente, el *encoder* produce una representación global de la secuencia a partir del *token* especial `[CLS]`, el cual se ubica al inicio de la entrada y está diseñado para agregar información de toda la reseña a través de los mecanismos de *self-attention*. Como resultado, el estado oculto asociado a este *token* actúa como un resumen contextual de la secuencia completa.

Formalmente, esta representación se denota como:

$$p_b \in \mathbb{R}^d,$$

donde p_b corresponde al vector asociado al *token* `[CLS]` en la reseña b. Esta representación global se utiliza en etapas posteriores del modelo para complementar y enriquecer los estados ocultos derivados de los *spans* de evidencia.

3.2.1.4 Detección de fragmentos de evidencia. A partir de los estados ocultos, el modelo estima de manera independiente la probabilidad de que cada *token* corresponda al inicio o al final de un fragmento de evidencia. Para ello, se emplean dos clasificadores lineales que producen *logits* definidos como

$$\ell_{b,t}^{\text{start}} = \mathbf{w}_s^\top h_{b,t} + b_s, \qquad \ell_{b,t}^{\text{end}} = \mathbf{w}_e^\top h_{b,t} + b_e,$$

donde $h_{b,t}$ denota el estado oculto del *token* t, mientras que $\mathbf{w}_s, \mathbf{w}_e \in \mathbb{R}^d$ y $b_s, b_e \in \mathbb{R}$ son los parámetros aprendidos.

Con el fin de ajustar la confianza de las probabilidades predichas, se introduce un factor de temperatura $T > 0$, el cual reescala los *logits* antes de aplicar la función *sigmoid*, siguiendo el enfoque de *temperature scaling* (Guo et al., 2017). En este trabajo se emplean valores $T > 1$ (véase Tabla 6), lo que produce distribuciones de probabilidad más suaves y menos concentradas.

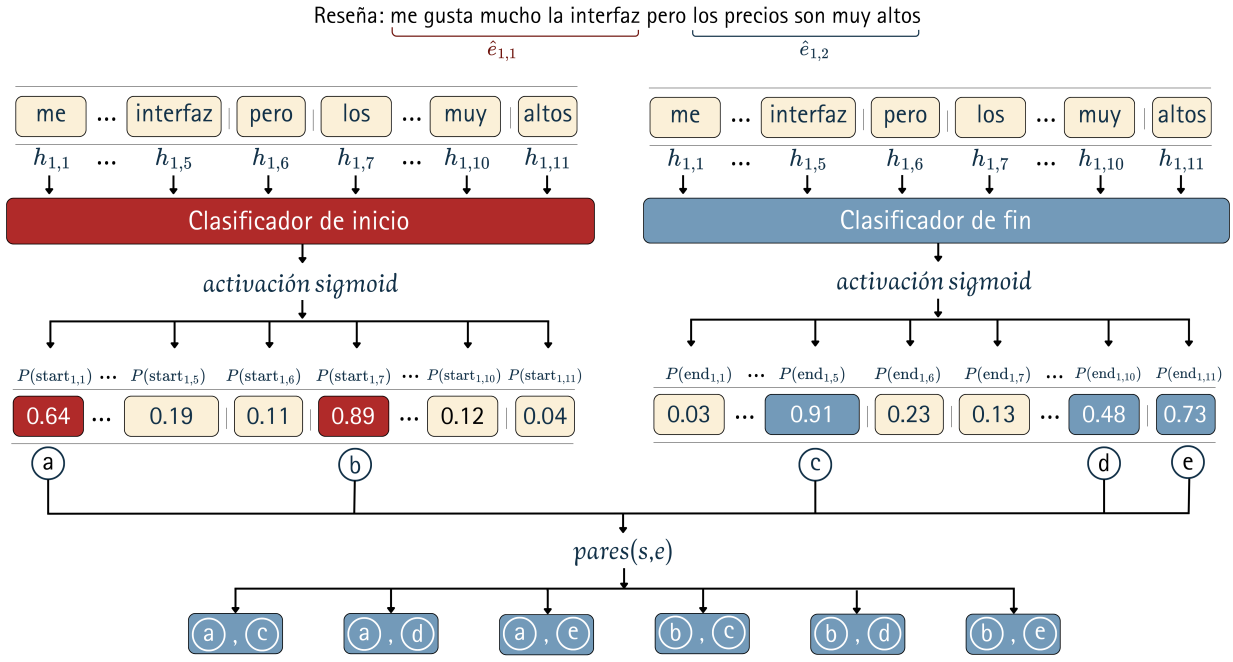
Este ajuste resulta particularmente útil en la tarea de detección de *spans*, ya que reduce la sobreconfianza del modelo en palabras más comunes y permite considerar un conjunto más diverso de candidatos de inicio y fin del *span*. Formalmente, las probabilidades de inicio y fin respectivamente se definen como:

$$P(\text{start}_{b,t}) = \sigma\left(\frac{\ell^{\text{start}}_{b,t}}{T}\right), \qquad P(\text{end}_{b,t}) = \sigma\left(\frac{\ell^{\text{end}}_{b,t}}{T}\right).$$

La Figura 13 presenta la construcción de un conjunto reducido de fragmentos candidatos, seleccionando las posiciones de inicio (ⓐ y ⓑ) y fin (ⓒ, ⓓ y ⓔ) más probables. A partir de estas posiciones, se generan pares (s, e) tales que $s \leq e$, que cumplan: $P(\text{s}) \geq \mu$ y $P(\text{e}) \geq \lambda$ (véase Tabla 6).

Figura 13

Primera sección del módulo de detección de spans



Nota. La figura presenta el flujo de trabajo de la primera sección del módulo de inferencia de *spans* compuesto por las dos cabezas clasificadoras.

Para cada candidato (s, e) de la reseña b, se propone un puntaje heurístico inspirado en enfoques de detección de fragmentos en arquitecturas *Transformer* (Devlin et al., 2019), así como en propiedades del espacio latente generado por el modelo. En particular, este puntaje extiende la combinación tradicional de probabilidades de inicio y fin mediante la incorporación de criterios adicionales orientados a capturar la coherencia semántica del fragmento y a regular su longitud. De esta forma, se obtiene una función que integra tres componentes complementarios. Además, cada uno de estos componentes se encuentra acotado en el intervalo [0,1], por lo que el puntaje resultante también pertenece a este rango.

$$\text{score}(b, s, e) = \underbrace{\frac{P(\text{start}_{b,s}) + P(\text{end}_{b,e})}{2}}_{\text{confianza}} \cdot \underbrace{\frac{1 + \cos(h_{b,s}, h_{b,e})}{2}}_{\text{coherencia}} \cdot \underbrace{\frac{1}{1 + \alpha \cdot (e - s)}}_{\text{penalización de longitud}}$$

El primer término corresponde a una medida de confianza en los límites del *span*. En tareas de extracción de fragmentos, es común asumir independencia entre las predicciones de inicio y fin y combinar ambas probabilidades para evaluar candidatos (Devlin et al., 2019). En este caso, se utiliza el promedio en lugar del producto para obtener una estimación más estable, evitando que una única predicción con baja probabilidad anule completamente el puntaje del *span*.

El segundo término corresponde a una medida de coherencia semántica basada en la similitud coseno entre los estados ocultos asociados a los tokens de inicio y fin del candidato. En particular, $\cos(h_{b,s}, h_{b,e})$ representa el coseno del ángulo formado entre los vectores de estados ocultos $h_{b,s}$ y $h_{b,e}$ en el espacio de representaciones aprendido por el modelo. Esta medida permite cuantificar el grado de alineación semántica entre ambos extremos del segmento: valores cercanos a 1 indican que los vectores apuntan en direcciones similares, sugiriendo una alta coherencia contextual entre el *token* inicial y el *token* final; valores cercanos a 0 reflejan una relación semántica débil; mientras que valores negativos indican direcciones opuestas y, por tanto, una baja compatibilidad semántica.

Dado que la similitud coseno toma valores en el intervalo $[-1, 1]$, se aplica la transformación $\frac{1+\cos(h_{b,s}, h_{b,e})}{2}$ para normalizar el resultado al rango $[0, 1]$. De esta manera, los candidatos cuyos extremos presentan representaciones contextuales más alineadas reciben una mayor puntuación de coherencia, favoreciendo la selección de segmentos semánticamente consistentes.

El tercer término corresponde a una penalización de longitud suave. Durante el análisis del comportamiento del modelo, se identificó un sesgo hacia la selección de *spans* excesivamente largos, incluso en casos donde la reseña contenía múltiples *spans* más cortos. Para mitigar este efecto, se introduce de manera heurística un término de penalización de longitud sobre el *score* de cada candidato. Esta penalización está controlada por el hiperparámetro α (véase Tabla 6) , el cual regula la intensidad con la que se castigan los *spans* a medida que aumenta su longitud. De este modo, se favorece un equilibrio en el proceso de selección: los *spans* largos pueden seguir siendo elegidos cuando existe evidencia suficiente, pero se evita que dominen sistemáticamente frente a alternativas más compactas y localizadas.

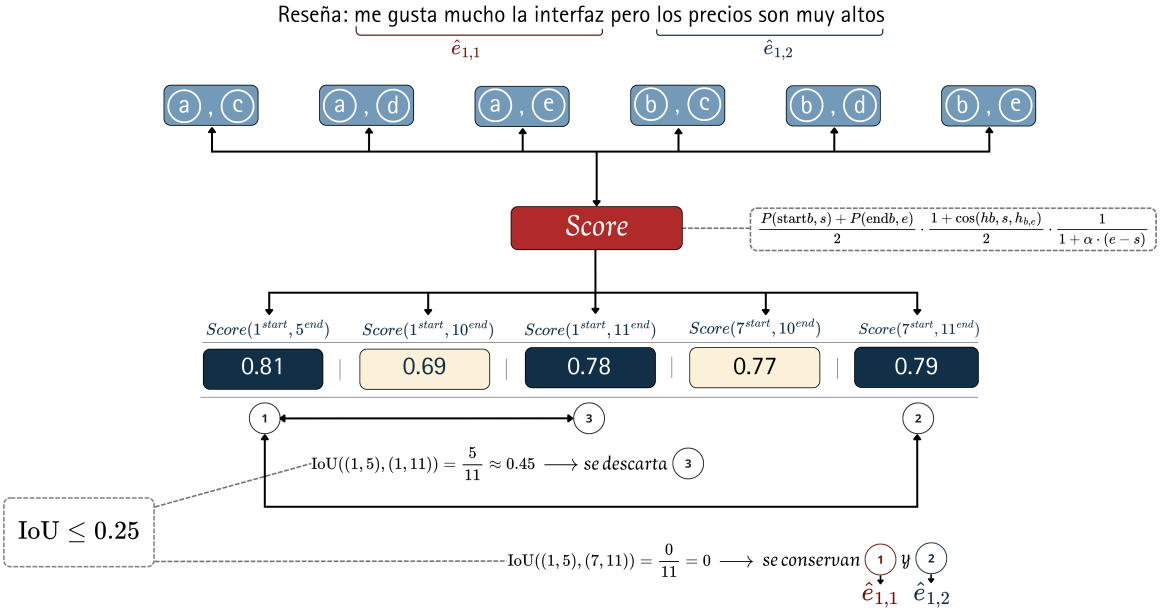
Una vez calculados los *score*, los *spans* candidatos se ordenan de forma descendente y se filtran mediante un procedimiento de supresión de no-máximos (NMS, del inglés *Non-Maximum Suppression*) basado en la intersección sobre unión (IoU, del inglés *Intersection of Union*) a nivel de caracteres. Este proceso ordena los candidatos de acuerdo con su puntaje de manera descendente. El *span* con la puntuación más alta es seleccionado automáticamente como *evidence span* definitivo, los candidatos restantes son evaluados secuencialmente, eliminando aquellos que presentan un alto grado de solapamiento con alguno de los *spans* previamente aceptados. Como resultado, se obtiene un conjunto final de evidencias que equilibra precisión en la localización, coherencia semántica y diversidad estructural dentro de la reseña. De esta manera, para cada reseña b se obtiene un conjunto de *spans* de evidencia definitivos, denotado como

$$\hat{e}_{b,k} = (t_s, t_e),$$

donde $\hat{e}_{b,k}$ representa el k-ésimo *span* detectado en la reseña b.

En la Figura 14 se ilustra un ejemplo del funcionamiento de la segunda sección del módulo de detección de *spans*. En este caso, tras la aplicación de la función *score*, la reseña genera tres *spans* candidatos, identificados como ①, ② y ③. De acuerdo con la configuración del modelo, un *span* candidato es aceptado únicamente si, al compararse con el *span* candidato de mayor puntaje, su valor de IoU es igual o inferior al umbral establecido (véase Tabla 6). En el ejemplo presentado, el *span* ③ excede dicho umbral, por lo que es descartado, mientras que ① y ② se conservan como *spans* definitivos.

Figura 14
Segunda sección del módulo de detección de spans



Nota. La figura presenta el flujo de trabajo de la segunda sección del módulo de detección de *spans* en la que se obtienen los *spans* candidatos y definitivos.

3.2.1.5 Representación vectorial del fragmento de evidencia. Una vez identificado un *evidence span* definido por los *tokens* $[t_s, t_e]$, se construye una representación vectorial agregada a partir de los estados ocultos correspondientes. Para ello se aplica una operación de *mean pooling*

sobre las representaciones contextuales generadas por el *encoder*:

$$v_{b,k} = \frac{1}{t_e - t_s + 1} \sum_{t=t_s}^{t_e} h_{b,t} \in \mathbb{R}^d.$$

Esta operación produce un vector que resume la información contextual de todos los *tokens* que conforman el fragmento de evidencia, permitiendo obtener una representación densa del *evidence span* completo en el espacio latente del modelo.

Con el fin de incorporar también información global de la reseña, el vector del *evidence span* se concatena con la representación p_b obtenida a partir del *token* especial `[CLS]`:

$$f_{b,k} = [\, v_{b,k} \| p_b \,] \in \mathbb{R}^{2d},$$

donde $p_b \in \mathbb{R}^d$ corresponde a la representación contextual de la reseña completa.

El vector resultante $f_{b,k}$ integra tanto información local del fragmento de evidencia como contexto global de la reseña, y se utiliza como entrada para las posteriores cabezas de clasificación.

3.2.1.6 Cabeza clasificadora de aspectos. El primer nivel de clasificación semántica consiste en predecir los aspectos asociados al fragmento de evidencia. Dado que un mismo fragmento puede estar relacionado con múltiples aspectos, esta tarea se modela como un problema de clasificación de etiquetas múltiples (del inglés, *multi-label*).

A partir de la representación del *span* $f_{b,k}$, los *logits* para cada aspecto se calculan mediante una transformación lineal:

$$z^{\text{asp}}_{b,k} = W_{\text{asp}} f_{b,k} + b_{\text{asp}} \in \mathbb{R}^{|\mathcal{A}|},$$

donde $W_{\text{asp}} \in \mathbb{R}^{|\mathcal{A}| \times 2d}$ es la matriz de pesos, $b_{\text{asp}} \in \mathbb{R}^{|\mathcal{A}|}$ es el vector de sesgos y $|\mathcal{A}|$ corresponde al número total de aspectos considerados por el modelo.

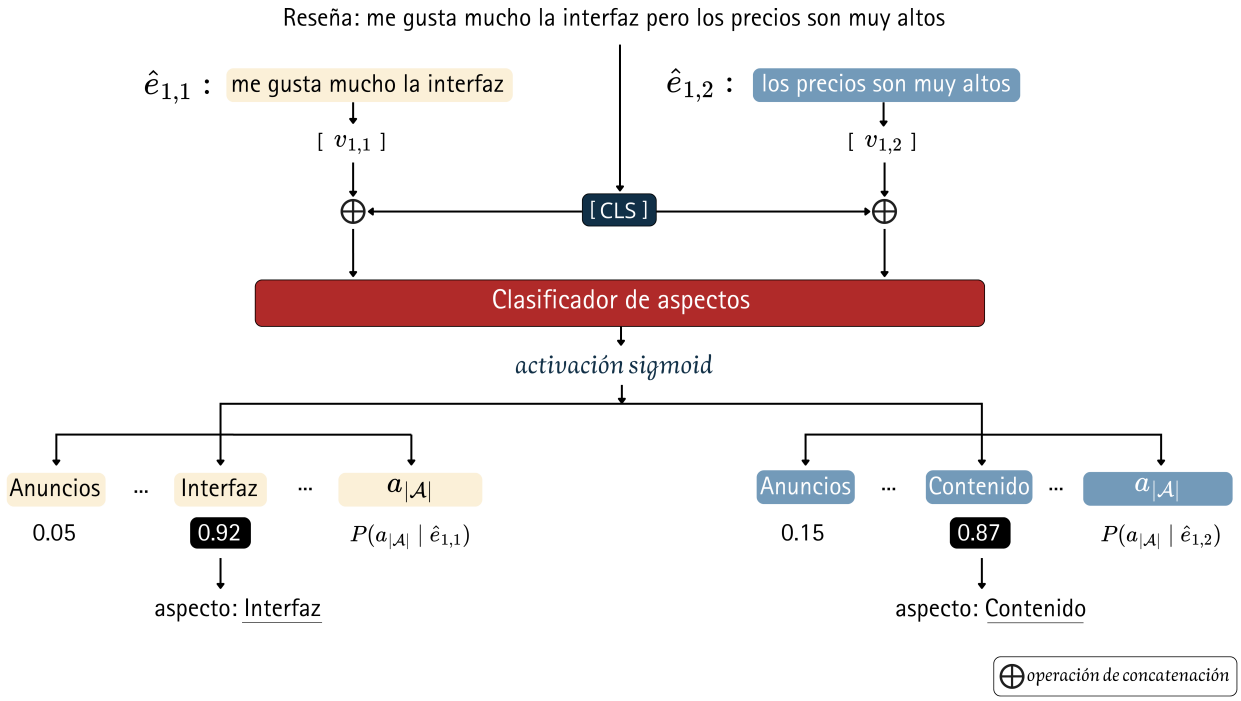
Posteriormente, se aplica una función *sigmoid* de forma independiente a cada componente para obtener la probabilidad de que el *evidence span* esté asociado a cada aspecto:

$$P(a_j \mid f_{b,k}) = \sigma(z^{\text{asp}}_{b,k,j}), \qquad j = 1, \ldots, |\mathcal{A}|.$$

De esta manera, el modelo puede asignar simultáneamente múltiples aspectos a un mismo fragmento de evidencia. Durante la inferencia, los aspectos con probabilidad superior a un umbral fijado se seleccionan (véase Tabla 6).

Siguiendo con el ejemplo de la Figura 14, en la Figura 15 se ilustra el proceso de clasificación de aspectos. A la cabeza clasificadora ingresa la representación agregada v de los *evidence spans*, cada uno concatenado con la representación global de la reseña, correspondiente al *token* [CLS]. Tras la aplicación de la función de activación *sigmoid*, se obtiene una probabilidad independiente para cada uno de los aspectos definidos en la guía de anotación (véase Tabla 3). Con base en estas probabilidades, el modelo selecciona el aspecto o los aspectos que mejor se ajustan a cada *span* en base a un umbral definido (veáse Tabla 6).

Figura 15
Módulo de detección de spans



Nota. La figura presenta el flujo de trabajo de la cabeza clasificadora de aspectos.

3.2.1.7 Cabeza clasificadora de categorías. La predicción de categorías se realiza de forma condicionada al aspecto detectado. Para ello, el modelo incorpora *embeddings* que representan los distintos aspectos y categorías:

$$E_a \in \mathbb{R}^{|\mathcal{A}| \times d}, \qquad E_c \in \mathbb{R}^{|C| \times d}.$$

Durante el entrenamiento, la representación del *evidence span* se concatena con el *embedding* correspondiente al aspecto real de la reseña utilizando la técnica de forzado del profesor (del inglés, *teacher forcing*), en la cual la salida correcta real (conocida en la literatura como *ground truth*) de un paso anterior se utiliza como entrada para el siguiente paso, en lugar de la propia predicción del modelo. Así pues, el vector que ingresa a la cabeza clasificadora está definido como:

$$u_{b,k}^{\text{cat}} = [\, f_{b,k} \| E_a[a_{b,k}^{\text{gold}}] \,] \in \mathbb{R}^{3d},$$

donde $E_a[a^{\text{gold}}]$ corresponde el *embedding* del aspecto presente en la anotación real *ground truth* de la etiqueta del *evidence span*, que concatenado con el vector $f_{b,k} \in \mathbb{R}^{2d}$ da lugar a un vector de dimensión $\mathbb{R}^{3d}$. Esta concatenación permite incorporar información semántica explícita sobre el aspecto al momento de predecir la categoría. Intuitivamente, el *embedding* de aspecto actúa como un vector de contexto que guía al modelo hacia las categorías más plausibles para dicho aspecto, reduciendo la ambigüedad inherente a la clasificación. Por ejemplo, un mismo fragmento de texto puede ser compatible con múltiples categorías en ausencia de contexto; sin embargo, al condicionar la predicción en el aspecto, el modelo restringe el espacio de búsqueda a aquellas categorías coherentes con dicho aspecto.

La predicción de categorías se obtiene mediante

$$z_{b,k}^{\text{cat}} = W_{\text{cat}} u_{b,k}^{\text{cat}} + b_{\text{cat}} \in \mathbb{R}^{|C|},$$

donde $W_{\text{cat}} \in \mathbb{R}^{|C| \times 3d}$ y $b_{\text{cat}} \in \mathbb{R}^{|C|}$ son los parámetros aprendidos del clasificador.

Luego, cada categoría se modela de forma independiente mediante una función *sigmoid*:

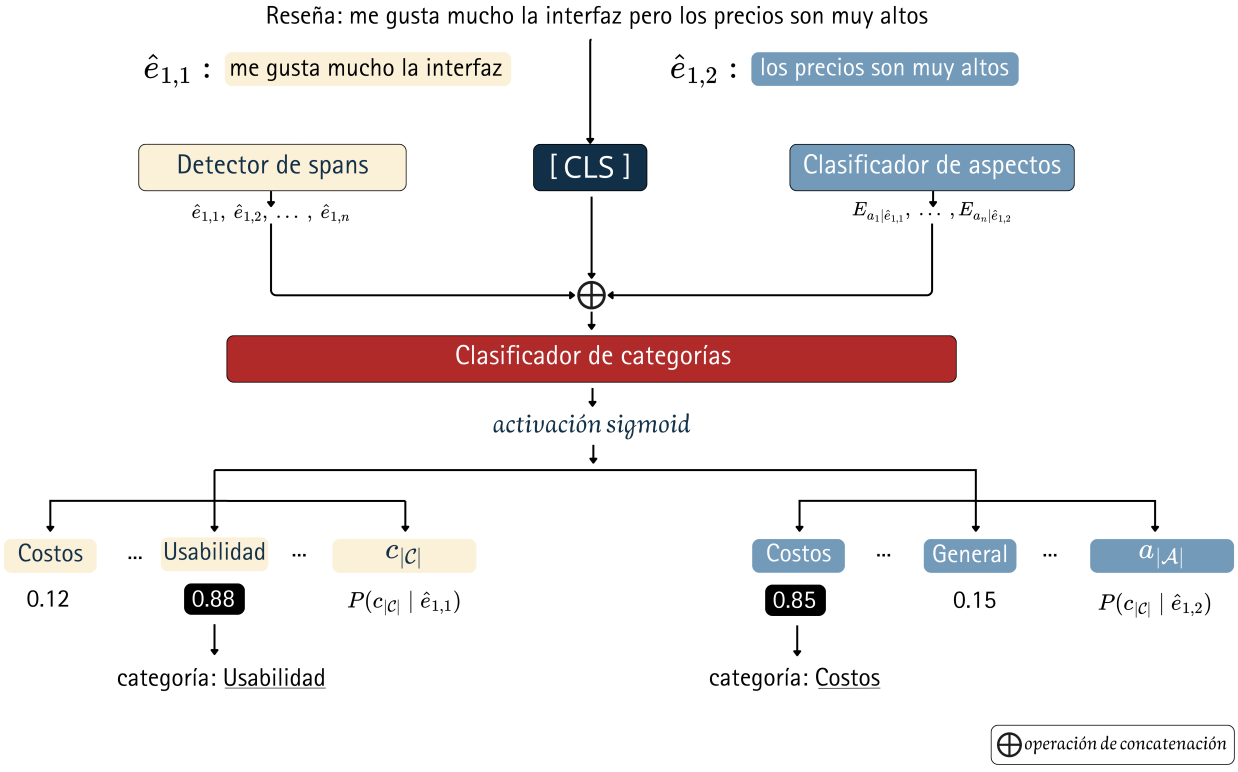
$$P(c_j \mid f_{b,k}, a) = \sigma(z_{b,k,j}^{\text{cat}}), \qquad j = 1, \ldots, |C|.$$

Finalmente, durante la inferencia, el *embedding* utilizado para condicionar esta predicción corresponde al aspecto previamente estimado por el modelo.

En la Figura 16 se ilustra el funcionamiento del módulo de clasificación de categorías, comenzando por la estructura de la entrada, compuesta por los *spans* detectados, los aspectos asociados y la representación global de la reseña. Tras la aplicación de la función de activación *sigmoid*, se obtienen probabilidades para cada una de las categorías definidas en la guía de anotación

(véase Tabla 2). Con base en dichas probabilidades, el modelo selecciona aquellas categorías con mayor valor, en función de cada *span*.

Figura 16
Clasificador de categorías



Nota. La figura presenta el flujo de trabajo de la cabeza clasificadora de categorías.

3.2.1.8 Cabeza clasificadora de sentimientos. Finalmente, el modelo predice la polaridad asociada al fragmento de evidencia condicionando simultáneamente en la representación del *evidence span*, el aspecto y la categoría seleccionados. Al igual que en la predicción de categorías (véase Sección 3.2.1.7), se usa la técnica de *teacher forcing*.

Para ello se construye la siguiente representación conjunta:

$$U_{b,k} = [\, v_{b,k} \| E_a[a^{\text{gold}}_{b,k}] \| E_c[c^{\text{gold}}_{b,k}] \,] \in \mathbb{R}^{3d},$$

donde $v_{b,k}$ captura la información contextual del fragmento de evidencia, mientras que $E_a[a_{b,k}^{\text{gold}}]$ y $E_c[c_{b,k}^{\text{gold}}]$ corresponden a los *embeddings* pertenecientes al *ground truth* del aspecto y la categoría, respectivamente.

Esta concatenación permite modelar de forma explícita la dependencia entre sentimiento, aspecto y categoría. Intuitivamente, la polaridad de un mismo fragmento puede variar dependiendo del aspecto considerado, y la categoría introduce un nivel adicional de especialización semántica que facilita una predicción más precisa.

La predicción de sentimiento se obtiene mediante una transformación lineal:

$$z_{b,k}^{\text{sent}} = W_{\text{sent}} U_{b,k} + b_{\text{sent}} \in \mathbb{R}^{|\mathcal{S}|},$$

donde $W_{\text{sent}} \in \mathbb{R}^{|\mathcal{S}| \times 3d}$ y $b_{\text{sent}} \in \mathbb{R}^{|\mathcal{S}|}$ son los parámetros aprendidos del clasificador.

Posteriormente, se aplica una función *softmax* para obtener una distribución de probabilidad sobre las clases de sentimiento:

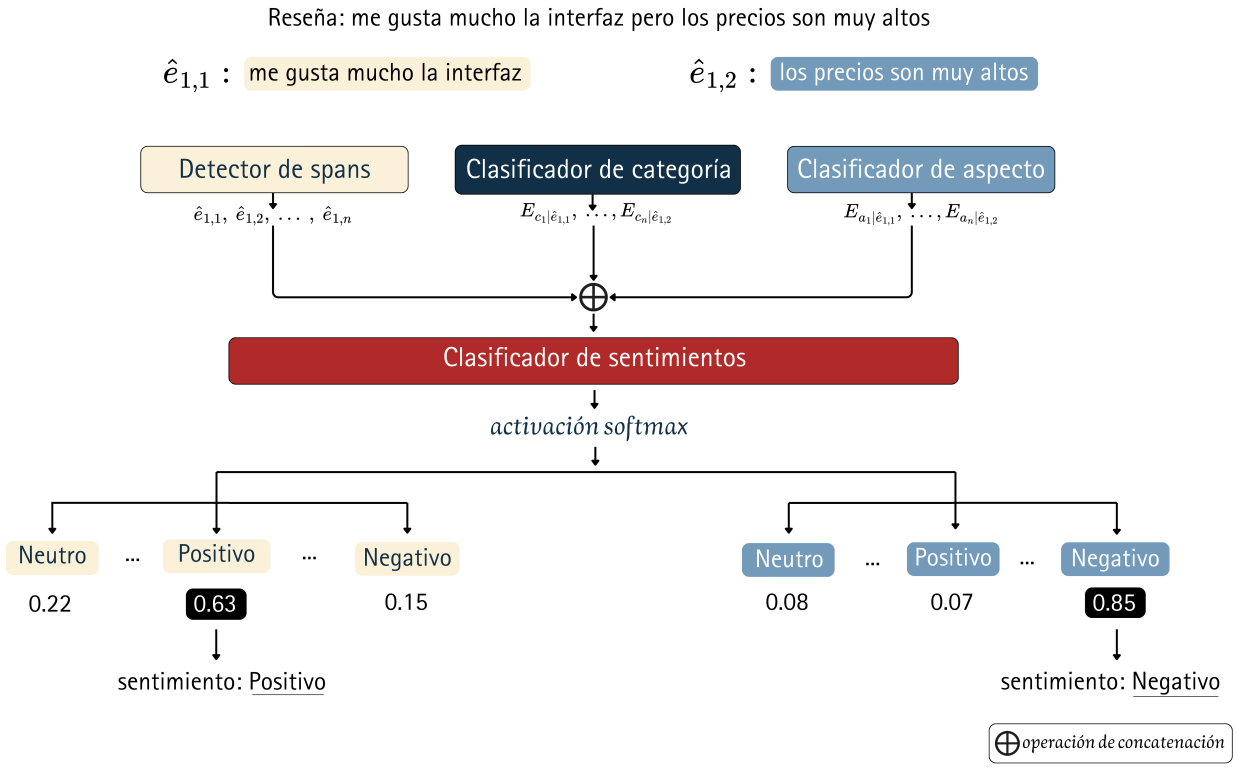
$$P(s_j \mid e_{b,k}, a, c) = \frac{\exp(z_{b,k,j}^{\text{sent}})}{\sum_{q=1}^{|\mathcal{S}|} \exp(z_{b,k,q}^{\text{sent}})}, \qquad j = 1, \ldots, |\mathcal{S}|.$$

Durante el entrenamiento, se emplea *teacher forcing*, mientras que en inferencia se utilizan las predicciones generadas en las etapas anteriores del modelo.

En la Figura 17 se ilustra el funcionamiento del módulo de clasificación de sentimientos, comenzando por la estructura de la entrada. A diferencia de los módulos anteriores, los *evidence spans* no se concatenan con la representación global de la reseña; en su lugar, la entrada está compuesta por los *evidence spans* detectados, junto con los aspectos y categorías asociados. En este caso, dado que a cada *evidence span* se le asigna una única etiqueta de sentimiento, se emplea la función de activación *softmax*. Como resultado, se obtiene una distribución de probabilidad sobre cada una de las polaridades definidas (véase Tabla 4), cuya suma es igual a 1. Finalmente,

se selecciona la polaridad con mayor probabilidad para cada *evidence span*, obteniendo así la clasificación total de cada *evidence span* y su tripleta asociada.

Figura 17
Módulo de detección de sentimientos



Nota. La figura presenta el flujo de trabajo de la cabeza clasificadora de sentimientos.

3.2.1.9 Función de pérdida. El entrenamiento del modelo se realiza optimizando de manera conjunta las distintas tareas que componen la arquitectura. En primer lugar, la detección del *evidence span* se supervisa con dos pérdidas binarias independientes (BCE, del inglés *Binary Cross-Entropy*), una para el *token* de inicio y otra para el *token* de fin:

$$\mathcal{L}_{\text{span}} = \mathcal{L}_{\text{BCE}}\left(t^{\text{start}}, y^{\text{start}}\right) + \mathcal{L}_{\text{BCE}}\left(t^{\text{end}}, y^{\text{end}}\right),$$

donde y^{start} y y^{end} son las etiquetas del *ground truth* a nivel de *token*.

La predicción de aspectos y categorías se entrena con BCE sobre cada componente:

$$\mathcal{L}_{\text{asp}} = \mathcal{L}_{\text{BCE}}\big(z^{\text{asp}}, y^{\text{asp}}\big), \qquad \mathcal{L}_{\text{cat}} = \mathcal{L}_{\text{BCE}}\big(z^{\text{cat}}, y^{\text{cat}}\big),$$

donde y^{asp} y y^{cat} son vectores *one-hot* construidos a partir de las anotaciones de cada *evidence span*.

Por su parte, la polaridad del sentimiento se supervisa con entropía cruzada multicategoría (CCE, del inglés *Categorical Cross-Entropy*):

$$\mathcal{L}_{\text{sent}} = -\sum_{j=1}^{|\mathcal{S}|} y_j^{\text{sent}} \log\left(\frac{\exp(z_j^{\text{sent}})}{\sum_{q=1}^{|\mathcal{S}|}\exp(z_q^{\text{sent}})}\right).$$

En consecuencia, la función de pérdida total queda dada por

$$\mathcal{L} = \mathcal{L}_{\text{span}} + \mathcal{L}_{\text{asp}} + \mathcal{L}_{\text{cat}} + \mathcal{L}_{\text{sent}}.$$

3.3 Estrategias de entrenamiento y evaluación del modelo propuesto

En esta sección se describen los parámetros empleados durante la fase de entrenamiento, junto con sus respectivos valores, así como las métricas utilizadas para evaluar el rendimiento del modelo. Para conocer, de forma detallada, los códigos construidos y empleados para el entrenamiento y cálculo de métricas, véase Apéndice C.

3.3.1 Estrategia de entrenamiento

El modelo se entrenó mediante un esquema de *fine-tuning* del *encoder* BETO junto con las cabezas específicas para la clasificación de las reseñas (véase Sección 3.2). Durante la optimización se actualizaron todos los parámetros del modelo, incluyendo tanto las capas *Transformer* como las capas de salida encargadas de la detección de *spans* y la predicción de aspecto, categoría y

sentimiento.

Tabla 6
Hiperparámetros de entrenamiento y predicción del modelo ABSA centrado en evidencia

Parámetro	**Valor**
Modelo base	`beto-uncased`
Estrategia sobre el encoder	*Fine-tuning* completo
Longitud máxima de secuencia	128 *tokens*
Tamaño de lote (*batch size*)	8
Tasa de aprendizaje	2×10^{-5}
Número de épocas	5
Dropout	0,1
Optimizador	AdamW
Semilla global	42
Umbral probabilidad de inicio μ	0,35
Umbral probabilidad de fin λ	0,35
Factor de temperatura T	1,5
Umbral de NMS	0,25
Umbral de aspectos	0,6
Umbral de categorías	0,6
Penalización de longitud α	0,04

Nota. Hiperparámetros utilizados para el entrenamiento y las predicciones del modelo final.

Las reseñas fueron tokenizadas utilizando el tokenizador asociado a BETO y posteriormente truncadas o rellenadas hasta una longitud máxima de 128 *tokens*. El entrenamiento se realizó con un tamaño de lote de 8 ejemplos durante 5 épocas, empleando el optimizador *AdamW* con una tasa de aprendizaje de 2×10^{-5}. Con el fin de garantizar la reproducibilidad del experimento, se fijó una semilla global igual a 42. El resumen de estos hiperparámetros se reportan en la tabla 6.

La función objetivo del modelo se definió como la suma de las pérdidas asociadas a cada subtarea (véase Sección 3.2.1.9).

3.3.2 Estrategia de evaluación

La naturaleza del modelo propuesto, orientado a la identificación explícita de fragmentos de evidencia junto con sus correspondientes etiquetas semánticas, exige un esquema de evaluación que considere, de manera conjunta, tanto la calidad de la clasificación como la precisión en la localización del texto relevante. En este contexto, la evaluación no se restringe a la asignación correcta de etiquetas, sino que incorpora la verificación de que dichas etiquetas estén sustentadas por el segmento textual adecuado dentro de la reseña.

Con el fin de obtener una estimación robusta del desempeño del modelo y reducir la dependencia de una partición específica de los datos, se empleó un esquema de validación cruzada de k particiones (del inglés, *k-fold cross-validation*) con $k = 5$. En este procedimiento, el conjunto de datos se divide en cinco subconjuntos disjuntos de tamaño similar. En cada iteración, uno de los subconjuntos se utiliza como conjunto de evaluación, mientras que los restantes se emplean para el entrenamiento del modelo.

Este proceso se repite cinco veces, de modo que cada subconjunto actúa exactamente una vez como conjunto de prueba. Las métricas descritas en las secciones siguientes se calculan de manera independiente en cada *fold*, y los resultados finales se reportan como el promedio de dichas métricas junto con su desviación estándar, proporcionando así una estimación más estable y representativa del comportamiento del modelo.

Bajo este enfoque, el proceso de evaluación se estructura en cuatro niveles complementarios: i) evaluación de la localización de evidencia, ii) métricas base de clasificación, iii) clasificación condicionada al span y iv) evaluación estricta de cuádruplas.

3.3.2.1 Evaluación de localización de evidencia. El primer nivel de evaluación mide la capacidad del modelo para localizar correctamente los fragmentos de texto que sustentan una opinión. Para ello, cada *evidence span* predicho $\hat{e}_{i,j}$ se compara con los *evidence spans* reales $e_{i,k}$

de la misma reseña mediante la métrica *Intersection over Union* (IoU), definida como

$$\text{IoU}(\hat{e}_{i,j}, e_{i,k}) = \frac{|\hat{e}_{i,j} \cap e_{i,k}|}{|\hat{e}_{i,j} \cup e_{i,k}|}.$$

En esta formulación, $\hat{e}_{i,j}$ y $e_{i,k}$ se interpretan como intervalos sobre *offsets* de caracteres. Un *evidence span* predicho se considera correctamente localizado en evaluación aproximada si existe un *span* real tal que

$$\text{IoU}(\hat{e}_{i,j}, e_{i,k}) \geq 0{,}50.$$

Además de esta evaluación aproximada, se emplea una evaluación exacta de spans, en la cual un *span* predicho se considera correcto únicamente si coincide exactamente con el *span* real:

$$\hat{e}_{i,j} = e_{i,k}.$$

De esta manera, se evalúa tanto la capacidad del modelo para aproximar la región textual relevante como su precisión en la delimitación exacta del fragmento de evidencia.

3.3.2.2 Métricas base de clasificación. Para evaluar la calidad de las predicciones semánticas del modelo, se emplean las métricas de *precision*, *recall* y *F1-score*. Estas métricas se calculan de manera consistente en las distintas subtareas de clasificación consideradas en este trabajo, a saber: predicción de aspecto, categoría, sentimiento y la combinación categoría–sentimiento.

Dado que cada una de estas subtareas puede involucrar múltiples clases, las métricas se definen inicialmente a nivel de clase individual. Para una clase i, se tiene:

3.3.2.2.1 Precision. La precisión mide la proporción de instancias clasificadas como pertenecientes a una clase que efectivamente corresponden a dicha clase:

$$\text{Precision}_i = \frac{TP_i}{TP_i + FP_i},$$

donde TP_i y FP_i representan los verdaderos positivos y falsos positivos, respectivamente. Una alta precisión indica que el modelo comete pocos errores al asignar la clase i.

3.3.2.2.2 Recall. El recall evalúa la proporción de instancias reales de una clase que son correctamente identificadas por el modelo:

$$\text{Recall}_i = \frac{TP_i}{TP_i + FN_i},$$

donde FN_i denota los falsos negativos. Esta métrica es especialmente relevante cuando se desea minimizar la omisión de instancias positivas.

3.3.2.2.3 F1-score. El *F1-score* combina precisión y recall mediante su media armónica:

$$F1_i = 2 \cdot \frac{\text{Precision}_i \cdot \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i}.$$

Esta métrica proporciona una medida balanceada del desempeño, penalizando escenarios en los que una de las dos componentes es significativamente inferior.

En el caso de las subtareas de aspecto y categoría, que se modelan como problemas *multi-label*, estas métricas se calculan de forma independiente para cada clase. Para la subtarea de sentimiento, que corresponde a una clasificación *single-label*, el cálculo sigue la formulación estándar.

3.3.2.2.4 Promedios micro y macro. Dado que el desempeño global del modelo depende de su comportamiento conjunto sobre todas las clases, las métricas definidas a nivel individual se

agregan mediante esquemas de promediado. En particular, se emplean los promedios *micro* y *macro*, los cuales se calculan de manera independiente para cada subtarea de clasificación, permitiendo caracterizar el desempeño del modelo en cada nivel semántico.

El promedio *micro* se obtiene acumulando los verdaderos positivos, falsos positivos y falsos negativos a través de todas las clases, para luego calcular las métricas globales:

$$\text{Precision}_{micro} = \frac{\sum_i TP_i}{\sum_i (TP_i + FP_i)},$$

$$\text{Recall}_{micro} = \frac{\sum_i TP_i}{\sum_i (TP_i + FN_i)},$$

$$F1_{micro} = 2 \cdot \frac{\text{Precision}_{micro} \cdot \text{Recall}_{micro}}{\text{Precision}_{micro} + \text{Recall}_{micro}}.$$

Por su parte, el promedio *macro* se calcula promediando las métricas obtenidas de manera independiente para cada clase, asignando el mismo peso a todas ellas:

$$\text{Precision}_{macro} = \frac{1}{N} \sum_{i=1}^{N} \text{Precision}_i,$$

$$\text{Recall}_{macro} = \frac{1}{N} \sum_{i=1}^{N} \text{Recall}_i,$$

$$F1_{macro} = \frac{1}{N} \sum_{i=1}^{N} F1_i.$$

Es importante destacar que estos esquemas de agregación se aplican de forma independiente en cada subtarea (aspecto, categoría, sentimiento y categoría–sentimiento), lo que permite analizar el desempeño del modelo tanto desde una perspectiva global como en términos de su comportamiento frente a clases específicas.

Finalmente, en coherencia con el enfoque basado en evidencia del modelo, todas las métricas de clasificación se calculan únicamente sobre aquellos *spans* que han sido previamente emparejados con un *span* real mediante el criterio de IoU. Por tanto, estas métricas no deben interpretarse como medidas globales del sistema completo, sino como una evaluación condicionada al correcto posicionamiento del fragmento de evidencia.

3.3.2.3 Clasificación condicionada al fragmento de evidencia. Con el fin de analizar de manera aislada el comportamiento de las cabezas semánticas del modelo, se evaluaron las predicciones de aspecto, categoría y sentimiento condicionadas a que el *span* de evidencia haya sido correctamente emparejado mediante IoU.

Sea $\mathcal{M}_i$ el conjunto de emparejamientos entre predicciones y anotaciones reales en la reseña r_i, definido como

$$\mathcal{M}_i = \{(\hat{\ell}_{i,j}, \ell_{i,k}) : \mathrm{IoU}(\hat{e}_{i,j}, e_{i,k}) \geq 0{,}50\}.$$

Sobre este conjunto se definen las siguientes exactitudes condicionadas:

$$\mathrm{Acc}_{\mathrm{asp|span}} = \frac{\sum_i \sum_{(\hat{\ell}_{i,j},\ell_{i,k})\in\mathcal{M}_i} \mathbb{I}(\hat{a}_{i,j} = a_{i,k})}{\sum_i |\mathcal{M}_i|},$$

$$\mathrm{Acc}_{\mathrm{cat|span}} = \frac{\sum_i \sum_{(\hat{\ell}_{i,j},\ell_{i,k})\in\mathcal{M}_i} \mathbb{I}(\hat{c}_{i,j} = c_{i,k})}{\sum_i |\mathcal{M}_i|},$$

$$\mathrm{Acc}_{\mathrm{sent|span}} = \frac{\sum_i \sum_{(\hat{\ell}_{i,j},\ell_{i,k})\in\mathcal{M}_i} \mathbb{I}(\hat{s}_{i,j} = s_{i,k})}{\sum_i |\mathcal{M}_i|},$$

$$\mathrm{Acc}_{\mathrm{cat+sent|span}} = \frac{\sum_i \sum_{(\hat{\ell}_{i,j},\ell_{i,k})\in\mathcal{M}_i} \mathbb{I}(\hat{c}_{i,j} = c_{i,k} \wedge \hat{s}_{i,j} = s_{i,k})}{\sum_i |\mathcal{M}_i|},$$

donde $\mathbb{I}(\cdot)$ es la función indicadora.

Estas métricas permiten evaluar la calidad de las predicciones semánticas de manera desacoplada del error de localización, proporcionando una visión más detallada del comportamiento

de cada componente del modelo.

3.3.2.4 Evaluación estricta de cuádruplas. La métrica principal del sistema corresponde a la evaluación estricta de cuádruplas, ya que resume el comportamiento *end-to-end* del modelo. Una predicción $\hat{\ell}_{i,j}$ se considera correcta únicamente si existe una anotación real $\ell_{i,k}$ tal que se cumplan simultáneamente las siguientes condiciones:

$$\mathrm{IoU}(\hat{e}_{i,j}, e_{i,k}) \geq 0{,}50, \qquad \hat{a}_{i,j} = a_{i,k}, \qquad \hat{c}_{i,j} = c_{i,k}, \qquad \hat{s}_{i,j} = s_{i,k}.$$

Es decir, una predicción cuenta como verdadera positiva únicamente cuando coincide tanto el *evidence span* como las etiquetas de aspecto, categoría y sentimiento. Si el *evidence span* logra emparejarse con una anotación real mediante IoU, pero alguna de las etiquetas no coincide, la predicción no se considera correcta bajo este criterio estricto. En consecuencia, esta métrica constituye la evaluación más exigente del sistema, al requerir la correcta identificación simultánea del fragmento de evidencia y de todas las etiquetas semánticas asociadas.

3.4 Desarrollo de un prototipo web para la visualización de resultados del modelo

Con el fin de complementar la etapa experimental y facilitar la interpretación de los resultados del modelo, se desarrolló un prototipo web orientado a la visualización interactiva de las predicciones generadas. La motivación principal de esta implementación fue disponer de una herramienta que permitiera una representación comprensible para el usuario final, en la que fuera posible inspeccionar reseñas individuales, observar tendencias globales y analizar los resultados mediante un panel tipo *dashboard*.

El prototipo se concibió como una interfaz de apoyo al proceso de evaluación del modelo. Por esta razón, se priorizó una aplicación ligera, reproducible y alineada con el ecosistema técnico utilizado durante las etapas de entrenamiento e inferencia.

3.4.1 Arquitectura del prototipo

Desde el punto de vista funcional, la aplicación fue diseñada bajo una arquitectura de tipo página única (SPA, del inglés *Single Page Application*). En este enfoque, todo el flujo de interacción se centraliza en una única interfaz que integra la carga de datos y la visualización de resultados.

La arquitectura lógica del prototipo se organiza en tres capas principales:

3.4.1.1 Capa de presentación. Corresponde a la interfaz de usuario, encargada de la interacción directa con el sistema. Incluye el módulo de carga de reseñas, el panel de métricas globales, las tablas de resultados y la visualización de evidencias resaltadas sobre el texto original.

3.4.1.2 Capa de orquestación. Actúa como intermediaria entre la interfaz y el módulo de inferencia. En esta capa se gestiona la entrada de datos, se valida el formato de las reseñas (texto plano o CSV), se coordina la ejecución del *pipeline* de inferencia y se transforman las salidas del modelo en estructuras tabulares y visualizaciones interpretables.

3.4.1.3 Capa de inferencia. La capa de inferencia encapsula el flujo completo de procesamiento del modelo. En esta etapa se realiza la carga del *checkpoint* entrenado, el procesamiento y tokenización de las reseñas de entrada, y la ejecución del modelo sobre dichas representaciones. Posteriormente, se lleva a cabo la decodificación de los *evidence spans* a partir de las predicciones del modelo, así como la asignación de etiquetas correspondientes a aspectos, categorías y polaridad de sentimiento.

3.4.2 Tecnologías empleadas

El desarrollo del prototipo se apoyó en un conjunto de tecnologías seleccionadas con el objetivo de dar soporte a las distintas capas de la arquitectura descrita previamente, abarcando desde la interfaz de usuario hasta el módulo de inferencia del modelo.

La tecnología principal utilizada para la construcción de la interfaz fue *Streamlit*. Este *framework* permitió implementar una aplicación web completamente desarrollada en *Python*, lo cual resultó consistente con el ecosistema utilizado en el entrenamiento e inferencia del modelo. Adicionalmente, su soporte nativo para componentes interactivos como formularios, tablas, métricas y gráficos facilitó la implementación de la capa de presentación sin requerir la separación del *frontend* y el *backend*.

Para el manejo y procesamiento de los datos en la capa de orquestación se empleó *Pandas*, el cual permitió la carga de archivos en formato CSV, la estructuración tabular de las predicciones y la agregación de resultados para su posterior visualización en el *dashboard*. Este componente actúa como puente entre la entrada del usuario y las salidas generadas por el modelo.

Finalmente, la capa de inferencia se implementó utilizando *PyTorch* y *Transformers*, bibliotecas que proporcionan las herramientas necesarias para la carga del *checkpoint* entrenado, el procesamiento de las reseñas y la ejecución del modelo. Estas tecnologías permiten encapsular el flujo de predicción dentro de un pipeline reproducible, que posteriormente alimenta la capa de presentación.

Cabe destacar que el prototipo desarrollado se encuentra diseñado para su ejecución en un entorno local (del inglés, *localhost*), lo que significa que su despliegue está orientado a fines demostrativos y de validación experimental.

4. Resultados

4.1 Base de datos obtenida

El conjunto de datos obtenido constituye el núcleo experimental de este trabajo. Este corpus integra tanto reseñas reales como datos sintéticos, los cuales fueron anotados siguiendo un esquema estructurado (véase Sección 3.1.1). Para visualizar detalladamente el conjunto de datos construido, véase Apéndice C. A continuación, se describen en detalle las propiedades estructurales

y estadísticas del corpus construido.

4.1.1 Formato del corpus

Cada registro del conjunto de datos contiene un campo *text*, correspondiente al contenido completo de la reseña, y un campo *label*, que agrupa las anotaciones asociadas a dicha instancia. Este último está compuesto por una lista de etiquetas, donde cada una representa una tripleta semántica anotada. En particular, cada etiqueta incluye el campo *aspect*, que indica el aspecto o identificado en el *evidence span* (véase Tabla 3); *evidence_span*, que corresponde a la posición del segmento textual donde se evidencia la opinión, representado como un par de índices de carácter (inicio y fin); *sentiment*, que denota la polaridad asociada al *evidence span*, pudiendo ser positiva (POS), negativa (NEG) o neutra (NEU); y *category*, que clasifica la opinión dentro de una dimensión específica (véase Tabla 2). Esta estructura permite capturar de manera explícita la relación entre el fragmento textual relevante y sus respectivas dimensiones semánticas, facilitando el modelado de tareas de análisis de sentimiento basado en aspectos.

En la Figura 18 se presenta un ejemplo de anotación extraída del conjunto de datos construido.

Figura 18
Ejemplo de instancia anotada en formato JSON

```
{
  "text": "desde que se actualizo no me deja ver las notificaciones",
  "label": [
    {
      "aspect": ["notificaciones"],
      "evidence_span": [4, 9],
      "sentiment": "NEG",
      "category": "funcionalidad"
    }
  ]
}
```

Nota. La figura presenta un ejemplo de anotación presente en el corpus, incluyendo aspecto, categoría, sentimiento y delimitación de *evidence span* a nivel de palabra.

4.1.2 Características estructurales del corpus

A continuación, se presentan las principales características cuantitativas del conjunto de datos, incluyendo su distribución, anotaciones y propiedades estructurales más relevantes.

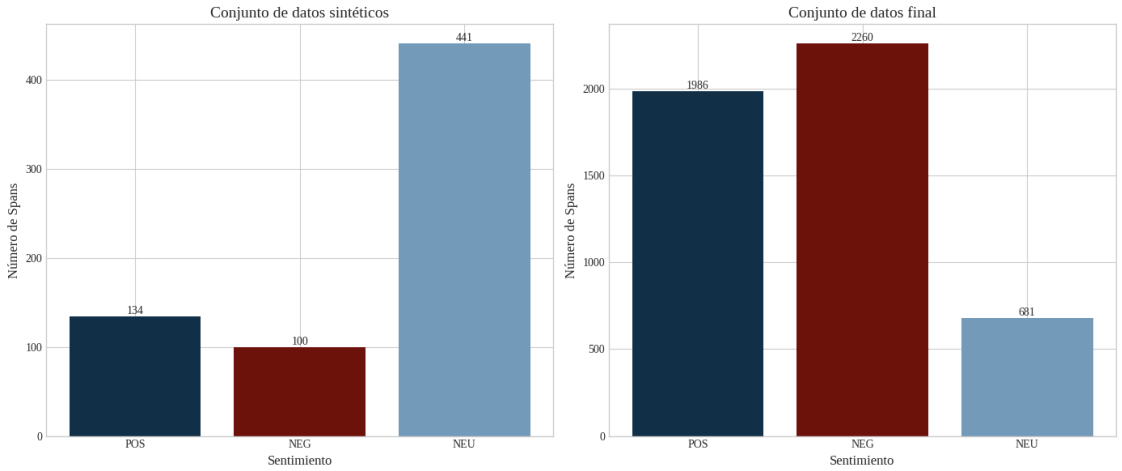
4.1.2.1 Contenido real versus contenido sintético. El corpus construido está conformado por un total de **3.493** reseñas, de las cuales **2.969** corresponden a reseñas reales y **524** a datos sintéticos generados, lo que equivale aproximadamente a un **85 %** de contenido real frente a un **15 %** de contenido sintético.

Tal como se expuso en la Sección 3.1.5.1, el conjunto de datos incial presentaba una subrepresentación considerable en cuanto *evidence spans* con polaridad neutral. Como resultado de la generación de datos sintéticos, obtuvo un conjunto de datos más balanceado, con una representación de sentimientos más homogénea. En la Figura 19 se presenta la distribución de la frecuencia de cada sentimiento de acuerdo con los *evidence spans* identificados, comparando el conjunto de

datos sintéticos con el conjunto final.

Figura 19

Distribución de fragmentos de evidencia por sentimiento: conjunto sintético vs. conjunto final

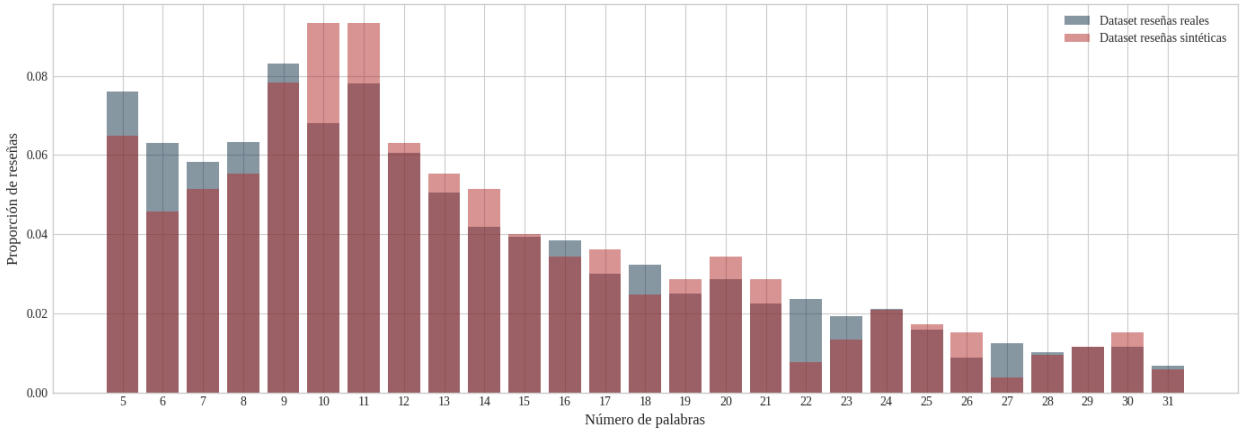


Nota. La figura muestra la frecuencia de los sentimientos según los *evidence spans* identificados en los conjuntos de datos sintético y final, permitiendo comparar su comportamiento y distribución.

Con el fin de analizar la estructura de los textos en ambos subconjuntos, la Figura 20 presenta la distribución del número de palabras por reseña para los conjuntos de datos real y sintético. Para ello, se calculó el número de palabras por reseña y se construyeron distribuciones de frecuencia sobre un rango común de longitudes comprendido entre 5 y 31 palabras, definido previamente como criterio para ambos conjuntos. Dado que los datasets difieren en tamaño, las frecuencias fueron normalizadas para obtener proporciones, permitiendo así una comparación directa.

Figura 20

Distribuciones de reseñas por número de palabras: Dataset real versus sintético



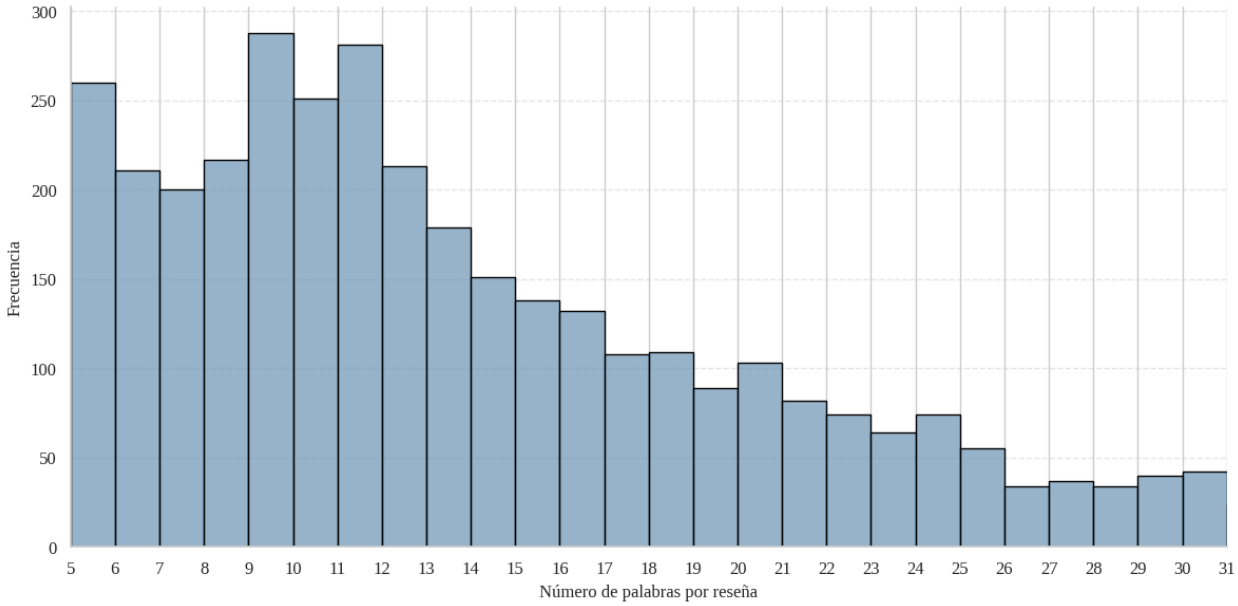
Nota. La figura presenta la distribución del número de palabras en las reseñas correspondientes a los conjuntos de datos real y sintético construidos.

A partir de los resultados, se observa un alto grado de solapamiento entre las distribuciones de longitud de las reseñas reales y sintéticas, lo que sugiere que el proceso de generación logró reproducir de manera adecuada la estructura textual del corpus original. En particular, ambas distribuciones concentran su mayor densidad en rangos similares de número de palabras, evidenciando que no existen desviaciones significativas hacia textos excesivamente cortos o largos en el conjunto sintético. No obstante, pueden identificarse ligeras variaciones en la frecuencia relativa de ciertas longitudes específicas, lo cual es esperable dado el carácter generado de los datos. En conjunto, estos resultados respaldan la consistencia del dataset sintético en términos de longitud, contribuyendo a su validez como complemento del corpus real.

4.1.2.2 Distribución de longitud de las reseñas. La Figura 21 presenta la distribución del número de palabras por reseña. Esta distribución permite observar la variabilidad en la extensión de los textos, evidenciando la coexistencia de opiniones cortas y descripciones más elaboradas, lo cual resulta relevante para el modelado de dependencias semánticas.

Figura 21

Distribución de reseñas por número de palabras

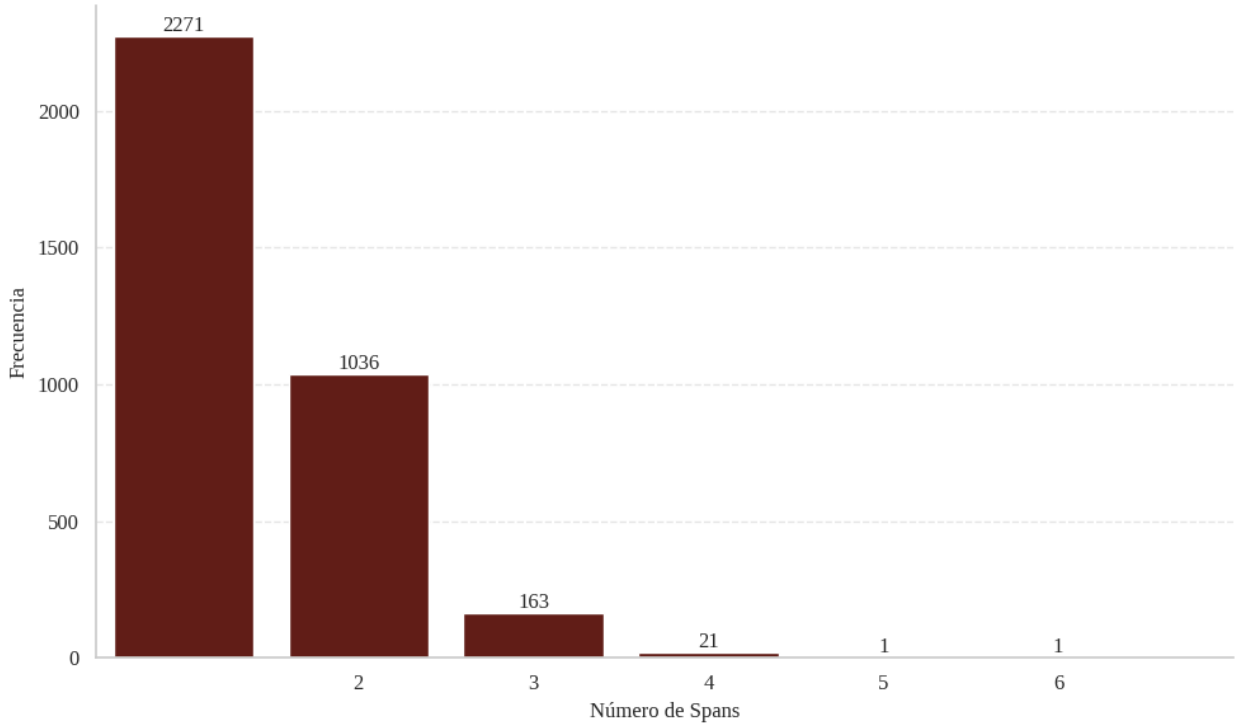


Nota. La figura presenta la distribución del corpus según la cantidad de palabras por reseña.

4.1.2.3 Distribución de spans por reseña. El corpus contiene un total de **4.927** etiquetas, lo que evidencia la presencia de múltiples anotaciones por reseña. En este sentido, **2.271** reseñas corresponden a instancias de una sola etiqueta (del inglés, *single-label*), mientras que **1.222** presentan múltiples etiquetas. La Figura 22 muestra la frecuencia del número de *evidence spans* anotados por reseña.

Figura 22

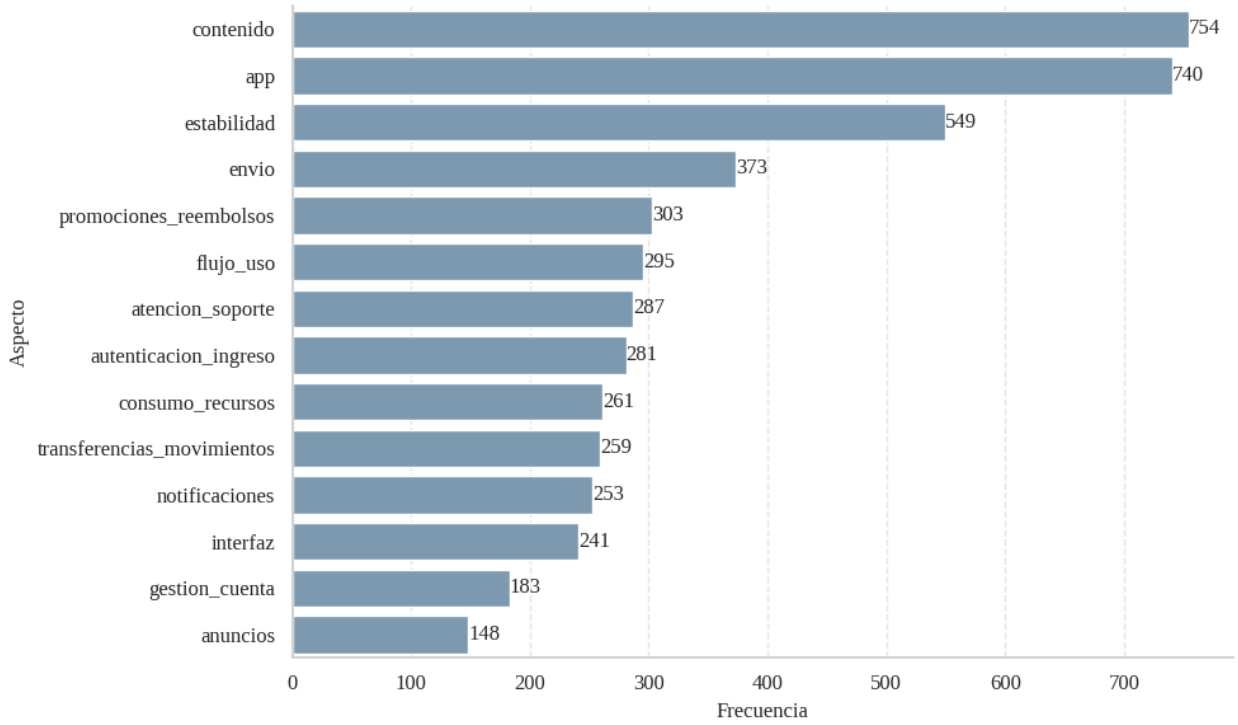
Frecuencia de la cantidad de fragmentos de evidencia por reseña



Nota. La figura presenta las frecuencias de la cantidad de *evidence spans* por reseña.

4.1.2.4 Distribución de aspectos. La Figura 23 muestra la frecuencia de los aspectos identificados en el corpus. Se observa que los aspectos más representativos son *contenido* (**754**), *app* (**740**) y *estabilidad* (**549**), lo que indica una alta concentración de opiniones relacionadas con la calidad del contenido y el comportamiento general de la aplicación. Otros aspectos relevantes incluyen *envio*, *promociones_reembolsos* y *autenticacion_ingreso*, reflejando la diversidad temática del corpus.

Figura 23
Frecuencia de aspectos en las etiquetas

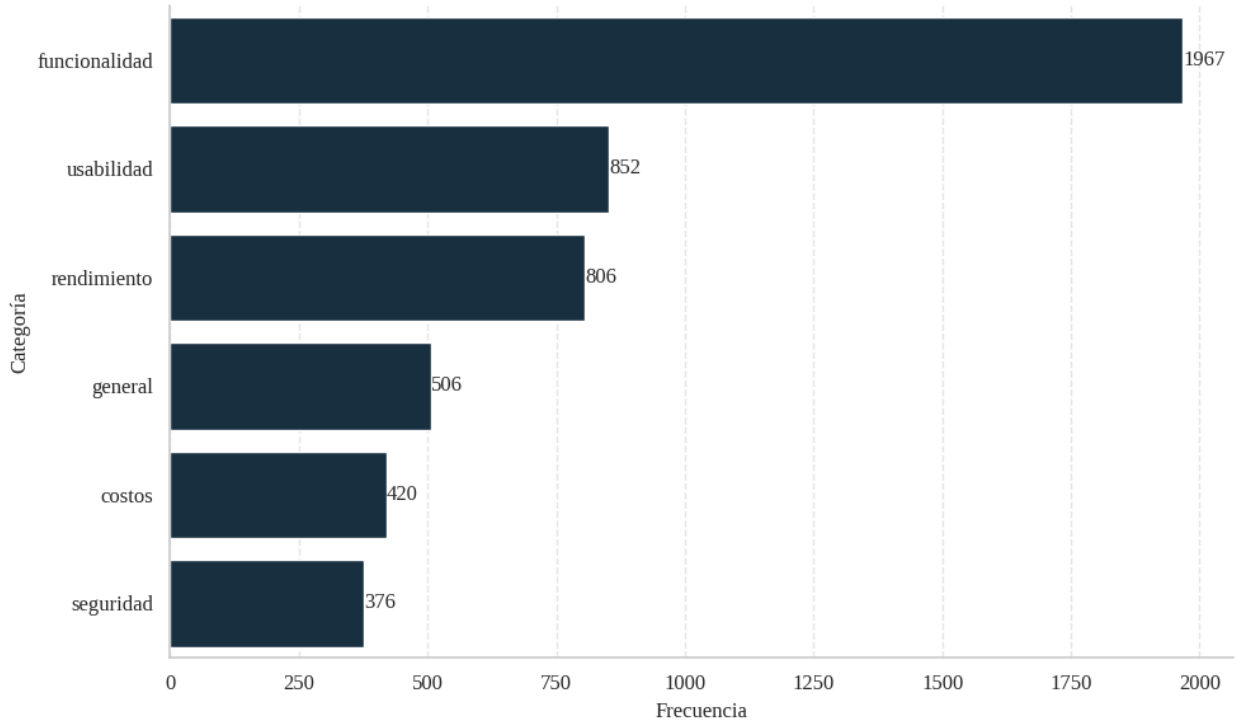


Nota. La figura presenta la frecuencia de aparición de los aspectos en las etiquetas del corpus.

4.1.2.5 Distribución de categorías. La Figura 24 presenta la distribución de las categorías definidas en el corpus. La categoría *funcionalidad* es la más frecuente (**1.967**), seguida de *usabilidad* (**852**) y *rendimiento* (**806**). Esto sugiere que la mayoría de las opiniones se centran en el funcionamiento general del sistema y la experiencia de uso, mientras que categorías como *costos* y *seguridad* presentan una menor, pero significativa, representación.

Figura 24

Frecuencia de categorías en las etiquetas

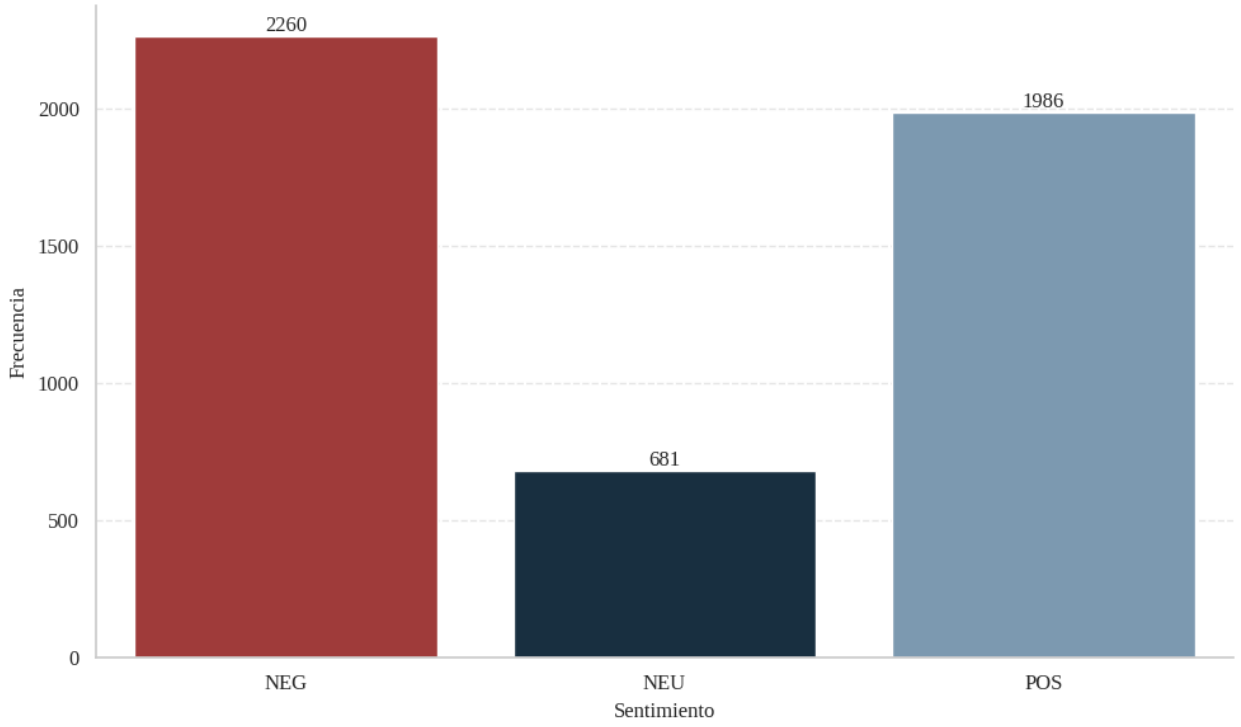


Nota. La figura presenta la frecuencia de aparición de las categorías en las etiquetas del corpus.

4.1.2.6 Distribución de sentimientos. La Figura 25 muestra la distribución de los sentimientos en el corpus. Se observa un predominio de instancias negativas (**2.260**), seguido de positivas (**1.986**) y, en menor medida, neutrales (**681**). Esta distribución sugiere una ligera inclinación hacia opiniones polarizadas, lo cual es consistente con el comportamiento típico de los usuarios en plataformas digitales.

Figura 25

Frecuencia de sentimientos en las etiquetas



Nota. La figura presenta la frecuencia de aparición de los sentimientos en las etiquetas del corpus.

4.1.2.7 Análisis del núcleo semántico. Con el fin de caracterizar la estructura semántica de las reseñas, se analizó la relación entre las etiquetas presentes en cada instancia, distinguiendo entre aquellas que comparten un mismo núcleo semántico y aquellas que presentan núcleos independientes, de acuerdo con la definición y tipología establecida (véase Sección 3.1.4.4).

Los resultados evidencian un claro predominio de reseñas con núcleos semánticos independientes, con un total de **3.309** instancias, frente a **184** reseñas que presentan núcleos semánticos compartidos. Esto sugiere que, en la mayoría de los casos, las opiniones expresadas por los usuarios tienden a segmentarse en unidades semánticas diferenciadas, aunque existe una proporción no despreciable de casos donde múltiples aspectos o categorías se articulan a partir de un mismo *evidence span*.

4.2 Métricas de evaluación

En esta sección, se presentan los resultados obtenidos por el modelo propuesto, siguiendo la estrategia de evaluación definida previamente (véase Sección 3.3). Dado el carácter del enfoque, los resultados se analizan de manera progresiva, comenzando por la capacidad de localización de evidencia, seguido del desempeño en las tareas de clasificación, y finalizando con la evaluación estricta de cuádruplas, que resume el comportamiento *end-to-end* del sistema.

Dado el esquema de validación cruzada utilizado, los resultados reportados corresponden al promedio de las métricas obtenidas en los cinco *folds*, junto con su desviación estándar.

4.2.1 Desempeño en la localización de evidencia

El primer componente evaluado corresponde a la capacidad del modelo para identificar correctamente los fragmentos de texto que sustentan una opinión. Para ello, se emplearon dos criterios: coincidencia aproximada mediante *Intersection over Union* (IoU) con umbral de 0.50, y coincidencia exacta de *spans* (véase Tabla 7).

Tabla 7
Desempeño en la localización de spans de evidencia

Métrica	Precisión	Recall	F1-score
Span IoU (≥ 0.50)	0.7932 ± 0.0172	0.7724 ± 0.0152	$\mathbf{0.7824 \pm 0.0070}$
Span exacto	0.5609 ± 0.0190	0.5460 ± 0.0112	$\mathbf{0.5532 \pm 0.0115}$

Nota. Resultados promedio sobre cinco particiones de validación cruzada. Se reporta la media y la desviación estándar para cada métrica.

Los resultados muestran que el modelo alcanza un desempeño sólido bajo el criterio de IoU, con un equilibrio adecuado entre precisión y *recall*. Esto indica que, en la mayoría de los casos, el modelo logra recuperar regiones textuales que se solapan significativamente con las anotaciones reales.

No obstante, al considerar el criterio de coincidencia exacta, se observa una disminución en el desempeño. Este comportamiento es consistente con la naturaleza de la tarea, ya que la delimitación precisa de los límites de un span resulta considerablemente más exigente que la identificación aproximada de la región relevante.

En conjunto, estos resultados evidencian que el modelo logra capturar de manera consistente la ubicación general de la evidencia en el texto, manteniendo un desempeño robusto entre particiones, aunque presenta margen de mejora en la precisión de los límites.

4.2.2 Desempeño en clasificación condicionada al fragmento de evidencia

Con el fin de aislar el comportamiento de las cabezas de clasificación, se evaluaron las predicciones de aspecto, categoría y sentimiento condicionadas a que el *evidence span* haya sido correctamente emparejado mediante IoU.

Tabla 8
Exactitud de clasificación condicionada al span

Subtarea	**Exactitud**
Aspecto	$0,7455 \pm 0,0143$
Categoría	$0,8649 \pm 0,0139$
Sentimiento	$0,9546 \pm 0,0094$
Categoría + Sentimiento	$0,8268 \pm 0,0149$

Nota. Métricas calculadas únicamente sobre *spans* correctamente emparejados mediante IoU. Los valores corresponden a la media y desviación estándar obtenidas en validación cruzada de 5 particiones.

Los resultados evidencian un desempeño diferenciado entre las distintas subtareas. En particular, la predicción de sentimiento presenta la mayor exactitud y, además, una baja variabilidad entre particiones, lo cual sugiere que, una vez identificado correctamente el contexto relevante, el modelo es altamente efectivo para determinar la polaridad de la opinión.

Por su parte, la predicción de categorías también alcanza un desempeño elevado, con una desviación estándar reducida, lo que indica que la incorporación de información contextual contribuye de manera estable a esta tarea. En contraste, la predicción de aspectos resulta más desafiante, reflejando la mayor granularidad y ambigüedad asociada a este nivel de representación semántica.

En conjunto, estos resultados permiten concluir que una parte importante del error global del sistema no proviene de la clasificación en sí misma, sino de la etapa de detección de la evidencia textual.

4.2.3 Análisis detallado por subtarea

Con el objetivo de profundizar en el comportamiento del modelo, se presentan los resultados desagregados por clase para cada una de las subtareas de clasificación: aspectos, categorías, sentimiento y la combinación categoría–sentimiento.

4.2.3.1 Predicción de aspectos.

Tabla 9
Resultados por clase en la predicción de aspectos

Aspecto	Precisión	Recall	F1
contenido	$0,5175 \pm 0,0206$	$0,5686 \pm 0,0196$	$\mathbf{0,5412 \pm 0,0071}$
app	$0,5653 \pm 0,0700$	$0,5914 \pm 0,0315$	$\mathbf{0,5738 \pm 0,0323}$
estabilidad	$0,5608 \pm 0,0372$	$0,6347 \pm 0,0437$	$\mathbf{0,5949 \pm 0,0355}$
envio	$0,6265 \pm 0,0914$	$0,4026 \pm 0,1183$	$\mathbf{0,4722 \pm 0,0600}$
promociones_reembolsos	$0,7695 \pm 0,0617$	$0,6136 \pm 0,0940$	$\mathbf{0,6744 \pm 0,0410}$
flujo_uso	$0,6223 \pm 0,0700$	$0,5486 \pm 0,0431$	$\mathbf{0,5812 \pm 0,0427}$
atencion_soporte	$0,5727 \pm 0,0577$	$0,5320 \pm 0,0623$	$\mathbf{0,5477 \pm 0,0404}$
autenticacion_ingreso	$0,4494 \pm 0,0471$	$0,6290 \pm 0,0962$	$\mathbf{0,5228 \pm 0,0596}$
consumo_recursos	$0,8239 \pm 0,0465$	$0,7325 \pm 0,0377$	$\mathbf{0,7749 \pm 0,0345}$
transferencias_movimientos	$0,7271 \pm 0,1289$	$0,2748 \pm 0,1165$	$\mathbf{0,3819 \pm 0,1186}$

Continúa en la siguiente página

Tabla 9 – *continuación*

Aspecto	Precisión	Recall	F1
notificaciones	$0,7403 \pm 0,0662$	$0,7221 \pm 0,0333$	$\mathbf{0,7294 \pm 0,0389}$
interfaz	$0,6616 \pm 0,0598$	$0,7142 \pm 0,0257$	$\mathbf{0,6859 \pm 0,0393}$
gestion_cuenta	$0,4820 \pm 0,0899$	$0,3704 \pm 0,0786$	$\mathbf{0,4114 \pm 0,0521}$
anuncios	$0,7262 \pm 0,0968$	$0,7174 \pm 0,0573$	$\mathbf{0,7203 \pm 0,0723}$

Nota. Métricas de inferencia de aspecto calculadas únicamente sobre *evidence spans* correctamente emparejados mediante IoU. Los valores corresponden a la media y desviación estándar obtenidas en validación cruzada de cinco *folds*.

El análisis por clase revela una variabilidad considerable en el desempeño entre distintos aspectos. Algunas clases presentan métricas equilibradas de precisión y *recall*, lo que indica una adecuada capacidad de generalización. Sin embargo, otras clases muestran un desempeño limitado, particularmente en términos de *recall*, lo que sugiere dificultades del modelo para identificar todas las instancias relevantes.

Este comportamiento puede atribuirse tanto al desbalance en la distribución de clases como a la complejidad semántica de ciertos aspectos, que pueden manifestarse de manera más diversa en el texto (véase Sección 4.1.2.4).

4.2.3.2 Predicción de categorías.

Tabla 10
Resultados por clase en la predicción de categorías

Categoría	Precisión	Recall	F1
funcionalidad	0,7085 ± 0,0136	0,6998 ± 0,0406	**0,7033 ± 0,0182**
usabilidad	0,7035 ± 0,0417	0,6770 ± 0,0162	**0,6895 ± 0,0249**
rendimiento	0,6869 ± 0,0234	0,6837 ± 0,0346	**0,6846 ± 0,0192**
general	0,6265 ± 0,0796	0,5751 ± 0,0429	**0,5951 ± 0,0387**
costos	0,6831 ± 0,0811	0,6252 ± 0,0640	**0,6470 ± 0,0354**

Continúa en la siguiente página

Tabla 10 – *continuación*

Categoría	Precisión	Recall	F1
seguridad	0,6453 ± 0,0628	0,6148 ± 0,0527	**0,6253 ± 0,0206**

Nota. Métricas de inferencia de categoría calculadas únicamente sobre *evidence spans* correctamente emparejados mediante IoU. Los valores corresponden a la media y desviación estándar obtenidas en validación cruzada de cinco *folds*

El desempeño en la predicción de categorías resulta más homogéneo en comparación con la tarea de aspectos, con valores de F1 relativamente consistentes entre clases. Esto sugiere que el modelo logra capturar de manera adecuada las distinciones a nivel categórico, favorecido por una menor granularidad semántica.

No obstante, ciertas categorías presentan una ligera caída en *recall*, lo que indica que aún existen casos en los que el modelo no logra capturar completamente la diversidad de expresiones asociadas a dichas clases.

4.2.3.3 Predicción de sentimiento.

Tabla 11
Resultados por clase en la predicción de sentimiento

Sentimiento	Precisión	Recall	F1
NEG	0,7361 ± 0,0355	0,7140 ± 0,0114	**0,7246 ± 0,0214**
POS	0,7494 ± 0,0236	0,7483 ± 0,0222	**0,7482 ± 0,0079**
NEU	0,8554 ± 0,0364	0,7823 ± 0,0394	**0,8163 ± 0,0272**

Nota. Métricas de inferencia de sentimiento calculadas únicamente sobre *evidence spans* correctamente emparejados mediante IoU. Los valores corresponden a la media y desviación estándar obtenidas en validación cruzada de cinco *folds*.

La clasificación de sentimiento constituye la subtarea con mejor desempeño global, evidenciando métricas altas y consistentes entre clases. En particular, se observa un equilibrio sólido entre precisión y *recall* en las clases mayoritarias, así como un desempeño robusto en la clase neutra.

Asimismo, la baja desviación estándar indica un comportamiento estable del modelo a través de los distintos *folds*, lo que refuerza la confiabilidad de estos resultados. En conjunto, esto sugiere que el modelo logra capturar de manera efectiva las señales lingüísticas asociadas a la polaridad.

4.2.3.4 Predicción conjunta categoría–sentimiento.

Tabla 12
Resultados por clase en la predicción conjunta categoría–sentimiento

Categoría + Sentimiento	Precisión	Recall	F1
funcionalidad | negativo	0,6591 ± 0,0458	0,6570 ± 0,0540	**0,6561 ± 0,0347**
funcionalidad | positivo	0,7081 ± 0,0359	0,6808 ± 0,0235	**0,6931 ± 0,0121**
rendimiento | negativo	0,6610 ± 0,0144	0,6609 ± 0,0680	**0,6598 ± 0,0412**
usabilidad | negativo	0,6632 ± 0,0667	0,6148 ± 0,0348	**0,6374 ± 0,0480**
general | positivo	0,6210 ± 0,0894	0,6318 ± 0,0510	**0,6216 ± 0,0555**
usabilidad | positivo	0,7067 ± 0,0640	0,7058 ± 0,0182	**0,7045 ± 0,0323**
funcionalidad | neutro	0,7098 ± 0,0961	0,7145 ± 0,0430	**0,7087 ± 0,0529**
rendimiento | positivo	0,6234 ± 0,0768	0,6329 ± 0,0383	**0,6248 ± 0,0433**
costos | positivo	0,6777 ± 0,1065	0,5977 ± 0,0730	**0,6286 ± 0,0651**
seguridad | positivo	0,6466 ± 0,0880	0,7018 ± 0,0889	**0,6728 ± 0,0874**
seguridad | negativo	0,5676 ± 0,1118	0,4843 ± 0,0703	**0,5128 ± 0,0545**
usabilidad | neutro	0,6686 ± 0,1085	0,6479 ± 0,0725	**0,6540 ± 0,0769**
costos | negativo	0,6032 ± 0,0842	0,6465 ± 0,0993	**0,6148 ± 0,0518**
general | negativo	0,4963 ± 0,1052	0,3598 ± 0,0808	**0,4138 ± 0,0837**
costos | neutro	0,5562 ± 0,2553	0,3592 ± 0,1624	**0,4290 ± 0,1875**
rendimiento | neutro	0,7527 ± 0,0998	0,6643 ± 0,1429	**0,6964 ± 0,0951**
general | neutro	0,5648 ± 0,3080	0,3990 ± 0,2200	**0,4516 ± 0,2288**
seguridad | neutro	0,4556 ± 0,0648	0,3860 ± 0,1808	**0,4011 ± 0,1248**

Nota. Métricas de inferencia conjunta de categoría-sentimiento calculadas únicamente sobre *evidence spans* correctamente emparejados mediante IoU. Los valores corresponden a la media y desviación estándar obtenidas en validación cruzada de cinco *folds*.

La predicción conjunta de categoría y sentimiento introduce un mayor nivel de complejidad al requerir la correcta combinación de ambas dimensiones semánticas, lo que se traduce en una disminución del desempeño respecto a las tareas individuales.

Aun así, el modelo mantiene resultados consistentes en las combinaciones más frecuentes, evidenciando su capacidad para capturar dicha relación. En contraste, las combinaciones menos representadas presentan una mayor variabilidad, reflejada en desviaciones estándar más elevadas, lo cual es consistente con la menor cantidad de ejemplos disponibles.

4.2.4 Evaluación estricta de cuádruplas

La evaluación estricta de cuádruplas constituye la métrica principal del modelo, al requerir la correcta identificación simultánea del span de evidencia, el aspecto, la categoría y el sentimiento.

Tabla 13
Evaluación estricta de cuádruplas

Métrica	**Precisión**	**Recall**	**F1-score**
Cuádruplas (IoU ≥ 0.50)	0,5287 ± 0,0156	0,5149 ± 0,0176	**0,5216 ± 0,0132**

Nota. Evaluación estricta considerando coincidencia simultánea de *evidence span*, aspecto, categoría y sentimiento.

El desempeño obtenido bajo este criterio refleja la dificultad inherente a la tarea *end-to-end*, donde los errores en cualquiera de los componentes afectan directamente la predicción final. En comparación con los resultados obtenidos en las tareas individuales de localización y clasificación, se observa una disminución en el desempeño, lo cual es esperable dado el carácter acumulativo del error en este escenario.

En conjunto, estos resultados evidencian que, aunque el modelo logra integrar de manera coherente las distintas etapas del *pipeline*, la tarea completa continúa representando un desafío significativo.

4.3 Prototipo web para la visualización de resultados del modelo

En esta sección se presentan los resultados de la implementación del prototipo web planteado (véase Sección 3.4). Para visualizar detalladamente el código fuente e instrucciones de uso, véase Apéndice C.

Con el fin de ejemplificar el funcionamiento, se llevó a cabo el proceso completo utilizando tres reseñas de prueba totalmente nuevas para el modelo, ingresadas a través de la opción de escritura en texto plano dentro de la interfaz. Las reseñas empleadas fueron: *"odio que la app me drene la bateria del cel"*, *"me gustan mucho los envios pero los precios son muy altos"* y *"podrian por favor cambiar el color de los botones"*. Este ejercicio permitió ilustrar el comportamiento del sistema frente a distintos tipos de opiniones y estructuras lingüísticas.

4.3.1 Carga de datos

La aplicación web desarrollada inicia con un panel de carga de datos (véase Figura 26), el cual permite dos modalidades de ingreso de información. La primera consiste en la carga de un archivo en formato CSV que debe contener una columna denominada *text*, en la que se incluyan las reseñas a procesar. Este archivo debe tener un tamaño máximo de 200 MB. La segunda modalidad corresponde a la entrada manual de reseñas en texto plano.

Figura 26
Módulo de carga de datos

Nota. La figura presenta el módulo de carga de datos del prototipo web, con tres ejemplos de reseñas tomadas del conjunto de datos inicial.

4.3.2 Inferencia

Una vez cargados los datos, la aplicación establece comunicación con el modelo para ejecutar el proceso de inferencia. El tiempo requerido para esta operación depende directamente del volumen de datos ingresados.

4.3.3 Visualización

Los resultados obtenidos se presentan mediante un *dashboard* interactivo, organizado en las siguientes secciones:

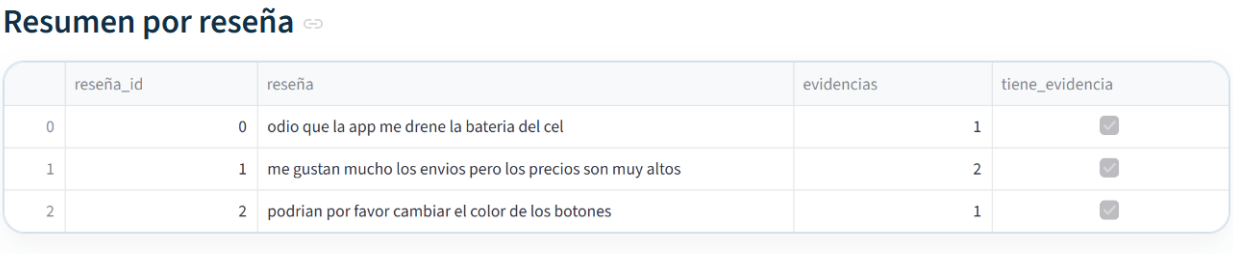
4.3.3.1 Resumen global. Esta sección presenta un conjunto de indicadores clave a través de tarjetas informativas, tales como el número total de reseñas cargadas, la cantidad de reseñas en las que se detectaron evidencias, el número total de spans de evidencia identificados, y la cobertura (definida como el porcentaje de reseñas que contienen al menos un span de evidencia respecto al total). Asimismo, se incluye el promedio de spans de evidencia por reseña. Adicionalmente, se incorporan gráficos interactivos que muestran la frecuencia de aparición de sentimientos, categorías y aspectos (véase Figura 27).

Figura 27
Tablero de resumen global

Nota. La figura presenta el resumen global luego de procesar las reseñas ingresadas a través del módulo de carga de datos del sitio web.

4.3.3.2 Resumen por reseña. En esta sección se listan las reseñas de manera individual, junto con su identificador correspondiente y el número de spans de evidencia detectados en cada una (véase Figura 28).

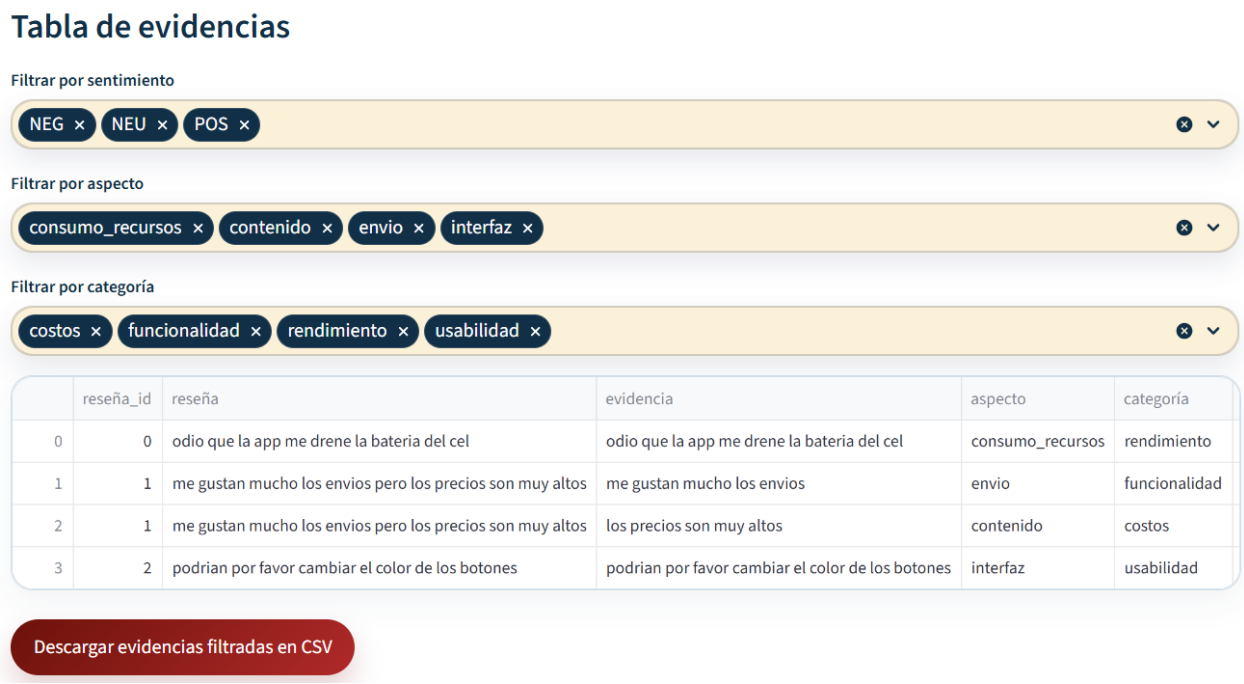
Figura 28
Cuadro resumen por reseña



Nota. La figura presenta el resumen por reseña luego de procesar las reseñas ingresadas a través del módulo de carga de datos del sitio web.

4.3.3.3 Tabla de evidencias. Se presenta una tabla que consolida los resultados de la inferencia realizada por el modelo, incluyendo el identificador de la reseña y los spans de evidencia detectados. En los casos en que una reseña contenga múltiples spans o spans con núcleo semántico compartido, se genera una fila independiente por cada uno de ellos. Esta sección incorpora filtros por aspecto, categoría y sentimiento. Además, permite la descarga de los resultados en formato CSV (véase Figura 29).

Figura 29
Tabla de evidencias



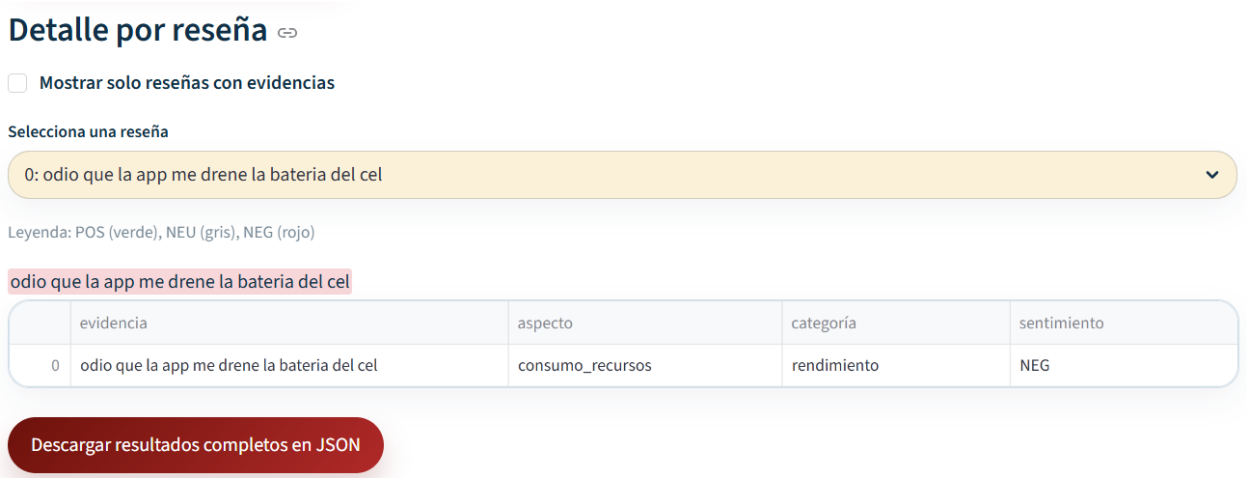
	reseña_id	reseña	evidencia	aspecto	categoría
0	0	odio que la app me drene la bateria del cel	odio que la app me drene la bateria del cel	consumo_recursos	rendimiento
1	1	me gustan mucho los envios pero los precios son muy altos	me gustan mucho los envios	envio	funcionalidad
2	1	me gustan mucho los envios pero los precios son muy altos	los precios son muy altos	contenido	costos
3	2	podrian por favor cambiar el color de los botones	podrian por favor cambiar el color de los botones	interfaz	usabilidad

Nota. La figura presenta la tabla de evidencias luego de procesar las reseñas ingresadas a través del módulo de carga de datos del sitio web.

4.3.3.4 Detalle por reseña. Esta sección permite seleccionar una reseña específica, independientemente de si contiene evidencias detectadas o no, para visualizar el detalle completo de la inferencia. El texto de la reseña se presenta con un sistema de subrayado codificado por color: verde para evidencias positivas, rojo para negativas y gris para neutras. Asimismo, se incluye una tabla con el detalle de cada span de evidencia identificado. Finalmente, se ofrece la opción de descargar los resultados completos en formato JSON (véanse Figuras 30, 31 y 32).

Figura 30

Detalle de la reseña 1



Nota. La figura presenta el detalle de la reseña 1 luego de procesar las reseñas ingresadas a través del módulo de carga de datos del sitio web.

Figura 31

Detalle de la reseña 2

Nota. La figura presenta el detalle de la reseña 2 luego de procesar las reseñas ingresadas a través del módulo de carga de datos del sitio web.

Figura 32
Detalle de la reseña 3

Nota. La figura presenta el detalle de la reseña 3 luego de procesar las reseñas ingresadas a través del módulo de carga de datos del sitio web.

5. Conclusiones

En el presente trabajo se propuso un enfoque integral para el análisis de reseñas en español colombiano, abordando desde la construcción del conjunto de datos hasta el desarrollo de un modelo basado en arquitecturas *Transformer* y la implementación de un prototipo funcional para la visualización de resultados.

En primer lugar, se diseñó un esquema de etiquetado propio, el cual difiere de las aproximaciones tradicionales reportadas en la literatura. Este esquema permitió adaptarse de manera más adecuada a la naturaleza específica de las reseñas analizadas, incorporando simultáneamente información de categoría, aspecto y sentimiento. Este enfoque de etiquetado multidimensional facilitó una representación más rica y estructurada del contenido textual.

Adicionalmente, se construyó un conjunto de datos en español colombiano compuesto por reseñas provenientes de diferentes aplicaciones y dominios. Este recurso constituye un aporte relevante, ya que ofrece un punto de partida para futuras investigaciones en procesamiento de lenguaje natural en español, particularmente en contextos latinoamericanos, donde este tipo de

recursos suele ser limitado. De manera complementaria, se realizó un proceso de aumento de datos mediante la generación supervisada de datos sintéticos utilizando *ChatGPT-5.0*, lo que permitió obtener un conjunto de datos más balanceado y robusto. Este proceso resultó satisfactorio, ya que se logró que el corpus sintético presentara estructuras semánticas altamente similares a las del corpus basado en datos reales, manteniendo coherencia contextual y preservando las características lingüísticas propias del dominio analizado.

En cuanto al modelo, se implementó una arquitectura basada en BETO, la cual demostró ser capaz de identificar de manera satisfactoria fragmentos de evidencia textual asociados a categorías, aspectos y sentimientos. La utilización de mecanismos de autoatención propios de la arquitectura *Transformer* permitió capturar dependencias contextuales relevantes dentro de las reseñas, mejorando la capacidad del modelo para interpretar relaciones semánticas complejas.

Asimismo, se abordó el tratamiento de reseñas con distinta naturaleza semántica, introduciendo los conceptos de *núcleo semántico compartido* y *núcleos semánticos independientes*. Estos conceptos permitieron analizar cómo distintas expresiones pueden converger o divergir en su significado subyacente, facilitando una mejor comprensión de la estructura interna de las reseñas y su correcta clasificación dentro del esquema propuesto.

Por otra parte, se desarrolló un prototipo web funcional que permite la interacción del usuario con los resultados del modelo de inferencia de manera intuitiva y amigable. Esta herramienta facilita la carga de datos, la ejecución del modelo y la visualización de resultados, contribuyendo a la interpretabilidad del sistema y a su potencial uso como herramienta de análisis exploratorio.

Finalmente, se destaca que el trabajo integra de manera coherente el ciclo completo de un sistema de análisis de texto basado en inteligencia artificial, desde la construcción del dataset hasta la implementación del modelo y su despliegue en un entorno interactivo, lo que demuestra la viabilidad de la solución propuesta y su potencial para ser extendida en futuras investigaciones.

6. Recomendaciones

Como líneas de trabajo futuro, se identifican diversas oportunidades de mejora y expansión del presente proyecto, tanto a nivel de los datos, el modelo propuesto y la plataforma web desarrollada.

En primer lugar, se propone ampliar la generalización del conjunto de datos, incorporando reseñas provenientes de nuevas fuentes y dominios distintos a los considerados en este estudio. Esto permitiría que el modelo sea capaz de reconocer categorías y aspectos fuera de los presentes en las aplicaciones seleccionadas, aumentando así su capacidad de generalización y robustez frente a escenarios reales más diversos. Adicionalmente, sería pertinente incluir reseñas con diferentes naturalezas semánticas, lo cual contribuiría a fortalecer la identificación de estructuras discursivas más complejas y de núcleos semánticos compartidos entre distintas expresiones, mejorando su correcta clasificación.

En esta misma línea, se plantea la exploración de modelos basados en LLMs en diferentes etapas del proceso. Estos podrían ser utilizados no solo en la fase de clasificación, sino también en el preprocesamiento de datos, la normalización de texto, la generación de representaciones semánticas más ricas, e incluso en la anotación automática de datos para futuras iteraciones del modelo. El uso de LLMs también abre la posibilidad de implementar estrategias de aprendizaje semisupervisado o débilmente supervisado, reduciendo la dependencia de datos etiquetados manualmente.

De manera complementaria, se propone aprovechar estas capacidades de los LLMs para la generación de datos sintéticos a partir de expresiones propias de la jerga colombiana. Este enfoque permitiría enriquecer el conjunto de datos con reseñas artificiales pero lingüísticamente plausibles, incrementando la capacidad de generalización del modelo frente a variaciones lingüísticas específicas del contexto local.

Otro aspecto relevante consiste en la mejora de la arquitectura del modelo, incorporando enfoques más avanzados de representación contextual, tales como mecanismos de atención más

sofisticados o modelos basados en transformadores optimizados para tareas de análisis de sentimiento y clasificación multietiqueta. Asimismo, sería interesante evaluar el desempeño del sistema bajo esquemas de entrenamiento incremental, que permitan su adaptación continua a nuevos tipos de datos.

Finalmente, en cuanto a la plataforma web desarrollada, se plantea como trabajo futuro su evolución hacia un sistema más robusto y escalable, orientado a su posible despliegue en entornos productivos.

Referencias Bibliográficas

Baishya, D., y Baruah, R. (2022). Recent Trends in Deep Learning for Natural Language Processing and Scope for Asian Languages. *2022 International Conference on Augmented Intelligence and Sustainable Systems (ICAISS)*, 408-411. https://api.semanticscholar.org/CorpusID:255996941

Bose, R., Dey, R., Roy, S., y Sarddar, D. (2019). Sentiment Analysis on Online Product Reviews. https://doi.org/10.1007/978-981-13-7166-0_56

Branch. (2025). Situación digital de Colombia en el 2025 [Publicado el 26 de mayo de 2025. Consultado el 20 de abril de 2026]. https://branch.com.co/marketing-digital/situacion-digital-de-colombia-en-el-2025/

Cañete, A., et al. (2020). BETO: Spanish BERT. *ArXiv*.

Chris, E., Favor, B., y Ramon, T. (2024). Aspect-Based Sentiment Analysis of Customer Reviews in E-Commerce.

de la Torre, J. (2025). Transformadores: Fundamentos teoricos y Aplicaciones. https://arxiv.org/abs/2302.09327

Devlin, J., Chang, M.-W., Lee, K., y Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. https://arxiv.org/abs/1810.04805

Feldman, R., y Sanger, J. (2006). *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press.

Feng, S. Y., Gangal, V., Wei, J., Chandar, S., Vosoughi, S., Mitamura, T., y Hovy, E. (2021). A Survey of Data Augmentation Approaches for NLP. https://arxiv.org/abs/2105.03075

Goodfellow, I., Bengio, Y., y Courville, A. (2016). *Deep Learning*. MIT Press.

Goyal, M., y Mahmoud, Q. (2024). A Systematic Review of Synthetic Data Generation Techniques Using Generative AI. *Electronics* 13, 3509. https://doi.org/10.3390/electronics13173509

Gunathilaka, S., y De Silva, N. (2022). Aspect-based Sentiment Analysis on Mobile Application Reviews. *2022 22nd International Conference on Advances in ICT for Emerging Regions (ICTer)*, 183-188. https://doi.org/10.1109/ICTer58063.2022.10024070

Guo, C., Pleiss, G., Sun, Y., y Weinberger, K. Q. (2017). On Calibration of Modern Neural Networks. https://arxiv.org/abs/1706.04599

Guzman, E., y Maalej, W. (2014). How Do Users Like This Feature? A Fine Grained Sentiment Analysis of App Reviews. *2014 IEEE 22nd International Requirements Engineering Conference, RE 2014 - Proceedings*, 153-162. https://doi.org/10.1109/RE.2014.6912257

Hassenzahl, M., y Tractinsky, N. (2006). User experience - A research agenda. *Behaviour and Information Technology* 25, 91-97. https://doi.org/10.1080/01449290500330331

He, R., Lee, W. S., Ng, H. T., y Dahlmeier, D. (2019). An Interactive Multi-task Learning Network for End-to-End Aspect-Based Sentiment Analysis. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 504-515. https://doi.org/10.18653/v1/P19-1048

Hellwig, N. C., Fehle, J., y Wolff, C. (2024). Exploring large language models for the generation of synthetic training samples for aspect-based sentiment analysis in low resource settings. *Expert Systems with Applications* 261, 125514. https://doi.org/10.1016/j.eswa.2024.125514

Hoang, M., Bihorac, O. A., y Rouces, J. (2019). Aspect-Based Sentiment Analysis using BERT. En M. Hartmann y B. Plank (Eds.), *Proceedings of the 22nd Nordic Conference on Computational Linguistics* (pp. 187-196). Linköping University Electronic Press. https://aclanthology.org/W19-6120/

Hua, Y. C., Denny, P., Wicker, J., y Taskova, K. (2024). A systematic review of aspect-based sentiment analysis: domains, methods, and trends. *Artificial Intelligence Review* 57. https://doi.org/10.1007/s10462-024-10906-z

ISO. (2019). ISO 9241-210: Ergonomics of Human-System Interaction—Human-Centred Design for Interactive Systems [ISO Standard].

ISO. (2024). *Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Quality model overview and usage* (Standard). International Organization for Standardization. Geneva, CH. https://iso.org

Jurafsky, D., y Martin, J. H. (2026). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, with Language Models* (3rd) [Online manuscript released January 6, 2026]. https://web.stanford.edu/~jurafsky/slp3/

LeCun, Y., y Hinton, G. (2015). Deep Learning. *Nature* 521, 436-44. https://doi.org/10.1038/nature14539

Li, X., Bing, L., Li, P., y Lam, W. (2019a). A Unified Model for Opinion Target Extraction and Target Sentiment Prediction. https://arxiv.org/abs/1811.05082

Li, X., Bing, L., Zhang, W., y Lam, W. (2019b). Exploiting BERT for End-to-End Aspect-based Sentiment Analysis. https://arxiv.org/abs/1910.00883

Lin, T., Wang, Y., Liu, X., y Qiu, X. (2022). A survey of transformers. *AI Open* 3, 111-132. https://doi.org/https://doi.org/10.1016/j.aiopen.2022.10.001

Liu, B. (2012). Sentiment Analysis and Opinion Mining. *Synthesis Lectures on Human Language Technologies* 5, 1-167. https://doi.org/10.2200/s00416ed1v01y201204hlt016

Liu, J., Tian, Y., y Song, Y. (2025). *Balanced Training Data Augmentation for Aspect-Based Sentiment Analysis*. https://arxiv.org/abs/2507.09485

Mosin, V., Samenko, I., Kozlovskii, B., Tikhonov, A., y Yamshchikov, I. P. (2023). Fine-tuning transformers: Vocabulary transfer. *Artificial Intelligence* 317, 103860. https://doi.org/https://doi.org/10.1016/j.artint.2023.103860

Nadăş, M., Dioşan, L., y Tomescu, A. (2025). Synthetic Data Generation Using Large Language Models: Advances in Text and Code. *IEEE Access* 13, 134615-134633. https://doi.org/10.1109/access.2025.3589503

Noronha, R., Alenchery, A., Deepa S, D., Jayapriya, J., y .M, V. (2025). Revolutionizing Legal Intelligence: Advances in Neural Networks and Language Mwodels for Legal NLP, 608-616. https://doi.org/10.1109/ISACC65211.2025.10969245

Peng, H., Ma, Y., Li, Y., y Cambria, E. (2020). Knowing What, How and Why: A Near Complete Solution for Aspect-Based Sentiment Analysis. *Proceedings of the AAAI Conference on Artificial Intelligence* 34, 8600-8607. https://doi.org/10.1609/aaai.v34i05.6406

Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I., y Manandhar, S. (2014). SemEval-2014 Task 4: Aspect Based Sentiment Analysis. En P. Nakov y T. Zesch (Eds.), *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)* (pp. 27-35). Association for Computational Linguistics. https://doi.org/10.3115/v1/S14-2004

Puspita, S., Ardhani, A., Retnaningrum, D., Firmansyah, A., y Rolliawati, D. (2024). ANALYSIS OF SOFTWARE QUALITY USING THE FURPS+ MODEL. *JURTEKSI (Jurnal Teknologi dan Sistem Informasi)* 11, 131-138. https://doi.org/10.33330/jurteksi.v11i1.3233

Ruder, S., Peters, M. E., Swayamdipta, S., y Wolf, T. (2019). Transfer Learning in Natural Language Processing. En A. Sarkar y M. Strube (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials* (pp. 15-18). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-5004

Schouten, K., y Frasincar, F. (2015). Survey on Aspect-Level Sentiment Analysis. *IEEE Transactions on Knowledge and Data Engineering* 28, 1-1. https://doi.org/10.1109/TKDE.2015.2485209

Sennrich, R., Haddow, B., y Birch, A. (2016). Improving Neural Machine Translation Models with Monolingual Data. https://arxiv.org/abs/1511.06709

Šmíd, J., y Král, P. (2025). Cross-lingual aspect-based sentiment analysis: A survey on tasks, approaches, and challenges. *Information Fusion* 120, 103073. https://doi.org/https://doi.org/10.1016/j.inffus.2025.103073

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., y Polosukhin, I. (2017). Attention is all you need. *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17, 6000-6010.

Wang, Y., Huang, M., Zhao, L., y Zhu, X. (2016). Attention-based LSTM for Aspect-level Sentiment Classification. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 606-615. https://doi.org/10.18653/v1/D16-1058

Wei, J., y Zou, K. (2019). EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. https://arxiv.org/abs/1901.11196

Xu, H., Liu, B., Shu, L., y Yu, P. S. (2018). Double Embeddings and CNN-based Sequence Labeling for Aspect Extraction. https://arxiv.org/abs/1805.04601

Xu, L., Chia, Y. K., y Bing, L. (2021). Learning Span-Level Interactions for Aspect Sentiment Triplet Extraction. En C. Zong, F. Xia, W. Li y R. Navigli (Eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 4755-4766). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.367

Xu, L., Li, H., Lu, W., y Bing, L. (2020). Position-Aware Tagging for Aspect Sentiment Triplet Extraction. *Proceedings of EMNLP 2020*, 2339-2349. https://doi.org/10.18653/v1/2020.emnlp-main.183

Young, T., Hazarika, D., Poria, S., y Cambria, E. (2018). Recent Trends in Deep Learning Based Natural Language Processing. https://arxiv.org/abs/1708.02709

Zhang, L., Wang, S., y Liu, B. (2018). Deep Learning for Sentiment Analysis : A Survey. https://arxiv.org/abs/1801.07883

Zhong, Q., Li, H., Zhuang, L., Liu, J., y Du, B. (2024). Iterative Data Generation with Large Language Models for Aspect-based Sentiment Analysis. https://arxiv.org/abs/2407.00341

Zhou, Y., Huang, L., Guo, T., Han, J., y Hu, S. (2019). A Span-based Joint Model for Opinion Target Extraction and Target Sentiment Classification. *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, 5485-5491. https://doi.org/10.24963/ijcai.2019/762

Ziv, I., Unger, M., y Geva, H. (2025). The Impact of LLM-Generated Reviews on Recommender Systems: Textual Shifts, Performance Effects, and Strategic Platform Control. https://arxiv.org/abs/2601.02362

Apéndices

Apéndice A. Prompt maestro empleado para la generación de datos sintéticos

Figura A1. *Prompt maestro estructurado para generación de reseñas sintéticas con evidence spans para ABSA*

Objetivo de la tarea:
Genera exactamente 10 reseñas de aplicaciones móviles en español colombiano que puedan ser usadas para entrenar un modelo de análisis de sentimientos por aspectos (Aspect-Based Sentiment Analysis, ABSA). Cada reseña debe reflejar fielmente las combinaciones de aspecto, categoría y sentimiento definidas en las etiquetas de entrada.

Formato de entrada:
El modelo recibe como entrada una reseña y debe identificar tripletas semánticas compuestas por un **aspecto**, una **categoría** y una **polaridad**.

- `aspect`: elemento específico mencionado en la reseña (ver Tabla 3).
- `evidence_span`: índices de palabra inicial y final que delimita el fragmento mínimo del texto que justifica la asignación del aspecto, la categoría y el sentimiento.
- `category`: dimensión a la que pertenece el aspecto (ver Tabla 2).
- `sentiment`: orientación del sentimiento expresado (ver Tabla 4).

Indicaciones de anotación de *evidence spans*:

- Cada span debe ser el fragmento contiguo más corto que justifique la(s) tripleta(s) semántica(s).
- Una reseña puede contener múltiples spans.
- Algunos spans pueden estar asociados a múltiples aspectos y categorías (núcleo semántico compartido).

Formato de salida:
Cada predicción debe seguir estrictamente el siguiente formato JSON:

```
{
  "text": "contenido de la reseña",
  "label": [
    {
      "aspect": ["aspect"],
      "evidence_span": [start_index, end_index],
      "sentiment": "POS|NEG|NEU",
      "category": "categoria"
    },
    ...
  ]
}
```

Notas: - Devuelve solo la predicción, sin comentarios adicionales. - Asegúrate de que los índices de spans coincidan con la posición de palabras reales en la reseña generada.

Ejemplos anotados:

```
{
  "text": "me parece una buena aplicacion pero no me gusta la estetica",
  "label": [
    {
      "aspect": ["app"],
      "evidence_span": [0, 4],
      "sentiment": "POS",
      "category": "general"
    },
    {
      "aspect": ["interfaz"],
      "evidence_span": [6, 10],
      "sentiment": "NEG",
      "category": "usabilidad"
    }
  ]
},

{
  "text": "me encantan los precios y los productos pero la interfaz es un
   poco confusa",
  "label": [
    {
      "aspect": ["contenido"],
      "evidence_span": [0, 6],
      "sentiment": "POS",
      "category": "costos"
    },
    {
```

```
      "aspect": ["contenido"],
      "evidence_span": [0, 6],
      "sentiment": "POS",
      "category": "funcionalidad"
    },
    {
      "aspect": ["flujo de uso"],
      "evidence_span": [8, 13],
      "sentiment": "NEG",
      "category": "usabilidad"
    }
  ]
}...
```


Apéndice B. Criterios de aceptación para reseñas sintéticas

El presente apéndice describe los criterios de aceptación definidos para la incorporación de reseñas sintéticas dentro del corpus, así como los análisis que respaldan dichas decisiones. Dado que la generación automática de texto mediante LLMs puede introducir inconsistencias, sesgos o construcciones poco naturales, se estableció un proceso de validación orientado a garantizar la calidad, coherencia y representatividad lingüística de las instancias generadas.

Tabla B1.

Criterios y umbrales de validación para reseñas sintéticas

Criterio	Regla de validación	Descripción
Distribución de longitud por rango válido	$[5, 31]$ palabras	Se utilizó el rango de palabras del dataset real como referencia. Las reseñas sintéticas debían ubicarse dentro de estos límites para ser consideradas representativas.

Continúa en la siguiente página

Tabla B1 – *continuación*

Criterio	Regla de validación	Descripción
Longitud textual promedio del corpus real	$\mu \in [11{,}36, 15{,}36]$ palabras	Se utilizó el promedio de palabras del dataset real ($\mu = 13{,}36$) como referencia central. Por cada lote de reseñas generadas, se calculó el promedio de palabras por reseña, el cual debía ubicarse dentro de un rango de variación de ± 2 palabras respecto al valor observado en el conjunto real (véase Figura 20).
Coherencia semántica	Similaridad temática consistente	Las reseñas debían mantener consistencia global de significado; desviaciones semánticas evidentes implicaban reescritura o descarte.
Uso de registro lingüístico colombiano	Inclusión controlada de variación léxica	Se permitió el uso moderado de expresiones propias del español colombiano, siempre que no afectaran la claridad ni la naturalidad del texto.
Ruido textual controlado	≤ 1 error ortográfico o variación leve	Se admitió como máximo una variación ortográfica o estilística por reseña, con el fin de introducir naturalidad en el texto sintético sin comprometer su calidad.
Consistencia de polaridad	Sin contradicciones internas de sentimiento	Las reseñas no debían contener cambios contradictorios de polaridad dentro del mismo texto.
Fluidez lingüística	Lectura natural sin estructuras rígidas	Se exigía coherencia sintáctica general; textos artificiales o altamente mecanizados eran reescritos.
Diversidad léxica	No repetición excesiva de patrones	Se controló la repetición de estructuras o frases genéricas entre reseñas generadas.

Nota. Los umbrales definidos fueron utilizados como criterios de control para la aceptación, modificación o descarte de reseñas sintéticas, garantizando calidad, naturalidad y representatividad del corpus generado.

Apéndice C. Acceso al repositorio

En el presente apéndice se presenta el enlace de acceso al repositorio de ***Github***. Allí, se encuentran los códigos fuentes del modelo ABSA propuesto (véase Sección 3.2) y del sitio web desarollado (véase Sección 4.3). Adicionalmente, está el conjunto de datos final, descrito en la Sección 4.1.