



UNIVERSIDAD
NACIONAL
DE COLOMBIA

Prototipo de modelo de aprendizaje automático para estimar la viabilidad de productos tecnológicos mediante la predicción de la satisfacción del consumidor

Edgar Daniel González Díaz

Universidad Nacional de Colombia
Facultad de Ingeniería, Departamento de Ingeniería de Sistemas y Computación
Bogotá D.C., Colombia
2025

Prototipo de modelo de aprendizaje automático para estimar la viabilidad de productos tecnológicos mediante la predicción de la satisfacción del consumidor

Edgar Daniel González Díaz

Tesis final presentada(o) como requisito parcial para optar al título de:
Magister en Ingeniería de Sistemas

Directora:
Ing. Elizabeth León Guzmán, PhD.

Línea de Investigación:
Procesamiento de lenguaje natural
Grupo de Investigación: MIDAS

Universidad Nacional de Colombia
Facultad de Ingeniería, Departamento de Ingeniería de Sistemas y Computación
Bogotá D.C., Colombia
2025

Profundos agradecimientos a mi familia y a la Universidad Nacional de Colombia por su constante apoyo en mi crecimiento profesional.

Resumen

Prototipo de modelo de aprendizaje automático para estimar la viabilidad de productos tecnológicos mediante la predicción de la satisfacción del consumidor

En este trabajo se propone y construye un prototipo de modelo predictor de viabilidad de productos tecnológicos, cuyo objetivo es transformar información textual estructurada en un indicador numérico. Este indicador es el resultado de la interpretación cuantitativa de la satisfacción del consumidor, expresada a través de reseñas de usuarios en el contexto del comercio electrónico, y está orientado a apoyar la planeación de productos tecnológicos, particularmente en etapas tempranas de análisis del entorno y caracterización del producto. El prototipo se fundamenta en un modelo de aprendizaje de máquina que analiza los componentes textuales presentes tanto en la información descriptiva de los productos como en las reseñas publicadas en línea, permitiendo estimar finalmente la satisfacción del consumidor como un indicador numérico. Como aproximaciones cuantificables para la construcción de este indicador, se emplean la calificación numérica otorgada por los usuarios y los votos de utilidad asociados a las reseñas.

Para la construcción del prototipo, se inició con la adquisición de un conjunto de datos compuesto por información de productos y reseñas provenientes de la plataforma de comercio en línea Amazon. Sobre este conjunto se ejecutaron procesos de exploración, limpieza y transformación de datos para obtener un conjunto de entidades y atributos representativos de las percepciones de usuarios en internet.

Por otro lado, a partir de las definiciones propuestas por diversos autores en el contexto del análisis de productos, se definió un conjunto de métricas que permiten la conversión de información semántica codificada en el indicador numérico propuesto. En función de la caracterización de dichas métricas objetivo, se construyó el prototipo de modelo de aprendizaje de máquina, el cual integra la codificación de información textual en vectores numéricos, la agrupación de productos similares para una comparación consistente de entidades y la selección de estrategias supervisadas y formuladas basadas en la calificación numérica de reseñas de usuario. Como resultado, el modelo seleccionado logró representar la viabilidad de un producto con un nivel de precisión de hasta 96%, de acuerdo con las interpretaciones de negocio analizadas. Finalmente, la estructura del prototipo fue seleccionada a partir de la evaluación de métricas de desempeño por modelo y la ejecución de flujos de validación manual.

Palabras clave: Comercio de productos, indicadores de satisfacción del cliente, viabilidad numérica de productos, reseñas de usuario.

Abstract

Prototype of a Machine Learning Model to Estimate the Viability of Technological Products through the Prediction of Consumer Satisfaction

This work proposes and develops a prototype of a predictive model for the viability of technological products, whose objective is to transform structured textual information into a numerical indicator. This indicator results from the quantitative interpretation of consumer satisfaction, expressed through user reviews in e-commerce contexts, and is intended to support technological product planning, particularly during early stages of environmental analysis and product characterization.

The prototype is based on a machine learning model that analyzes textual components present in both product descriptions and online user reviews, ultimately enabling the estimation of consumer satisfaction as a numerical indicator. As quantifiable approximations for constructing this indicator, the model employs the numerical ratings assigned by users and the helpfulness votes associated with the reviews.

For the construction of the prototype, a dataset composed of product information and reviews obtained from the Amazon e-commerce platform was collected. This dataset underwent exploration, cleaning, and transformation processes to derive a set of entities and attributes representative of user perceptions on the internet.

Additionally, based on definitions proposed by various authors in the context of product analysis, a set of metrics was defined to enable the conversion of encoded semantic information into the proposed numerical indicator. According to the characterization of these target metrics, the machine learning prototype was constructed by integrating textual information encoding into numerical vectors, clustering of similar products to ensure consistent entity comparison, and the selection of supervised learning strategies formulated based on user review ratings. As a result, the selected model achieved up to 96% accuracy in representing product viability, according to the analyzed business interpretations. Finally, the structure of the prototype was determined based on the evaluation of performance metrics per model and the execution of manual validation workflows.

Keywords: Product commerce, customer satisfaction indicators, numerical product viability, user reviews.

Este Trabajo Final de maestría fue calificado en **ABRIL** de 2026 por el siguiente evaluador:

IVAN MAURICIO RUEDA CACERES
Profesor Facultad de Ingeniería
Universidad Nacional de Colombia, Sede Bogotá

Contenido

	Pág.
Resumen.....	4
Abstract.....	5
Contenido.....	7
Lista de figuras.....	10
Lista de tablas	13
1. Introducción.....	1
1.1 Necesidad	1
1.2 Conjunto de datos.....	2
1.3 Objetivos	2
1.4 Contribuciones	2
1.5 Estructura del documento.....	3
2. Estado del arte	4
2.1 Contexto del negocio.....	4
2.2 Contexto tecnológico.....	5
2.3 Trabajos previos	7
3. Caracterización de información de productos y reseñas	14
3.1 Recolección y acceso a los datos.....	14
3.1.1 Origen de los datos	14
3.1.2 Estrategia de recolección	15
3.2 Descripción del conjunto de datos	15
3.2.1 Información de productos.....	15
3.2.2 Información de reseñas	17
3.3 Diagnóstico de calidad del conjunto de datos	18
3.3.1 Valores nulos y arreglos vacíos	18
3.4 Análisis exploratorio de los datos	21
3.4.1 Exploración de campos textuales	21
3.4.2 Exploración de esquemas en detalles de publicaciones.....	24
3.4.3 Distribución de calificaciones de usuarios	27
3.4.4 Distribución de cantidad de reseñas por producto.....	27
3.1.1 Distribución de subcategorías de producto.....	28

3.4.5	Relación de variables numéricas en reseñas de producto.....	29
3.5	Selección de variables relevantes.....	30
3.6	Limpieza y preprocesamiento.....	33
3.7	Conclusiones de capítulo.....	34
4.	Métricas para medir la satisfacción del consumidor.....	36
4.1	Recopilación de métricas candidato.....	36
4.1.1	Métrica basada en metodología SHELL.....	36
4.1.2	Métrica basada en metodología JTBD.....	36
4.1.3	Métrica basada en puntuación de usuario.....	37
4.2	Construcción y definición de métricas candidato.....	37
4.3	Conclusiones de capítulo.....	41
5.	Implementación de modelo predictor de viabilidad numérica.....	43
5.1	Metodología de modelamiento.....	43
5.2	Identificación de algoritmos y herramientas.....	44
5.2.1	Módulos pre-entrenados.....	44
5.3	Codificación de entidades.....	44
5.4	Búsqueda de productos similares.....	47
5.4.1	LSH (Locality Sensitive Hashing).....	47
5.5	Plan de entrenamiento y selección de hiperparámetros.....	48
5.6	Arquitecturas de modelado para predicción de viabilidad numérica.....	49
5.6.1	Modelo de regresión mediante valoración binaria.....	49
5.6.2	Modelo de clasificación mediante valoración categórica.....	50
5.6.3	Modelo de regresión mediante valoración categórica (incluyendo información de categoría).....	51
5.6.4	Valoración mediante modelado formulado.....	52
5.7	Conclusiones de capítulo.....	53
6.	Evaluación de modelo predictor de viabilidad numérica.....	54
6.1	Planteamiento de métricas de desempeño.....	54
6.2	Evaluación de desempeño de modelos.....	55
6.2.1	Modelo de clasificación mediante valoración binaria.....	55
6.2.2	Modelo de clasificación mediante valoración categórica.....	56
6.2.3	Modelo de clasificación mediante valoración binaria (incluyendo información de categoría).....	57
6.2.4	Modelo de regresión formulado.....	59
6.3	Ejemplificación de caso de uso.....	59
6.4	Análisis de resultados y selección de mejores modelos.....	62
6.5	Conclusiones de capítulo.....	63
7.	Conclusiones y trabajo futuro.....	64
7.1	Conclusiones.....	64
7.2	Recomendaciones.....	65
7.3	Trabajo Futuro.....	65
8.	Referencias.....	69

Lista de figuras

	Pág.
Figura 2-1. Tasa de fracaso de empresas emergentes por sector, según estudio llevado a cabo por [1] (Imagen adaptada).	5
Figura 2-2. Metodología CRISP-DM ampliamente adaptada para la estructuración del flujo de trabajo en el contexto del procesamiento de datos e inferencia a partir de los mismos. Adaptado de [9].	6
Figura 2-3. Opciones de arquitectura de modelo Universal Sentence Encoder (USE): (a) Variante Deep Averaging Network y (b) Variante basada en transformadores. Adaptado de [12].	7
Figura 2-4. Metodología SHELL adaptada para la evaluación de riesgos y causas de fracaso en Startups. Metodología planteada por [1] (Imagen adaptada).	8
Figura 2-5. Ejemplo de resultado de aplicación del modelo Jobs-To-Be-Done (JTBD) en un negocio real. Tomado de: [5] (Imagen adaptada).	9
Figura 2-6. Flujo de reconocimiento de características en textos de descripciones de productos. Adaptado de [13].	10
Figura 2-7. Propuesta de modelo de Procesamiento de Lenguaje Natural para la asociación de reseñas de usuario con parámetros de diseño de producto encontrado en trabajos previos. [17] (Imagen adaptada).	11
Figura 3-1. Diagrama descriptivo de la arquitectura medallion. Adaptado de [20]	15
Figura 3-2. Valores nulos en cada una de las columnas del conjunto de datos original para la información de productos.	19
Figura 3-3. Listas vacías para columnas con formato de lista en el conjunto de datos de productos original.	19
Figura 3-4. Listas de imágenes vacías en el conjunto de datos de reseñas.	20
Figura 3-5. Productos sin detalles especificados.	20
Figura 3-6. Distribución de cantidades de palabras en descripciones de productos.	22

Figura 3-7. Distribución de cantidades de palabras en títulos de productos.....	22
Figura 3-8. Distribución de cantidad de palabras en campos de características de productos.....	23
Figura 3-9. Distribución de TTR (Token-type ratio) para textos unificados con información de producto.	23
Figura 3-10. Distribución de largos textuales para información de reseñas	24
Figura 3-11. Distribución de campos detalle en descripciones de producto.	25
Figura 3-12. Muestra de campos de características diligenciados en detalle de productos para las 4 subcategorías más frecuentes.....	27
Figura 3-13. Distribución general de calificaciones de usuario en reseñas del conjunto de datos.	27
Figura 3-14. Distribución logarítmica de cantidad de reseñas de usuario por producto. ...	28
Figura 3-15. Distribución de subcategorías de producto.	29
Figura 3-16. Correlación de variables: Calificación, votos útiles y número de palabras. ...	30
Figura 3-17. Descripción gráfica de flujo de preprocesamiento.....	34
Figura 4-1. Proceso de búsqueda de productos con propósitos de asociación de productos similares.....	39
Figura 4-2. Proceso de cálculo de viabilidad en función de la calificación dada a un producto en reseñas individuales. (a) y (b) proponen formulaciones distintas para el tratamiento de la variable objetivo. (c) presenta un esquema del proceso de inferencia para un nuevo producto.....	40
Figura 4-3. Proceso de cálculo de viabilidad a partir de las funcionalidades conocidas para un determinado grupo de productos y las definiciones dadas por JTBD [5].....	41
Figura 5-2. Generación de codificaciones mediante PCA para información de productos	46
Figura 5-3. Generación de codificaciones mediante PCA para información de reseñas. ...	46
Figura 5-5. Modelo para la predicción de viabilidad en función de calificaciones dadas en reseñas de usuario	50
Figura 5-6. Modelo para la predicción de viabilidad en función de regresión real sobre información de calificación de reseña.....	51
Figura 5-7. Modelo para la predicción de categoría de producto.	52
Figura 5-8. Modelo para la predicción de viabilidad booleana de producto con contexto de categoría de producto elegida. La segunda entrada del modelo, proveniente de la salida del	

modelo en la Figura 5-7, corresponde a la capa de entrada a la derecha, cerca de la salida final.52

Figura 6-1. Métricas de evaluación para modelo de predicción de viabilidad en función de clasificación binaria sobre información de calificación de reseña.....55

Figura 6-2. Matriz de confusión para modelo de predicción de viabilidad en función de clasificación binaria sobre información de calificación de reseña.....56

Figura 6-3. Métricas de evaluación para modelo de predicción de viabilidad en función de clasificación multiclase sobre información de calificación de reseña.57

Figura 6-4. Matriz de confusión para modelo de predicción de viabilidad en función de clasificación multiclase sobre información de calificación de reseña.57

Figura 6-5. Métricas de evaluación para modelo de predicción de viabilidad basado en etiqueta booleana modificado para inclusión de información de categoría de producto.58

Figura 6-6. Matriz de confusión para modelo de predicción de viabilidad basado en etiqueta booleana modificado para inclusión de información de categoría de producto.59

Figura 6-7. Representación de flujo completo de procesamiento de un nuevo ejemplo entrante al proceso de inferencia construido.....59

Lista de tablas

	Pág.
Tabla 2-1. Resumen de propuestas tecnológicas relacionadas con el entendimiento de las características y capacidades de productos de uso final.	13
Tabla 3-1. Campos presentes en entidad de producto en conjunto de datos original.	17
Tabla 3-2. Campos presentes en entidad de reseña en conjunto de datos original.	18
Tabla 3-3. Resumen de inclusión y exclusión de campos de conjunto original según análisis realizado.	32
Tabla 3-4. Subcategorías seleccionadas para análisis.	33
Tabla 4-1. Ponderación de componentes textuales de producto para generación de representación final.	38
Tabla 6-1. Estructura de ejemplo generado como entrada para flujo de modelo.....	60
Tabla 6-2. Estructura de productos similares asociados a ejemplo de la Tabla 6-1.	61
Tabla 6-3. Resultados de predicción para ejemplo de flujo completo planteado.	62

1. Introducción

El comercio de productos, tanto en el ámbito de empresas consolidadas en el mercado como en el de startups emergentes, constituye un mecanismo de generación de valor que, particularmente en el sector tecnológico, implica asumir un riesgo elevado acompañado del potencial de obtener beneficios significativos [1].

En este sentido, debido a la alta competitividad del mercado y la influencia de múltiples factores externos, alrededor de un 71% de los nuevos startups fracasan dentro de sus primeros diez años de operación [2]. Por su parte, el trabajo en [3] ofrece una perspectiva aún más pesimista, al exponer hallazgos en la literatura que indican que solo 1 de cada 5.000 lanzamientos de nuevos productos logra alcanzar el éxito comercial. No obstante, este mismo estudio identifica a la planeación estratégica y la mejora en la calidad del producto como factores clave durante la evolución de la iniciativa.

En relación con la planeación estratégica, [4] destaca el modelo Canvas como herramienta fundamental para descubrir necesidades en el mercado, aprovechar oportunidades y establecer una visión a corto, mediano y largo plazo. De manera similar, diversos modelos proponen estructuras formales para definir propuestas de valor y diseñar modelos de negocio completos dentro de un enfoque orientado hacia la innovación y la generación de valor a través de la exploración de nuevos mercados y la expansión en los mercados conocidos.

Sin embargo, el estudio presentado en [5] advierte sobre las limitaciones de estos enfoques, ya que tienden a introducir sesgos respecto a las verdaderas necesidades del cliente, al asumir que el valor del producto ofrecido responde efectivamente a dichas necesidades. El mismo trabajo hace énfasis en la importancia de una identificación sistemática de los requerimientos del cliente y de anticipar el desempeño del producto desde la perspectiva del usuario final, centrándose en las funcionalidades y las acciones que el consumidor desea realizar. La validación temprana se convierte, por tanto, en un desafío fundamental, especialmente cuando las decisiones se basan en supuestos no soportados con evidencia empírica.

1.1 Necesidad

Ante este panorama, se hace evidente la necesidad de una herramienta que facilite la validación de las necesidades del cliente a partir de las actividades que este busca realizar. Esta herramienta permitiría evaluar la viabilidad de un producto tecnológico mediante la identificación de necesidades expresadas por potenciales clientes y usuarios de soluciones existentes, así como la homologación de dichas necesidades con las características de una mejora o un nuevo producto que se pretende lanzar al mercado. En este trabajo, la *viabilidad de producto* se entiende como un *indicador cuantitativo* asociado a la *satisfacción esperada*

del consumidor, la cual es inferido a partir de la información textual relacionada con la experiencia de uso reportada por los usuarios de este producto.

1.2 Conjunto de datos

Dado el ecosistema digital actual, la información necesaria para el desarrollo y validación de la herramienta propuesta fue extraída de una plataforma de comercio en línea, específicamente del conjunto de datos Amazon Reviews [6]. Este conjunto fue recopilado a partir de la plataforma de comercio homónima e incluye información textual sobre productos y reseñas realizadas por usuarios, todas ellas en idioma inglés. El uso de este conjunto de datos permite incorporar un enfoque global que incluye la percepción y experiencia del usuario final frente a productos del sector tecnológico. El subconjunto de productos en este trabajo corresponde a un total de cuatro categorías principales correspondientes a los sectores de Software, Electrónicos, Celulares y accesorios y videojuegos.

1.3 Objetivos

El propósito principal de este trabajo es la generación de un prototipo de modelo de aprendizaje de máquina que, a través de información de productos en el sector tecnológico en formato textual, permita la predicción de un indicador numérico que represente la viabilidad del producto en función de la satisfacción esperada del consumidor. A fin de cumplir con este objetivo, se plantearon los siguientes objetivos específicos:

- Procesar un conjunto de datos recopilado desde un repositorio público que contiene información sobre productos y valoraciones de usuarios, mediante técnicas de limpieza, transformación y normalización de datos.
- Seleccionar un conjunto de indicadores cuantitativos que permitan estimar la satisfacción esperada del consumidor en función de las características del producto.
- Construir un modelo de aprendizaje automático capaz de transformar las características del producto en un indicador numérico que refleje la satisfacción esperada del consumidor.
- Evaluar el desempeño del modelo mediante métricas cuantitativas basadas en la precisión e indicadores específicos del modelo final, así como métricas cualitativas basadas en pruebas manuales.

1.4 Contribuciones

- Conjunto de datos preprocesado con información consolidada y seleccionada representativa de entidades de producto y reseña. Como resultado final del proceso se cuenta con un conjunto de 64.618 productos y 2'984.015 reseñas correspondientes. Se incluyen además conjuntos de datos relevantes para el resultado final.
- Repositorio con implementación de flujos de análisis, procesamiento, entrenamiento y evaluación de conjunto de datos. Se incluyen todos los notebooks de Jupyter asociados, así como tecnologías que incluyen PySpark y Tensorflow.

-
- Un sistema predicción de viabilidad numérica en función de información textual de producto, con resultados que alcanzan hasta un 96% de precisión bajo una implementación basada en redes neuronales para clasificación binaria.
 - Artículo académico pendiente por publicación. Se presentan los resultados descritos en este documento, así como el proceso general realizado.

1.5 Estructura del documento

El presente documento se estructura de la siguiente manera:

- **Capítulo 2: Estado del arte.** Resumen de trabajos previos y marco teórico relevante.
- **Capítulo 3: Características del conjunto de datos.** Análisis descriptivo del conjunto de datos, discusión sobre variables objetivo y planteamiento de flujo de preprocesamiento en función de resultados obtenidos.
- **Capítulo 4: Definición de métrica para estimación de indicador de viabilidad.** Se evalúan alternativas inspiradas por la literatura estudiada que permitan cuantificar la viabilidad de productos en función de la satisfacción de consumidor. Se asocia además el cálculo de dichas alternativas desde la perspectiva de la estructura del conjunto de datos conocida.
- **Capítulo 5: Implementación de modelo predictor de viabilidad numérica.** En función de las definiciones dadas en el capítulo anterior, se diseña, implementa y entrenan los candidatos de modelo predictor para cada una de las etapas necesarias que hacen parte del cálculo de las métricas definidas.
- **Capítulo 6: Evaluación de modelo predictor de viabilidad numérica.** Se evalúan individualmente las etapas del modelo final, el flujo de inferencia completo a través de pruebas manuales con información del conjunto de datos, y los resultados generales del trabajo.
- **Capítulo 7: Conclusiones.** Se listan conclusiones del proyecto y el trabajo futuro a desarrollar.

2. Estado del arte

2.1 Contexto del negocio

La generación de valor a través del comercio de productos ha sido una estrategia ampliamente adoptada por las compañías. Como propone [3], este comportamiento busca tanto expandir su influencia en el mercado como para sobrevivir a los cambios inherentes que este puede experimentar. Sin embargo, revisiones de la literatura llevadas a cabo por dicho trabajo evidencian una alta tasa de fracaso en iniciativas de mercado, destacando que solo 1 de cada 5.000 productos logra representar un caso de éxito. Asimismo, incluso en negocios grandes y bien establecidos, estos lanzamientos suelen enfrentar expectativas de éxito bajas o moderadas.

Billah en [3] identifica factores relacionados con la baja inversión en marketing dirigido al consumidor, posicionamiento débil, baja calidad en relación con la competencia y la distribución de producto deficiente o limitada. Estos factores abarcan distintas etapas del proceso de evolución de un producto, comenzando con la fase de planeamiento estratégico y extendiéndose hasta aspectos logísticos y de distribución del producto final.

Por otro lado, [5] agrupa estudios que sugieren una tasa de fracaso de hasta el 95% en el desarrollo de nuevos productos, así como un menor énfasis por la generación de valor a través de esta estrategia en algunos sectores, como el sector automotriz en Alemania, donde apenas el 14% de participación en la generación de valor se realiza a través de soluciones personalizadas.

Los autores presentan un análisis similar al realizado por [3], atribuyendo la baja tasa de éxito en innovación a causas relacionadas con el sobreajuste del diseño del producto respecto a las necesidades del consumidor, una retroalimentación limitada sobre las tareas y necesidades reales del cliente, y la baja aceptación de los productos lanzados al mercado. Estas causas se originan principalmente en etapas tempranas de planteamiento del producto. En respuesta al riesgo latente de fracaso representado por las cifras ya discutidas, los trabajos relacionados proponen prácticas y metodologías orientadas a la generación de modelos de negocio estructurados. Entre estas metodologías se encuentran los Product Roadmaps, Segmentaciones de Mercado e Ingeniería de Requerimientos.

Adicionalmente, múltiples indicadores han sido definidos para medir el desempeño de nuevos modelos de negocio antes de ser lanzado un producto. Algunos de estos indicadores incluyen el beneficio neto, la cuota de mercado (market share), intensidad de la innovación y la complejidad de la oferta de mercado. Paralelamente, se ha estudiado cómo los efectos que tiene la innovación dentro del mercado competitivo generan efectos de réplica sobre los competidores directos e indirectos dentro del segmento de negocio [7].

De manera complementaria, se han propuesto indicadores adicionales diseñados para cuantificar características de productos en función de múltiples dimensiones de interés de sus clientes objetivo. En este contexto, [8] estudia la definición de indicadores de desempeño de un producto tecnológico inteligente en etapas tempranas de planeación, considerando características como autonomía, capacidad de aprendizaje, reactividad, cooperación con otros dispositivos, interacción con seres humanos y rasgos de personalidad.

Más allá de las apreciaciones específicas, los autores en [8] describen la *satisfacción del consumidor* como un indicador transversal ampliamente desarrollado por múltiples

investigaciones. La *satisfacción del consumidor* se define como la evaluación afectiva de un producto o servicio por parte del consumidor, en conjunto con el grado en el cual el producto o servicio satisface o excede las expectativas del mismo.

En consecuencia, el problema se reduce a la generación de un método por medio del cual, a partir de la información de productos y un punto de referencia en particular, pueda estimarse la satisfacción del consumidor, en particular para el contexto de productos en el sector tecnológico, como un indicador tangible que pueda reflejar la viabilidad de un producto en términos de la maximización de dicha satisfacción o solvencia de necesidades.

Finalmente, en relación con la delimitación del problema, el análisis llevado a cabo por [1] y la revisión de trabajos previos han permitido identificar al sector tecnológico como uno de los más afectados por altas tasas de fracaso en el lanzamiento de productos. En este sector destacan industrias como aplicaciones móviles, herramientas de Big Data, software y plataformas de comercio electrónico, tal como se ilustra en la Figura 2-1.

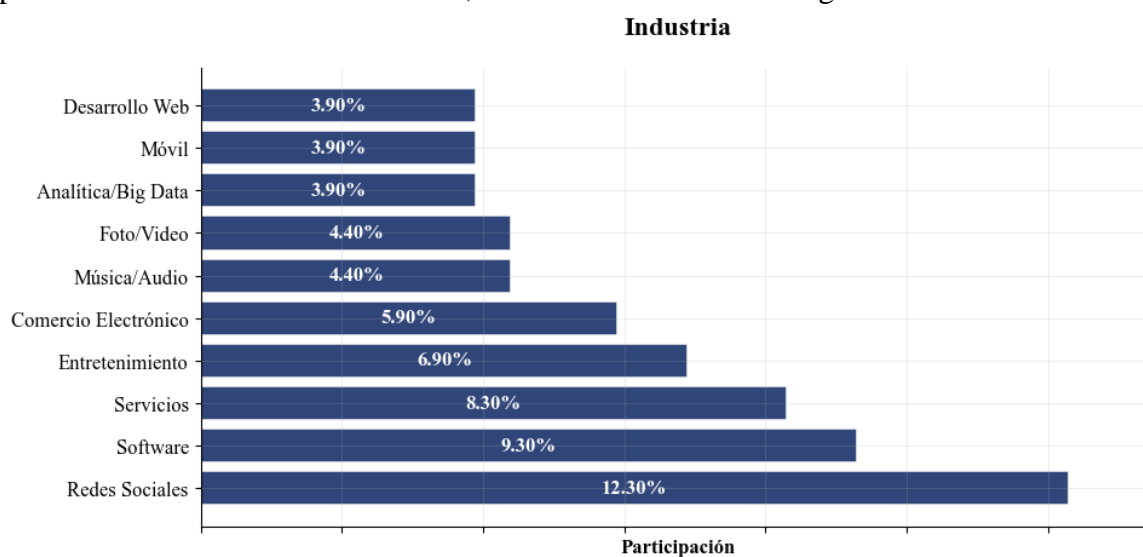


Figura 2-1. Tasa de fracaso de empresas emergentes por sector, según estudio llevado a cabo por [1] (Imagen adaptada).

2.2 Contexto tecnológico

Anteriormente se discutió el objetivo general a nivel de negocio; sin embargo, es necesario resaltar el contexto tecnológico que permite articular las prácticas y arquitecturas existentes con el desarrollo del presente trabajo. En este sentido, CRISP-DM [9] propone una metodología ampliamente estandarizada para el desarrollo e implementación de un flujo de datos robusto independientemente de la tecnología usada o el negocio destino. De esta manera, facilita la agilidad y la implementación de buenas prácticas a lo largo del ciclo de vida de proyectos de análisis e inferencia.

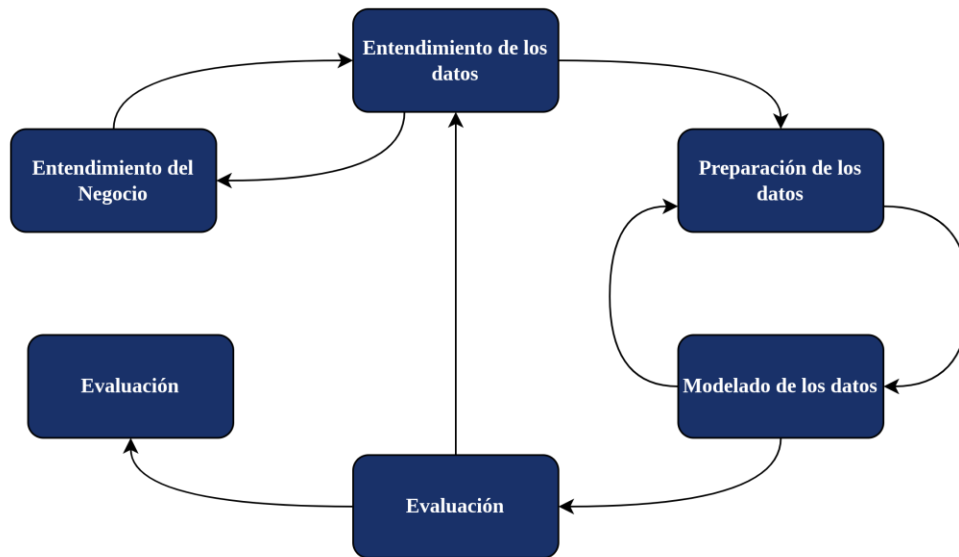


Figura 2-2. Metodología CRISP-DM ampliamente adaptada para la estructuración del flujo de trabajo en el contexto del procesamiento de datos e inferencia a partir de los mismos. Adaptado de [9].

La Figura 2-2 presenta las etapas definidas por el modelo CRISP-DM y el flujo de transición entre estas. El proceso inicia con la fase de entendimiento del negocio, en la cual se establecen los objetivos del proyecto y se identifican los requerimientos que se busca satisfacer. A continuación, la etapa de entendimiento de los datos permite analizar la estructura, distribución y características generales de la información disponible, de acuerdo con los objetivos previamente definidos.

Posteriormente, se desarrolla la etapa de preparación de los datos, que comprende actividades de limpieza, selección de atributos y construcción de nuevas características que puedan ser utilizadas como entrada para las técnicas de modelado. Las fases de modelamiento y evaluación incluyen el diseño, entrenamiento, ajuste y validación de modelos de aprendizaje de máquina orientados a resolver el problema planteado. Finalmente, la metodología contempla una etapa de despliegue, asociada a la puesta en práctica del modelo resultante en un entorno real, donde se evalúa su capacidad de generalización a partir de datos no observados previamente.

Ahora bien, vale la pena definir una de las herramientas tecnológicas que facilitarán el cumplimiento del objetivo de este trabajo, principalmente, en el ámbito del procesamiento de lenguaje natural para la condensación de información de tipo textual.

Universal Sentence Encoder es propuesto para la codificación numérica de la semántica de oraciones según dos modelos preliminares basados en redes neuronales y cuya diferencia recae en la complejidad de la estructura subyacente [10]. Particularmente, se cuenta con las arquitecturas descritas en la Figura 2-3. Ambas alternativas facilitan la codificación numérica de la semántica de oraciones en idioma inglés, sin embargo, su diferencia recae en la complejidad del modelo usado para dicho propósito. La Figura 2-3 (a) presenta una arquitectura ligera basada en codificación de tokens y una red neuronal simple para la generación del vector representativo de una oración. Por otra parte, la Figura 2-3 (b) presenta una arquitectura más compleja implementada a través de transformadores [11].

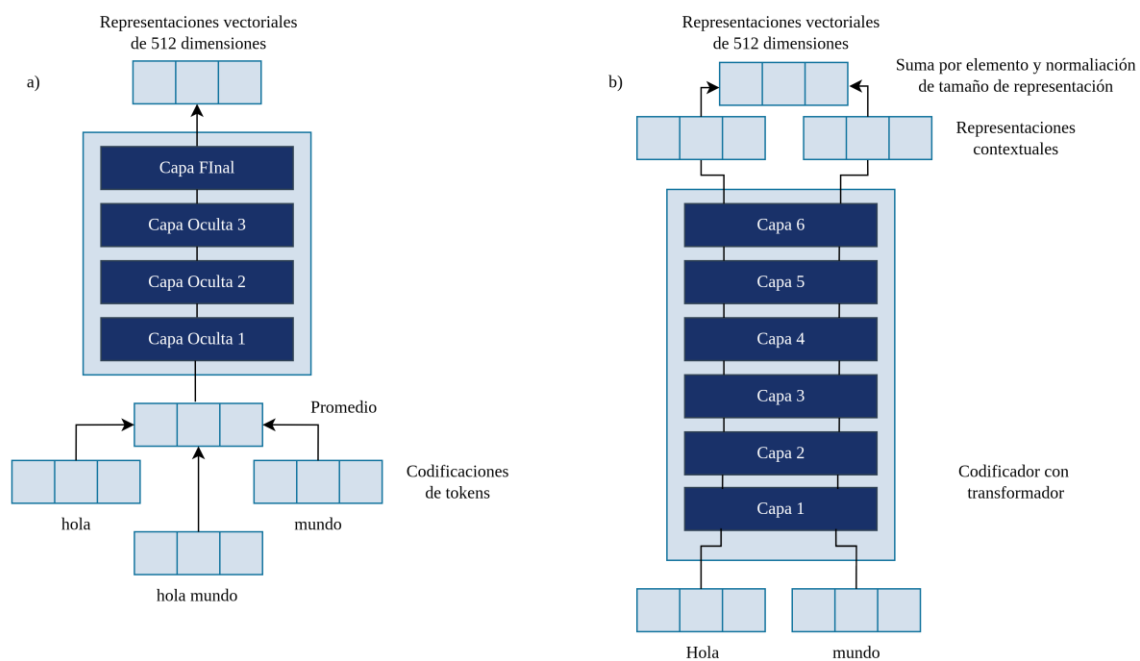


Figura 2-3. Opciones de arquitectura de modelo Universal Sentence Encoder (USE): (a) Variante Deep Averaging Network y (b) Variante basada en transformadores. Adaptado de [12].

2.3 Trabajos previos

Un análisis a alto nivel con relación al planteamiento de la viabilidad de productos fue llevado a cabo por [1]. Se propone a SHELL como una metodología adaptada desde contextos externos para cualificar y cuantificar los riesgos a los que se enfrenta un negocio dentro del mercado. La Figura 2-4 representa las generalidades del modelo como una metodología nacida de la necesidad de proponer posibles causas que expliquen el fracaso de un producto lanzado al mercado.



Figura 2-4. Metodología SHELL adaptada para la evaluación de riesgos y causas de fracaso en Startups. Metodología planteada por [1] (Imagen adaptada).

En términos generales, SHELL genera una apreciación de riesgos y causas de fracaso en función de calificaciones proporcionadas por personal especializado para cada uno de los 5 componentes presentados en la Figura 2-4, así como a las causas específicas de cada componente. Estas apreciaciones se dan en términos de porcentajes que posteriormente se comparan para definir los elementos de mayor relevancia y cuantificar el riesgo como un resultado numérico final.

De manera similar, [5] propone la metodología JTBD (Jobs-To-Be-Done) como una aproximación estructurada a partir de la evaluación cuantitativa de las actividades o “Jobs” que los usuarios buscan realizar mediante un producto. Adicionalmente, se incorporan factores contextuales como el crecimiento del sector y la voluntad de compra, siendo estos últimos cuantificados de manera controlada por evaluaciones externas e internas por personal especializado. El trabajo mencionado define al valor agregado de un “Job” o actividad para los interesados como $V_{J_n M_k}$, descrito mediante la ecuación (1), donde S_{ik} es el valor agregado para un interesado i dentro del segmento de mercado k , y m es el total de interesados. Originalmente, los autores presentan una evaluación sencilla de S_{ik} como una valoración en el rango 1-5 proveída por usuarios potenciales, sin embargo, la medición podría adaptarse a una estimación mediante una escala equivalente.

$$V_{J_n M_k} = \frac{\sum_{i=1}^m S_{ik}}{m} \quad (1)$$

Por otra parte, para cada segmento de mercado k , se condensa la voluntad de compra (WB) y la capacidad de crecimiento (GP) en un indicador llamado atraktividad del segmento de mercado (A_k), como se presenta en la ecuación (2). Cabe resaltar que la voluntad de compra y la capacidad de crecimiento son evaluados de manera cualitativa por un equipo interno, como lo señalan los autores.

$$A_k = \frac{WB_k + GP_k}{2} \quad (2)$$

Como etapa final, se evalúan los segmentos de mercado identificados mediante una representación gráfica del valor agregado para el cliente vs la atraktividad del segmento, es decir, comparando $V_{J_n M_k}$ con A_k . La Figura 2-5 presenta un ejemplo de aplicación de este método. Cada punto representa un segmento de mercado analizado.

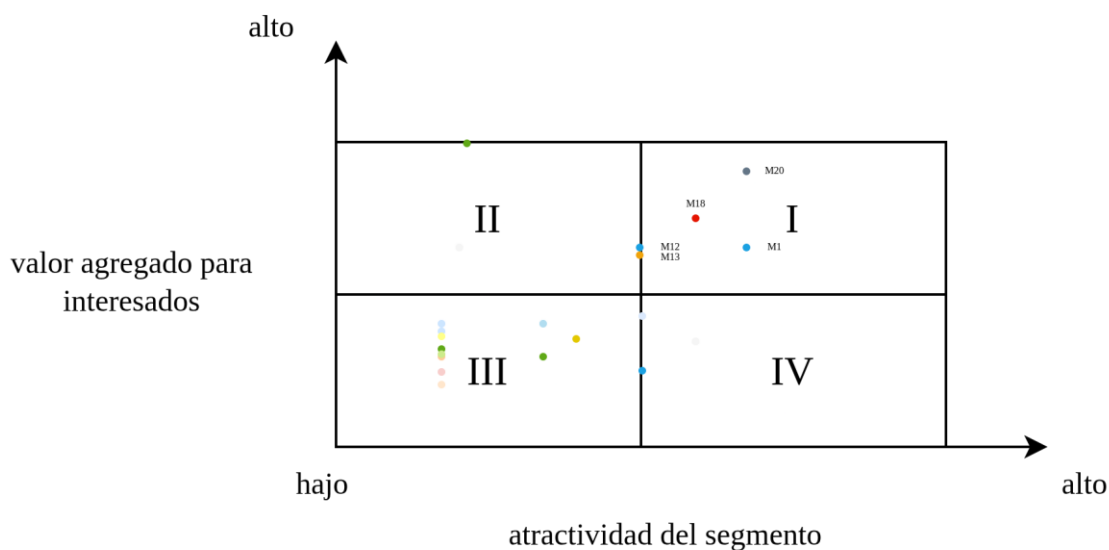


Figura 2-5. Ejemplo de resultado de aplicación del modelo Jobs-To-Be-Done (JTBD) en un negocio real. Tomado de: [5] (Imagen adaptada).

Una vez definidos algunos estándares para la cuantificación y cualificación de las características de un negocio plasmadas en un producto, es necesario identificar metodologías y técnicas que permitan la extracción y procesamiento de elementos de definición dados generalmente durante el planteamiento de una idea de negocio o producto. El trabajo en [13] propone un flujo de procesamiento basado en el reconocimiento de características textuales representadas en formatos booleanos, textuales y numéricos, a partir dentro de las descripciones dadas a productos publicados en internet. Este trabajo supone la existencia de un diccionario de búsqueda que reduzca los nombres de características a encontrar y la aplicación de métodos determinísticos basados en expresiones regulares para búsqueda de atributos de producto especificados en componentes textuales extensos. La Figura 2-6 presenta con mayor claridad este flujo a través de un ejemplo de reconocimiento de características en un texto extraído de una tienda en línea.

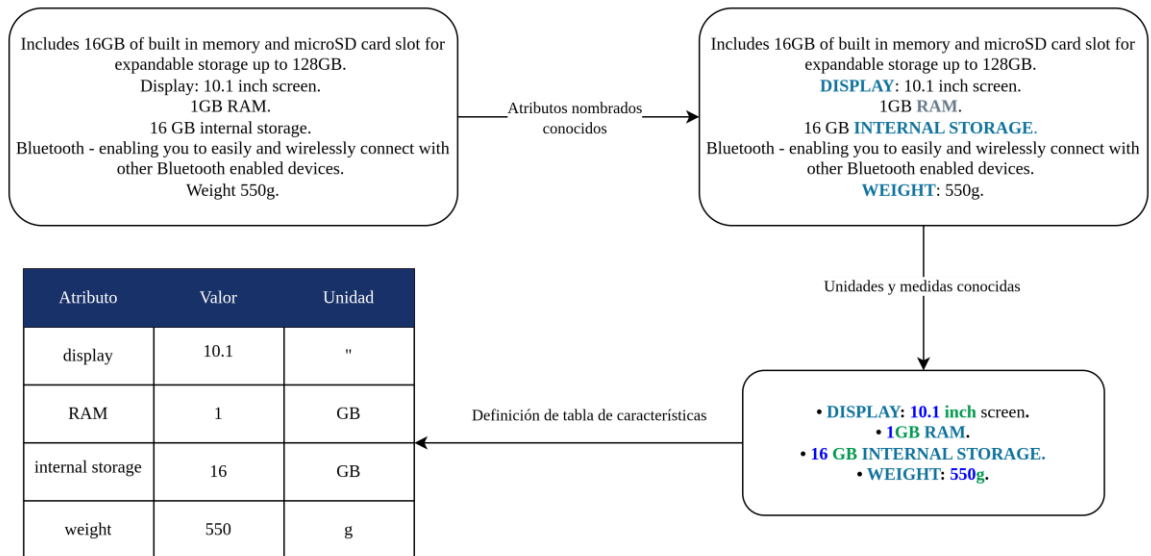


Figura 2-6. Flujo de reconocimiento de características en textos de descripciones de productos. Adaptado de [13].

De manera complementaria, el trabajo en [14] trata la relevancia de los modelos de recomendación y su desempeño real sobre las decisiones finales tomadas por clientes, siendo este un excelente representante del impacto real que se esperaría obtener a partir de modelos de aprendizaje de máquina especializados en la identificación de productos y relacionamiento con posibles clientes.

Por otro lado, [15] y [16] describen a grandes rasgos las generalidades y consideraciones del procesamiento de lenguaje natural, abarcando desde la extracción de elementos de oración hasta el análisis de sentimientos para la identificación de percepción del usuario respecto a un producto a través de reseñas.

En este contexto, [17] presenta una de las aproximaciones más relevantes en relación con el objetivo de este trabajo, describiendo un modelo de Deep Learning destinado a transformar necesidades de clientes en especificaciones de diseño de productos. A través del procesamiento de lenguaje natural a partir de herramientas como las redes neuronales recurrentes (RNNs), los autores generaron un sistema capaz de identificar características de diseño de productos que satisfacen necesidades expresadas por reseñas de usuario provenientes de un conjunto de datos de e-commerce en una plataforma de comercio en China.

El trabajo en [17] se validó sobre un subconjunto de la información de la tienda en línea, correspondiente a información de dispositivos móviles. La Figura 2-7 presenta la arquitectura general del modelo propuesto por los autores. Se describe además la identificación de palabras clave presentes en descripciones de productos como un método de filtrado de las oraciones más relevantes, para, posteriormente, entrenar un modelo clasificador que permita asociar parámetros de diseño de productos con las descripciones textuales de necesidades de usuario inferidas a partir de referencias reales provenientes de reseñas de producto. Cabe resaltar el uso de redes neuronales recurrentes a través de celdas

LSTM bidireccionales para el procesamiento de textos y la generación de codificaciones textuales.

Vale la pena destacar las afirmaciones que los autores realizan sobre algunos estudios relacionados, donde se justifica la viabilidad del uso de referencias de usuario en internet como información representativa de las necesidades de clientes reales. Entre las lecciones aprendidas, se señala el crecimiento de la complejidad en la identificación de necesidades de usuario, así como la necesidad de implementar modelos que se puedan adaptar a esta complejidad con un rango de información mayor al conjunto de datos usado para el entrenamiento del modelo desarrollado en el estudio.

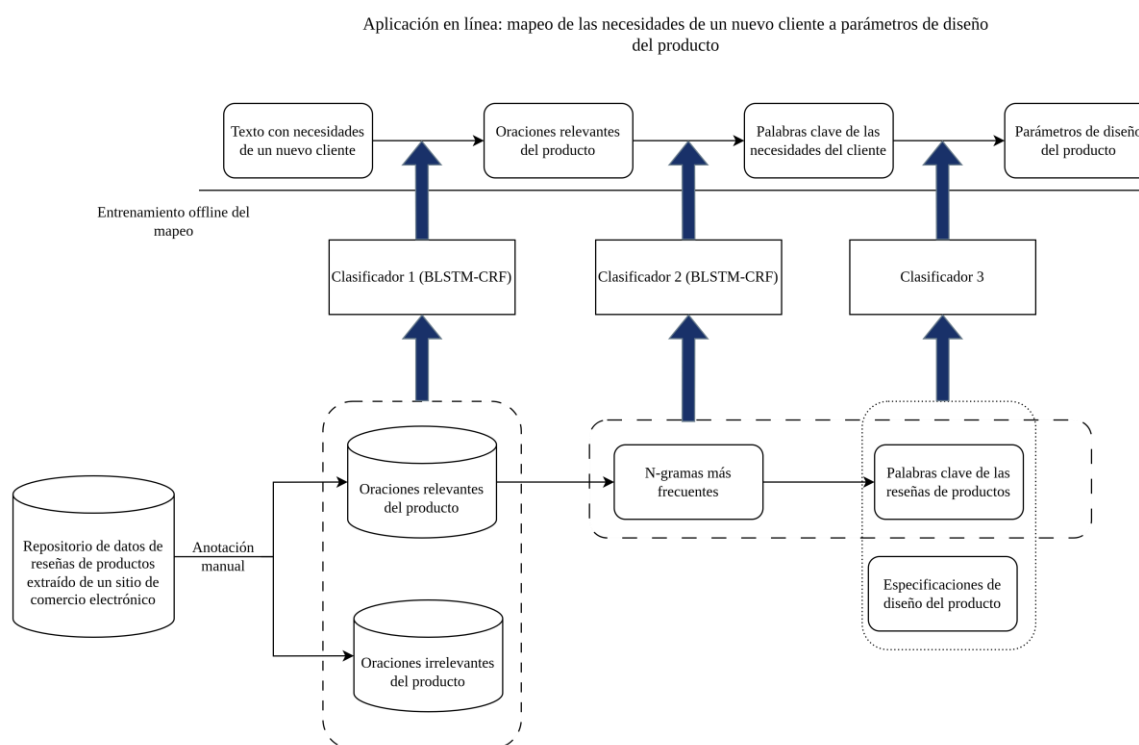


Figura 2-7. Propuesta de modelo de Procesamiento de Lenguaje Natural para la asociación de reseñas de usuario con parámetros de diseño de producto encontrado en trabajos previos. [17] (Imagen adaptada)

Una aproximación similar fue llevada a cabo por [18], cuyo trabajo tuvo como propósito la creación de un modelo probabilístico basado en Naive Bayes con la capacidad de asociar requerimientos de usuario con configuraciones de producto. Nuevamente se expresa la necesidad de parametrizar atributos de diseño de productos desde las perspectivas de clientes potenciales, en este caso, recolectadas a través de la realización de encuestas.

Cabe destacar que las tecnologías subyacentes previamente descritas para los dos modelos pueden replantearse a la luz de enfoques más recientes, como el Universal Sentence Encoder (USE) expuesto en [10] y cuya arquitectura se describió a detalle en la Figura 2-3, el cual ofrece ventajas en términos de representación semántica y escalabilidad frente a modelos basados exclusivamente en RNN.

El trabajo desarrollado en [19] constituye uno de los antecedentes más cercanos a la medición de satisfacción de usuarios en función de reseñas publicadas en la plataforma de comercio en línea Amazon.

En términos generales, [19] propone un indicador numérico para medir la satisfacción del consumidor a partir de la calificación en estrellas otorgada a cada reseña (en una escala discreta de 0 a 5). El trabajo considera además atributos como la fecha de publicación de la reseña, reputación del usuario que publica y la calificación de utilidad o “helpful votes”. Las ecuaciones (3), (4) y (5) describen el cálculo de la satisfacción final asociada a un producto.

$$L_t = \frac{\frac{2}{N+1} \sum_{i=(k-1)t}^{(k-1)t+N} \left(\frac{N-1}{N+1}\right)^i ER_{(k-1)t+N-i}}{\frac{2}{N+1} \sum_{i=(k-1)t}^{(k-1)t+N} \left(\frac{N-1}{N+1}\right)^i \cdot E_{(k-1)t+N-i}} \quad (3)$$

$$E_i = (1 + \theta) \left(1 + \varphi \frac{H_i}{\sqrt{T_i}}\right) \quad (4)$$

$$ER_i = E_i R_i \quad (5)$$

Como se mencionó previamente, se considera la reputación del usuario θ , medida como una ponderación adicional dada al usuario cuando hace parte del programa Amazon Vine Voice, es decir, usuarios especializados en la realización de reseñas útiles e insesgadas. De forma similar, φ representa la ponderación de la cantidad de votos H_i , donde cada voto señala que una reseña fue útil para un usuario de la plataforma. El total de votos de todas las reseñas para el producto padre de la reseña i se representa como T_i . Mientras que E_i representa la efectividad de una reseña i , así como ER_i representa la efectividad ponderada de una única reseña con un valor en estrellas R_i . Finalmente, L_t refleja la satisfacción ponderada de un producto para un instante t . Cabe mencionar que algunas variables adicionales, principalmente aquellas que definen el recorrido de la sumatoria principal de valoraciones, representan la recolección de calificaciones propuesta por el documento a través de intervalos fijos de publicación de reseñas.

La Tabla 2-1 presenta un resumen de los trabajos previos y aproximaciones técnicas a través de los cuáles se realizan procesos de caracterización de información de productos, codificación de necesidades de clientes, e incluso predicción de viabilidad en función de componentes específicos.

Autor	Tecnologías descritas
Cer, D. et al. [10]	Modelo de propósito general para la codificación de secuencias de tokens a vectores representativos. Útil para tareas de aumentación de datos o incluso incorporación de otros modelos a través de Transfer Learning.
Gallin et al. [14]	Los modelos de recomendación como oportunidades de entendimiento de las necesidades del cliente y las percepciones generadas por un producto sobre su usuario final.
Chiche et al. [15]	Diferentes métodos para la extracción de categorías gramaticales enfocado hacia la mejora de modelos complejos de procesamiento de lenguaje natural.

Praveen et al [16]	Aprovechamiento de los modelos de procesamiento de lenguaje natural para la decodificación y entendimiento de las referencias de usuarios sobre los productos que han usado.
Wang et al. [17]	Modelo de Deep Learning para la inferencia de especificaciones de diseño de productos a partir de necesidades de usuario caracterizadas desde productos tecnológicos en línea. Se propone el uso de referencias de artículos en línea como fuente de información de necesidades de clientes.
Jiao et al. [18]	Modelo de clasificación basado en Naive-Bayes para la asociación de requerimientos de cliente con configuraciones pre-establecidas de productos en el comercio electrónico.
Liu et al. [19]	Estructuración matemática de métrica para estimar la satisfacción del consumidor en función de la calificación efectiva a partir de las reseñas de producto en una tienda en línea.

Tabla 2-1. Resumen de propuestas tecnológicas relacionadas con el entendimiento de las características y capacidades de productos de uso final.

Retomando la definición propuesta por [8], la satisfacción del consumidor puede entenderse como la evaluación afectiva que realiza el usuario sobre un producto o servicio, en función del grado en que este cumple o supera sus expectativas. Al integrar esta perspectiva con los aportes de [16], [18] y [19], relacionados con el análisis de reseñas en línea, se identifica una oportunidad para construir una métrica que permita reconocer fortalezas y debilidades de un producto en función de las necesidades expresadas por usuarios de productos similares a nivel global.

En síntesis, se han identificado múltiples modelos orientados a estimar la viabilidad de un producto o servicio, centrados principalmente en la percepción del usuario y en sus necesidades manifiestas. No obstante, muchos de estos enfoques se basan en definiciones procedimentales que dependen de muestras del mercado objetivo, lo cual implica costos asociados a la planificación, recopilación de información y toma de decisiones, además del riesgo de sesgos derivados del alcance limitado de las consultas y la interpretación de necesidades. Asimismo, modelos computacionales como los propuestos en [17] y [18] presentan limitaciones relacionadas con la escala de los datos y la restricción a categorías específicas de producto, así como con el uso de arquitecturas que pueden ser superadas por enfoques más recientes, como los modelos basados en transformadores [11].

Tras la revisión de las metodologías y modelos existentes, se identifica la necesidad de una herramienta que permita estimar de forma cuantitativa la viabilidad de un producto a partir de sus características y de las necesidades reales expresadas por potenciales clientes. En este contexto, el sector tecnológico, caracterizado por su dinamismo y alta tasa de innovación, se presenta como un escenario idóneo para el diseño y validación de dicha herramienta.

3. Caracterización de información de productos y reseñas

En este capítulo se presenta el proceso de extracción, caracterización, exploración y procesamiento del conjunto de datos de productos y reseñas utilizado durante el desarrollo de este trabajo. En este sentido, a partir del conjunto de datos Amazon Reviews [6], el cual contiene información de productos y reseñas de la plataforma de comercio en línea Amazon, se realiza un análisis exploratorio que deriva en la caracterización y definición de estructura de la información.

El análisis exploratorio de datos llevado a cabo incluye la descripción de los resultados obtenidos tras el consumo de la información en términos de calidad del conjunto de datos, distribución de variables y correlación entre estas.

Posteriormente, con base en los resultados del análisis exploratorio, se plantea un esquema de limpieza y preprocesamiento de información basado en Medallion Architecture [20], generando finalmente un conjunto de datos con una estructura viable para posteriores pasos de modelamiento.

3.1 Recolección y acceso a los datos

El conjunto de datos seleccionado corresponde a Amazon Reviews [6], originalmente recolectado por el grupo McAuley Lab de la University of California, San Diego [21], referente al sector de negocio de comercio electrónico de Amazon. Este conjunto contiene reseñas de productos y valoraciones proporcionadas por usuarios durante la operación del comercio electrónico en la plataforma hasta el año 2023.

3.1.1 Origen de los datos

La información se encuentra disponible en un repositorio público accesible a través de internet y cuya publicación se describe en una página pública con enlaces de descarga segmentados por categoría de producto. Esta segmentación permite descargar únicamente subconjuntos específicos de la información total, reduciendo así la carga computacional asociada al preprocesamiento y al análisis exploratorio.

3.1.2 Estrategia de recolección

Los archivos que contienen la información de reseñas y datos de productos se encuentran almacenados en formato *jsonl* (*JavaScript Object Notation - Lines*), una extensión del conocido formato JSON adaptado para grandes volúmenes de información por archivo mediante un formato separado por líneas individuales.

Dado que el repositorio de información permite la descarga libre de archivos, la estrategia de recolección se reduce a la descarga de las categorías preseleccionadas de productos, para su posterior cargue y transformación en archivos de menor tamaño a través de *PySpark* [21]. De manera particular, se escoge el formato *parquet* como la opción más conveniente como formato destino dada su versatilidad, capacidad de compresión y compatibilidad con herramientas relacionadas a este trabajo.

Las categorías de producto preseleccionadas para examinación, elegidas de manera manual bajo el contexto de productos relacionados al sector tecnológico, corresponden a software, electrónicos, celulares y accesorios, y *videojuegos*.

Se definió un flujo de preprocesamiento basado en una arquitectura Medallion [20]. Esta decisión permite estructurar no solo la presente fase de entendimiento y transformación de los datos, sino también las etapas posteriores de modelamiento y evaluación, aprovechando la versatilidad de los distintos niveles del catálogo de datos para múltiples propósitos. La Figura 3-1 presenta una caracterización a gran escala de la arquitectura de implementación usada.

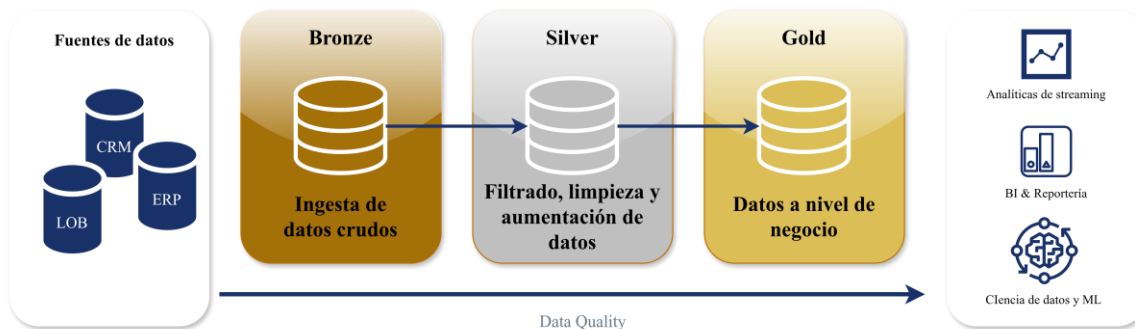


Figura 3-1. Diagrama descriptivo de la arquitectura medallion. Adaptado de [20]

3.2 Descripción del conjunto de datos

El conjunto de datos se encuentra dividido en dos secciones correspondientes a la información de productos y la información de reseñas de usuarios a dichos productos. Se describe a continuación a detalle cada una de estas secciones.

3.2.1 Información de productos

La información de cada producto se compone principalmente de las descripciones textuales y definiciones en lenguaje natural que describen las características de este en lenguaje inglés. Sin embargo, se incluyen datos adicionales relativos al precio, la tienda que lo publica y subcategorías adicionales usadas para su clasificación en los sistemas de búsqueda de la

plataforma. De manera detallada, se describen las columnas de la entidad producto en la Tabla 3-1.

Columna	Descripción
Título	Título de la publicación. Nombre en conjunto de datos original: <i>title</i>
Categoría principal	Segundo nivel de categoría del producto. Nombre en conjunto de datos original: <i>main_category</i>
Características	Lista de funcionalidades y capacidades del producto. Nombre en conjunto de datos original: <i>features</i>
Descripción	Descripción condensada en una lista de párrafos. Nombre en conjunto de datos original: <i>description</i>
Calificación promedio	Calificación promedio del producto dada por usuarios. Nombre en conjunto de datos original: <i>average_rating</i>
Cantidad de calificaciones	Cantidad de calificaciones realizadas por usuarios en la plataforma. Nombre en conjunto de datos original: <i>rating_number</i>
Precio	Precio asignado al producto por parte del publicador. Nombre en conjunto de datos original: <i>Price</i>
Tienda	Tienda asociada a la venta del producto. Amazon se encarga de establecer la plataforma para facilitar el enlace de clientes y vendedores, los vendedores suelen ser empresas que se encargan de la fabricación en masa o la generación de los productos expuestos. Nombre en conjunto de datos original: <i>store</i>
Identificador de producto padre	Dada la posibilidad de que un producto expuesto pueda corresponder a una variante de un SKU genérico, se define un único código padre numérico que permita agrupar productos descartando sus particularidades asociadas a diferencias como color, tamaño, presentación, etc. Nombre en conjunto de datos original: <i>parent_asin</i>
Categorías asociadas	Se asignan subcategorías adicionales al producto para facilitar su clasificación a detalle dentro de la plataforma. Nombre en conjunto de datos original: <i>categories</i>
Detalles	Existe una sección donde se listan los detalles <i>nombrados</i> del producto, es decir, las características específicas del mismo y que no son aplicables de manera general a otros productos. Por ejemplo, el tamaño de memoria RAM solo aplica y se registra para computadores. Nombre en conjunto de datos original: <i>details</i>
Imágenes	El conjunto de datos incluye además información de las imágenes adjuntadas originalmente al producto. Esta

información se presenta como una lista de enlaces a las imágenes asociadas. Nombre en conjunto de datos original: *images*

Tabla 3-1. Campos presentes en entidad de producto en conjunto de datos original.

Las columnas con información de *características, descripción, categorías asociadas e imágenes* se representan mediante listas de textos de tamaño variable. En el caso de las descripciones de producto, la información se encuentra segmentada por párrafos de la descripción original.

3.2.2 Información de reseñas

La información de cada reseña incluye datos de la descripción y título de la reseña en lenguaje natural, así como campos adicionales como la fecha de publicación, la calificación en la escala 0-5 dada por el usuario, la relevancia o utilidad de la reseña en términos de votos de otros usuarios y las imágenes (en caso de haber sido adjuntadas) asociadas a la misma. De manera detallada, se describen las columnas de la entidad reseña en la Tabla 3-2.

Columna	Descripción
Calificación	Calificación numérica en la escala 1-5 de la reseña. Nombre en conjunto de datos original: <i>rating</i>
Título	Título de la reseña. Nombre en conjunto de datos original: <i>title</i>
Texto	Cuerpo o descripción de la reseña en lenguaje natural. Nombre en conjunto de datos original: <i>text</i>
Fecha	Fecha de publicación de la reseña. Nombre en conjunto de datos original: <i>timestamp</i>
Calificación de utilidad de la reseña	Algunas reseñas pueden resultar de utilidad para otros usuarios por diversos motivos, como resaltar algún detalle importante que no se encuentre en las características del producto. Esta utilidad es medida en términos de la cantidad de votos de otros usuarios indicando que la reseña fue útil para el desarrollo de la compra. Nombre en conjunto de datos original: <i>helpful_votes</i>
Identificador de producto padre	Se asocia el identificador del producto sobre el cuál se realizó la reseña. Nombre en conjunto de datos original: <i>parent_asin</i>
Imágenes	Se incluye además información de las imágenes adjuntadas en la reseña. En algunos casos las calificaciones de los usuarios incluyen imágenes del producto real para tanto validar su funcionamiento como para mostrar una falla notoria de manera visual. Esta información se presenta como una lista de enlaces a las imágenes asociadas. Nombre en conjunto de datos original: <i>images</i>

Tabla 3-2. Campos presentes en entidad de reseña en conjunto de datos original.

Se puede evidenciar la relación de las reseñas con los productos y sus características a través del identificador del producto padre. Sin embargo, vale la pena resaltar que las reseñas no se adaptan a una variedad de producto en específico, en su lugar, se asocian únicamente a la publicación descartando opciones de configuración específicas de la venta.

3.3 Diagnóstico de calidad del conjunto de datos

Se evalúan a continuación diferentes aspectos relacionados a la calidad del conjunto de datos, entre ellos campos vacíos, arreglos sin elementos, valores de precios inválidos, entre otros.

3.3.1 Valores nulos y arreglos vacíos

Se realiza una revisión de los valores nulos encontrados en el conjunto de datos para las categorías preseleccionadas enumeradas anteriormente. La evaluación se ejecuta mediante conteos de valores nulos en cada una de las columnas de los dos conjuntos de datos disponibles. Además, se presenta el porcentaje de valores faltantes en una columna con respecto al total de filas en el conjunto de datos correspondiente.

La Figura 3-2 presenta los resultados de este análisis sobre las columnas del conjunto de datos representativo de productos. Se observa en general una baja presencia de valores nulos a excepción del campo precio, donde se evidencia un porcentaje de valores faltantes de hasta aproximadamente el 71%. Lo anterior sugiere la consideración de este campo como candidato a descarte por escasez de información o el planteamiento de un método de llenado adicional, como se discutirá más adelante.

Por otra parte, dado que algunas columnas del conjunto de datos representan información en formato de lista de elementos, se realizó la misma operación anterior calculando la cantidad de registros con listas vacías para cada una de las columnas con este formato, obteniendo así los resultados de la Figura 3-3.

La Figura 3-3 refleja una carencia de hasta el 46.6 % de información para descripciones de producto, así como un 30,7 % para listados de características de producto. En este caso, y al tratarse de un campo relevante para el análisis de características y viabilidad general de productos, se propone como medidas de tratamiento el descarte de los registros que carezcan de elementos en este campo o la consolidación de todos los campos textuales en un único campo que condense la información necesaria. De manera similar, se descartarán productos con subcategorías no diligenciadas, las cuales representan hasta un 7,4 % del conjunto de datos. La carencia de esta información puede justificarse debido a la no obligatoriedad de llenado de este campo durante la creación de una publicación.

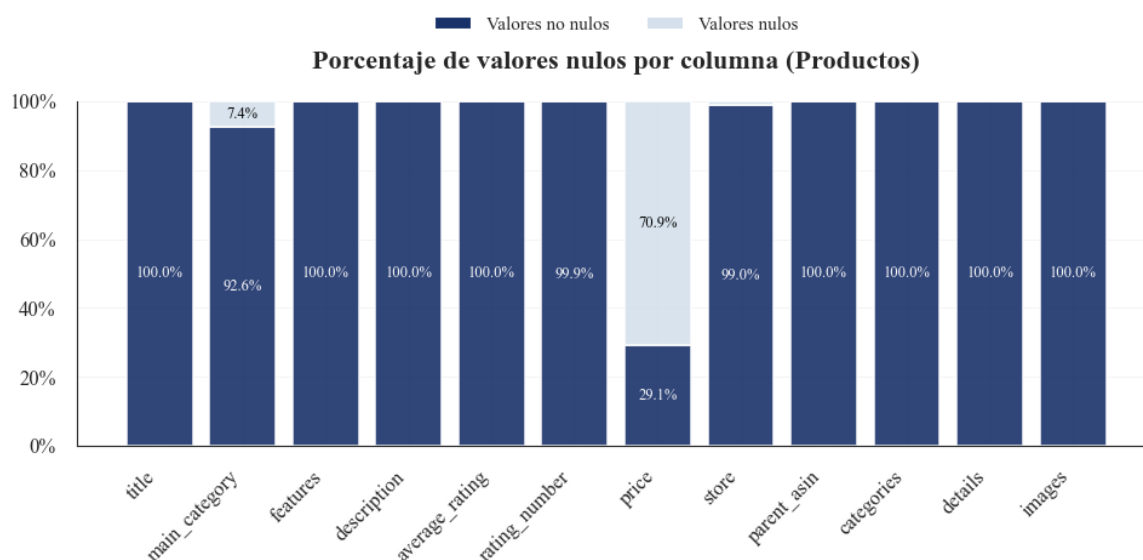


Figura 3-2. Valores nulos en cada una de las columnas del conjunto de datos original para la información de productos.

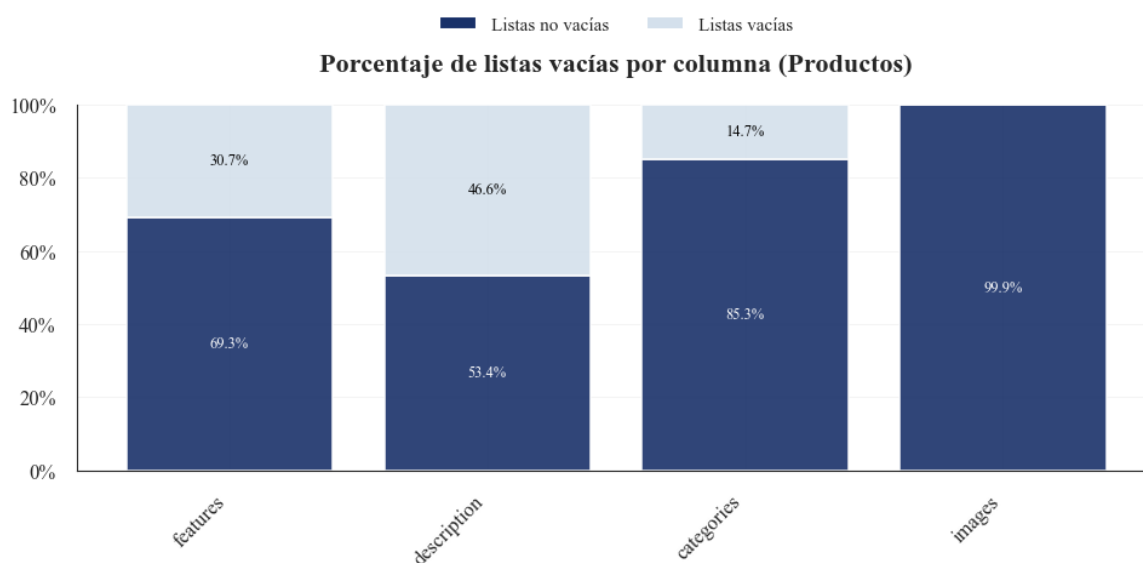
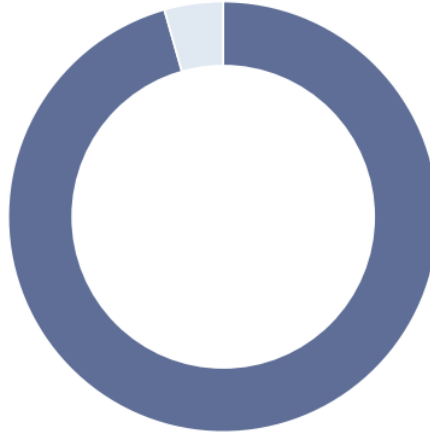


Figura 3-3. Listas vacías para columnas con formato de lista en el conjunto de datos de productos original. La misma operación fue realizada sobre el conjunto de datos de reseñas de producto. Se encontró que ninguna columna posee un porcentaje significativo de valores nulos, siendo inferior a un 0.01% del total de productos agrupados. Así mismo, en la Figura 3-4 se observa que el único campo asociado a un tipo de dato lista, es decir, la lista de imágenes adjuntas a la reseña, posee una alta cantidad de registros sin elementos, llegando a incluir esta información para solamente un 1.43 % del total de reseñas disponibles. Sin embargo, dado que el tratamiento de imágenes no se considerará dentro del alcance inicial del presente trabajo, no se ve necesaria la toma de medidas o tratamientos sobre este atributo del conjunto de datos.

Porcentaje de listas de imágenes vacías (Reseñas)

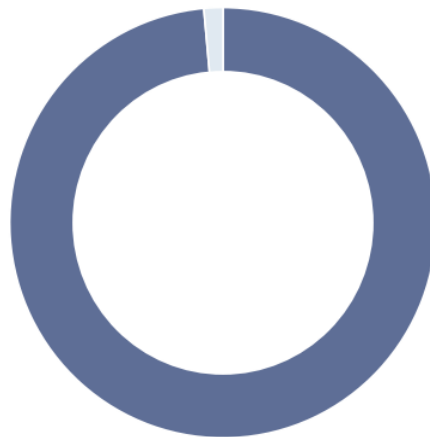
■ Listas vacías (1.432 %) ■ Listas no vacías (98.568 %)

**Figura 3-4.** Listas de imágenes vacías en el conjunto de datos de reseñas.

El campo de detalles de la información de productos alberga el conjunto de atributos asociados a la publicación en formato llave-valor indicando las características puntuales de cada ítem. Se presenta este dato como un documento JSON resumido como cadena de texto. Por tanto, se realizó una deserialización del documento y se extrajeron datos de nombres de características definidas para cada ítem. La Figura 3-5 busca reflejar la presencia de este campo en el conjunto de datos, siendo no especificado en menos del 2% del total de productos disponibles, haciéndolo viable para uso dentro del proceso de homologación de características y permitiendo su introducción en pasos de modelamiento siguientes.

Porcentaje de productos sin detalles especificados

■ Contiene detalles (98.568 %) ■ Sin detalles (1.432 %)

**Figura 3-5.** Productos sin detalles especificados.

Como conclusión de un primer análisis se encuentra una cantidad significativa de valores nulos en las columnas precio y descripción en el conjunto de datos correspondiente a productos. Respecto al precio del producto, se considera la posibilidad de descartar esta columna dado que, si bien puede contribuir al análisis y modelamiento, no representa un elemento crucial para este trabajo.

Por otra parte, resulta vital la inclusión de campos textuales no vacíos, dado que estos conforman la fuente de información principal para la inferencia de características y valoración positiva o negativa en reseñas. Por tanto, se propone el descarte de productos y reseñas que no posean en conjunto información de características o descripciones, ya que carecen de los datos esenciales para los propósitos del estudio. Sin embargo, para la información de imágenes y videos de producto no se encuentran las mismas conclusiones, ya que, si bien pueden ser útiles para la aumentación de datos y la extracción de información adicional, no son vitales para los propósitos de este trabajo, motivo por el cual no se hace necesaria su eliminación como columna de información ni el descarte de registros que no posean este dato.

3.4 Análisis exploratorio de los datos

Se realiza una revisión de las relaciones entre variables para los conjuntos de datos importados. En esta sección se revisa la distribución de tamaños de textos y subcategorías de producto viables para pasos posteriores de modelamiento, así como la relación de las mismas con las apreciaciones de los usuarios transmitidas a partir de reseñas.

3.4.1 Exploración de campos textuales

Vale la pena realizar una revisión sobre la distribución de los tamaños de textos en términos de cantidad de palabras. Este análisis busca estimar de manera preliminar la densidad de contenido textual presente en diferentes campos de este tipo. Cabe mencionar que algunos campos fueron unificados para condensar textos representados por múltiples párrafos o enumeraciones.

La Figura 3-6 presenta la distribución logarítmica de la cantidad de palabras en descripciones de producto. Para realizar esta operación se concatenaron todos los párrafos asociados a la descripción del ítem y se realizó un proceso de tokenización. Se puede observar una alta frecuencia de valores alrededor de las 100 palabras por descripción, esto indica un tamaño de texto significativo que puede ser útil para la extracción de características. Sin embargo, también puede considerarse la realización de un filtrado preliminar sobre los productos con descripciones relativamente cortas, por ejemplo, de menos de 20 palabras, todo ello a fin de conservar los ítems que presentan una estructura suficientemente extensa. Cabe mencionar que al proceso de filtrado se suma un paso de limpieza general que involucra la eliminación de caracteres especiales y la tokenización de ciertos elementos textuales, incluyendo ciertos componentes no verbales encontrados, principalmente símbolos asociados a emojis.

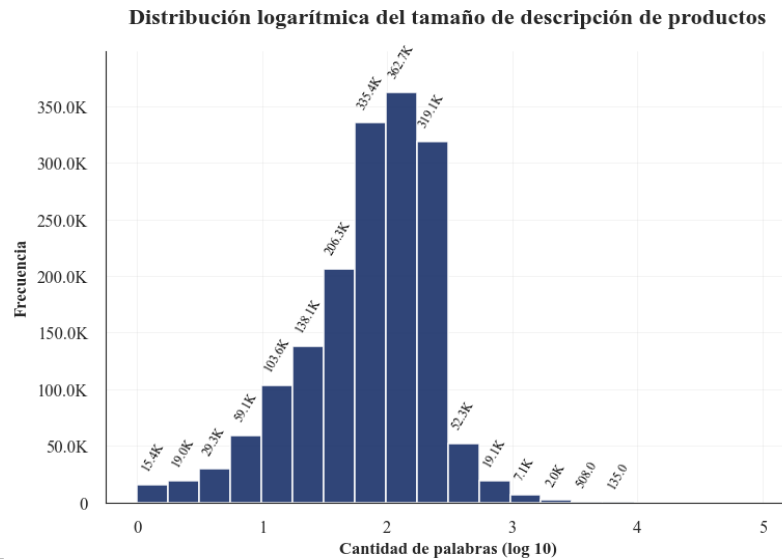


Figura 3-6. Distribución de cantidades de palabras en descripciones de productos.

De manera equivalente, las Figura 3-7 y Figura 3-8 presentan el mismo análisis para los campos de título y características de la publicación, respectivamente. El comportamiento de la distribución de contenido textual en características resulta similar al encontrado en la Figura 3-6. Lo anterior indica una paridad entre los campos descripción y características, incluso, se considera la posibilidad de complementar los valores faltantes en descripción con la información de características disponible.

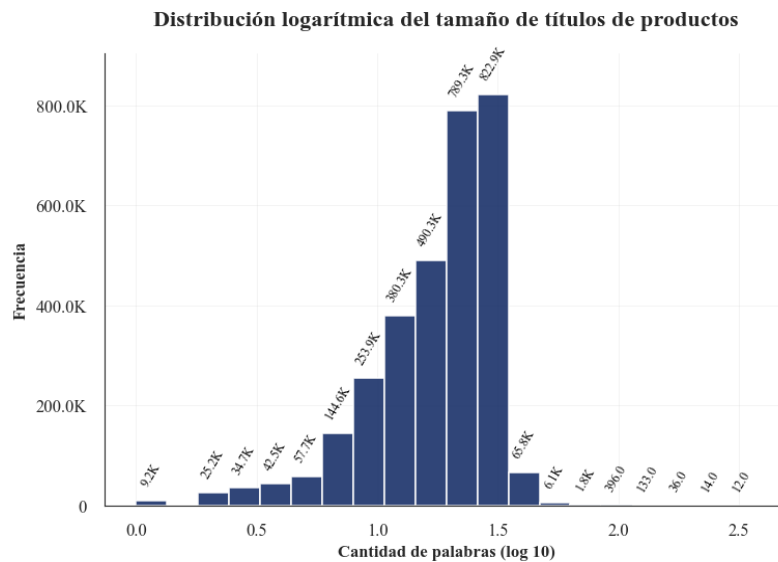


Figura 3-7. Distribución de cantidades de palabras en títulos de productos

Por otra parte, la Figura 3-7 señala una baja cantidad de tokens en los títulos de publicaciones, lo cual se considera un comportamiento esperado al tratarse de un contenido corto usualmente conformado por una sola frase. En este último caso se evidencia un tamaño de texto de entre alrededor de 10 y 35 palabras.

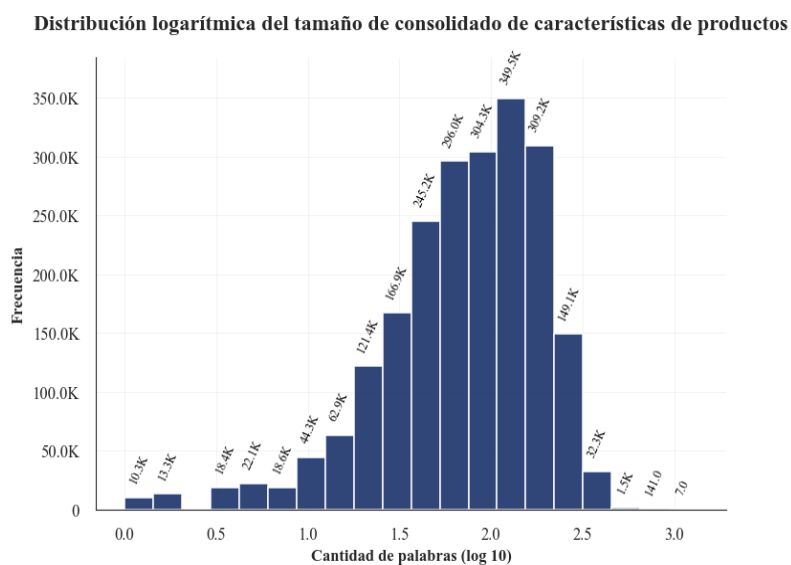


Figura 3-8. Distribución de cantidad de palabras en campos de características de productos.

Como se mencionó anteriormente, vale la pena considerar la unificación de campos textuales en un único campo consolidado que albergue cada uno de los elementos que componen las características del producto, de esta manera, se propone la concatenación del título, descripción y características especificadas del producto. Así mismo, sobre este campo consolidado se realizó un análisis extendido sobre cantidad de palabras y relación Type-token ratio.

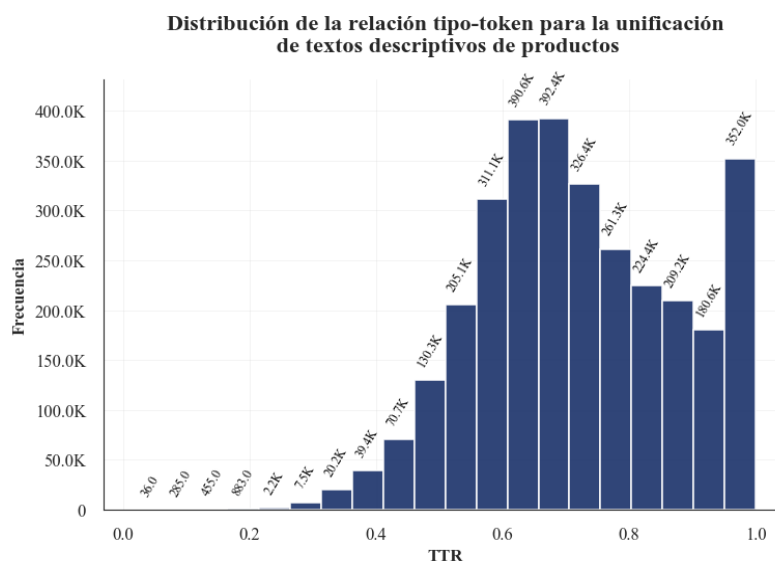


Figura 3-9. Distribución de TTR (Token-type ratio) para textos unificados con información de producto. La Figura 3-9 presenta la distribución de la relación tipo-token para el consolidado resultado de la concatenación de título, descripción y características para cada producto dado. La relación TTR o Type-Token Ratio se describe en la ecuación (6).

$$TTR = \frac{\text{Cantidad de tokens \u00fanicos}}{\text{Cantidad de tokens totales}} \quad (6)$$

Se encuentra que la mayor\u00eda de productos poseen informaci\u00f3n textual con un TTR de entre el 60% al 100%, lo cual indica una riqueza textual significativa, al presentar una relativa baja repetic\u00edon de palabras y, por tanto, una estimable baja redundancia en componentes textuales. Sin embargo, es importante retomar la consideraci\u00f3n de la longitud de textos y los faltantes encontrados en la Figura 3-3, pues ya se ha discutido la longitud relativamente corta de campos como t\u00edtulos de publicaci\u00f3n, la cual puede contribuir a la generaci\u00f3n de tokens \u00fanicos dentro de textos cortos. A continuaci\u00f3n, se relacionan algunos ejemplos de casos como este, extra\u00eddos directamente del conjunto de datos original:

- Jamie Dornan 2017 (English, French and German Edition)
- Sakar Compact Deluxe Gadget Bag - DC74

En este sentido, cobra relevancia la limpieza de registros con una cantidad de tokens poco significativa, es decir, el descarte de productos con una cantidad de tokens por debajo de un determinado l\u00edmite.

Con relaci\u00f3n a la informaci\u00f3n textual presente en rese\u00f1as, se realiz\u00f3 un an\u00e1lisis exploratorio similar al ejecutado para la informaci\u00f3n de productos. La Figura 3-10 presenta el resultado de este an\u00e1lisis, encontrando que la mayor\u00eda de los textos se encuentran en un rango de alrededor de 23 a 57 tokens, lo cual es significativamente m\u00e1s corto a comparaci\u00f3n con las descripciones de productos en general.

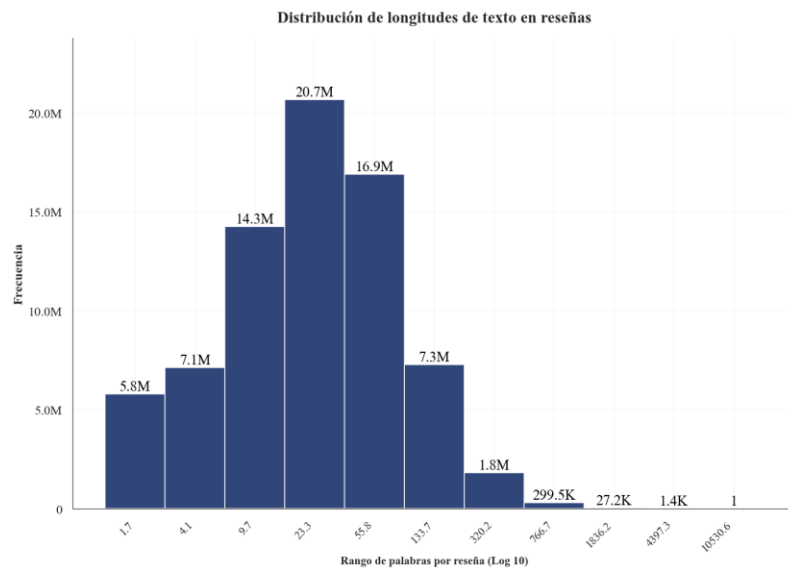


Figura 3-10. Distribuci\u00f3n de largos textuales para informaci\u00f3n de rese\u00f1as

3.4.2 Exploraci\u00f3n de esquemas en detalles de publicaciones

El campo de detalles en las publicaciones de productos ofrece informaci\u00f3n que puede resultar relevante para la extracci\u00f3n de caracter\u00edsticas de productos. Se trata de informaci\u00f3n agrupada en formato clave-valor que describe caracter\u00edsticas espec\u00edficas del producto. En la

Figura 3-11, se realiza una revisión sobre los esquemas disponibles y la frecuencia de aparición de ciertos campos, ya que estos varían entre productos diferentes.

Distribución de campos detalle en productos

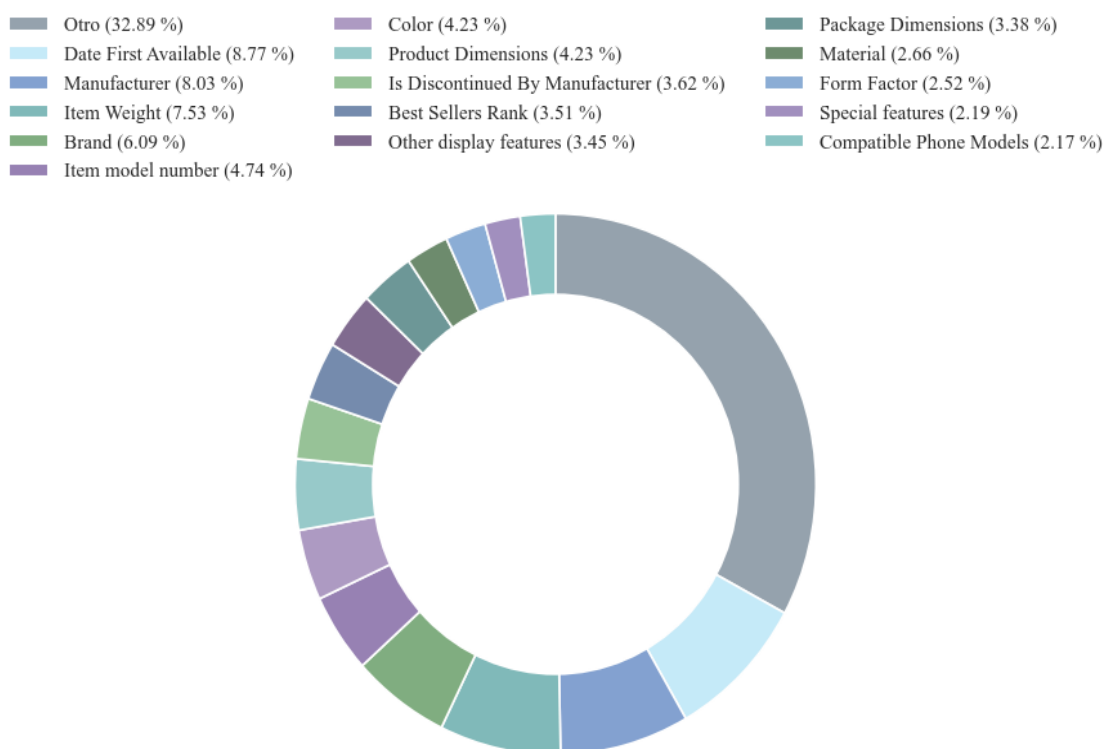


Figura 3-11. Distribución de campos detalle en descripciones de producto.

Se observa una amplia variedad de características proveídas para productos en general, sin embargo, existe una notable presencia de campos como fecha de primera publicación, compañía que manufactura, y modelo del ítem. Así mismo, se encuentran elementos usuales como color, peso, dimensiones, entre otros. Cabe mencionar que cada categoría de producto provee sus propias características específicas y, en algunos casos, únicas para la agrupación. Sin embargo, existen algunas características que carecen de relevancia para los propósitos de este trabajo, por ejemplo, no es de interés la fecha de publicación, tienda que publica o el estado de continuidad de su fabricación, razón por la cual no se espera considerar estos campos en posteriores pasos.

La Figura 3-12 logra reflejar consistencia entre las características esperadas de productos con las categorías a las que pertenecen. A manera de ejemplo, se presenta información de dispositivos compatibles para computadores, así como atributos de pantalla para celulares. Sin embargo, y como se mencionó anteriormente, existen algunas características compartidas entre categorías que no hacen parte relevante del estudio que se busca realizar, motivo por el cual se considerará su descarte.

Participación de detalles de producto por categoría principal

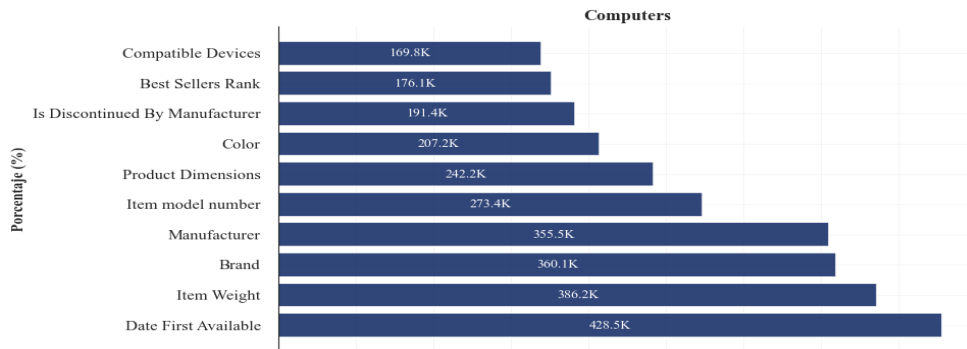
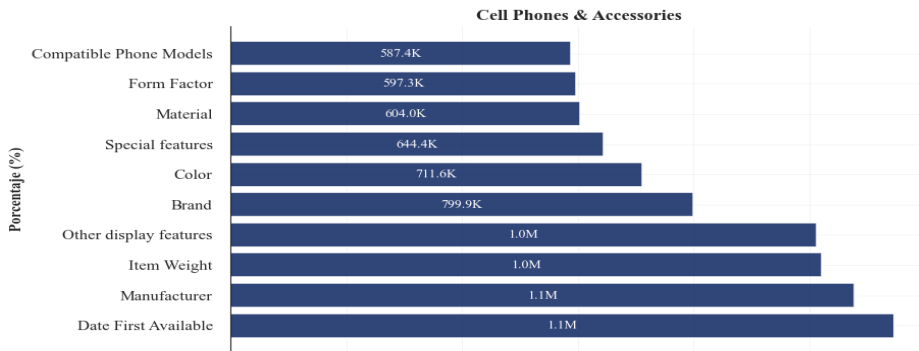
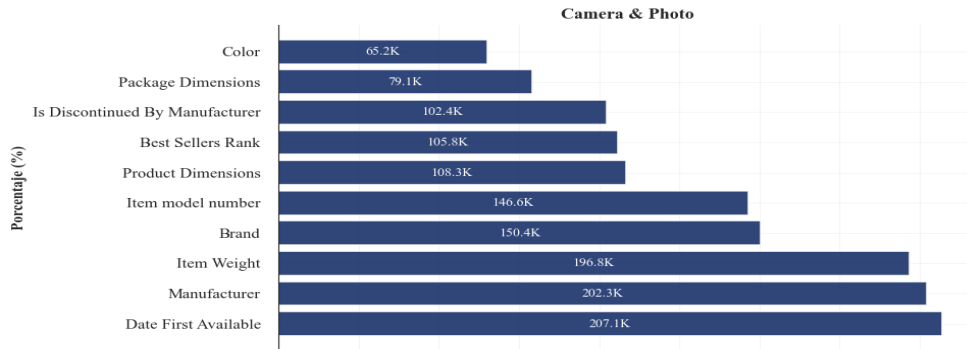
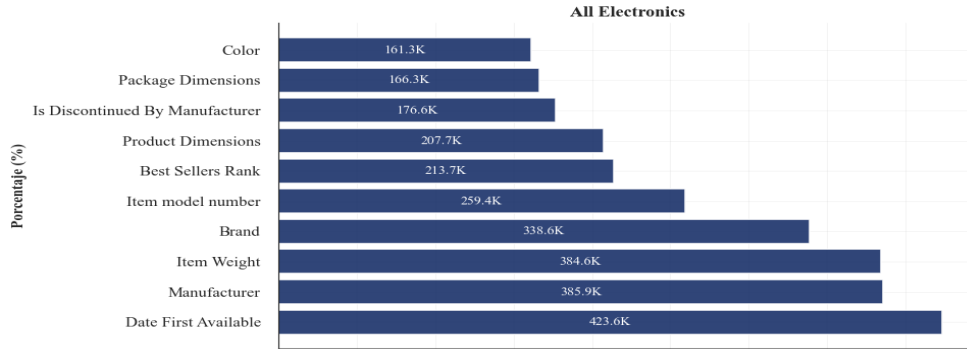


Figura 3-12. Muestra de campos de características diligenciados en detalle de productos para las 4 subcategorías más frecuentes.

3.4.3 Distribución de calificaciones de usuarios

Las calificaciones de usuarios representan las valoraciones numéricas discretas en la escala entera 0-5 dadas en la plataforma Amazon, y cuyo propósito es reflejar la satisfacción del usuario con un determinado producto comercializado a través de la plataforma. Se trata de un estándar implementado en tiendas en línea y usualmente permite a otros usuarios tener una vista preliminar de las experiencias de otros compradores con respecto a un producto.

La calificación de un usuario puede reflejar la intención de una reseña y la relación de la misma con los aspectos positivos y negativos del producto. La Figura 3-13 presenta la distribución general de todas las calificaciones en el rango 0-5 de las reseñas asociadas a los productos pre-seleccionados. Se evidencia de esta manera una predominancia de la calificación perfecta o cinco estrellas con alrededor del 64,5 % de participación.

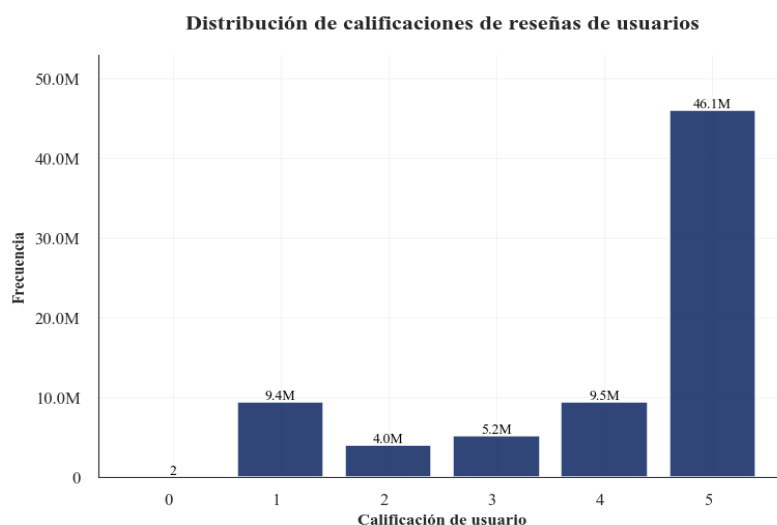


Figura 3-13. Distribución general de calificaciones de usuario en reseñas del conjunto de datos.

Se encuentra en la Figura 3-13 la predominancia de la calificación de usuario máxima sobre el total de reseñas disponibles. Para este caso, se considera necesaria la conservación de todas las calificaciones de usuario disponibles. Sin embargo, para etapas de modelamiento se contempla la estandarización de calificaciones únicamente en dos categorías según un determinado límite (tentativamente la calificación 3 o 4), donde se puedan caracterizar reseñas por encima de este límite como positivas o satisfactorias para el usuario, así como reseñas por debajo del mismo como calificaciones negativas o poco satisfactorias.

3.4.4 Distribución de cantidad de reseñas por producto

Se ha analizado hasta este punto el contenido de información textual en lenguaje natural tanto en reseñas como productos, sin embargo, vale la pena realizar una exploración sobre

la cantidad de reseñas presentes por producto dada la influencia que este aspecto podría tener sobre el entrenamiento de modelos de aprendizaje de máquina y el posible sesgo que pueda introducirse hacia determinadas categorías o productos con una mayor cantidad de reseñas en promedio.

La Figura 3-14 presenta la distribución de cantidades de reseñas por producto. Se evidencia que una porción de alrededor del 33% de productos totales no cuenta con una cantidad de reseñas significativa, variando principalmente en el rango de 1 a alrededor de 10 reseñas. Lo anterior refleja que es necesario posicionar una estrategia de preprocesamiento que permita el descarte de productos con un número de reseñas por debajo de un valor aceptable, en este caso se considera 5 reseñas como un valor coherente con los propósitos del trabajo, ya que se puede contar con información suficiente de calificaciones para satisfacer el análisis de preferencias de clientes para productos individuales.

De manera similar al caso de la información de productos, se concatenará la información de título con la descripción de la reseña a fin de aprovechar y enriquecer el contenido con el que se cuenta en este conjunto.

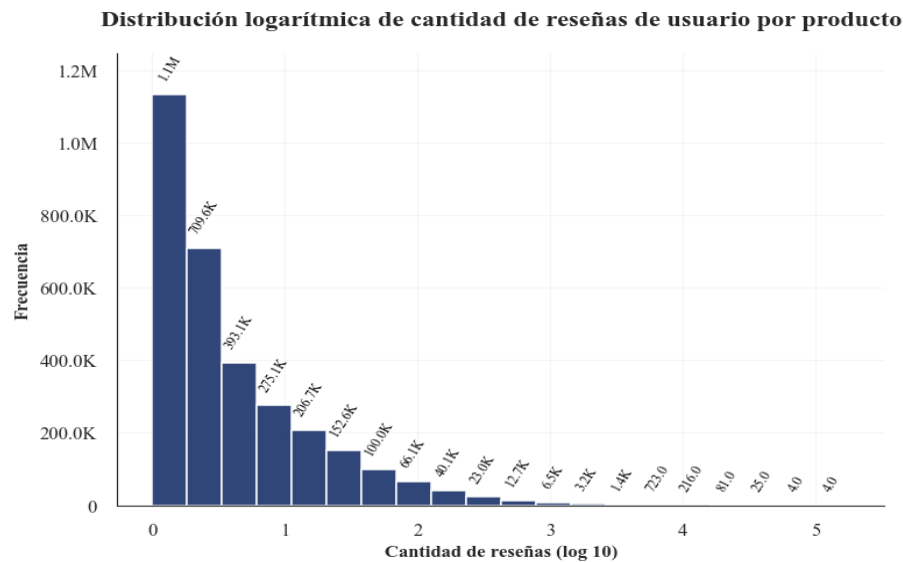


Figura 3-14. Distribución logarítmica de cantidad de reseñas de usuario por producto.

3.1.1 Distribución de subcategorías de producto

La Figura 3-15 revela que la mayoría de productos disponibles hacen parte de la categoría de *Celulares y accesorios*. Esta categoría llega a representar un 41% del total de registros en el conjunto de datos, convirtiéndose de esta manera en uno de los candidatos con mayor relevancia en la definición de criterios de selección final.

Por otra parte, resulta valiosa la selección de subcategorías que provean cierto nivel de generalidad para productos que puedan ser comercializados dentro de un segmento de clientes relativamente amplio. Por ejemplo, las aplicaciones de teléfono y en general productos de Software se consideran elementos personalizados que suelen asociarse a tecnologías

ampliamente conocidas, pero cuyo propósito no se ve especificado en las descripciones de una licencia comerciada en línea. Usualmente, el verdadero propósito y la explicación principal de su funcionamiento se encuentra en el software mismo o en el contexto general del cliente al que se busca llegar. Así mismo, los nombres de los productos no llegan a ser explicativos en relación a su propósito, dificultando de esta manera las oportunidades de inferencia y generalidad en el modelo final que se busca implementar.

Además, se observa que existen ciertas subcategorías de producto que se vieron introducidas probablemente de manera incorrecta a las categorías preseleccionadas, entre ellas, la más notable es *Amazon Fashion*, categoría asociada a productos de belleza.

En este sentido, se considera pertinente la selección de subcategorías fijas a conservar dentro del conjunto de datos final, como se discutirá en una próxima sección del presente trabajo.

Top 15 Subcategorías de producto disponibles

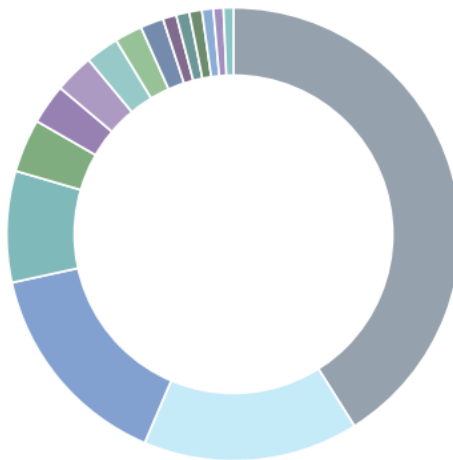
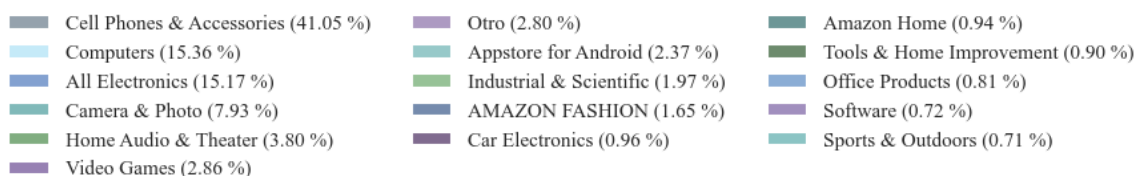


Figura 3-15. Distribución de subcategorías de producto.

3.4.5 Relación de variables numéricas en reseñas de producto

Las reseñas de producto con una cantidad de texto amplia pueden ser asociadas a una reseña más completa y significativa para el lector, siendo este hecho extensible para la mayoría de textos de opinión en general. En esta sección, se analizará la correlación entre variables asociadas a indicadores numéricos representativos de reseñas.

La Figura 3-16 presenta los resultados encontrados para un análisis de correlación lineal entre las variables:

- Votos a reseña por utilidad hacia otros usuarios
- Cantidad de tokens en texto de reseña
- Calificación dada al producto asociado en la reseña

Se encuentra de esta manera como aspecto relevante la presencia de una ligera relación lineal entre la cantidad de palabras en la reseña y la cantidad de votos de utilidad sobre la misma. Esto se sustenta bajo los argumentos ya mencionados al inicio de la presente sección, pues es común calificar textos altamente explicativos (y por tanto más extensos) como útiles dado su alto nivel de detalle.

Asociando lo anterior con los propósitos de este trabajo, es importante notar que, de la misma manera, los textos con un tamaño considerable ofrecen una riqueza superior en términos de contenido semántico y características extraíbles.

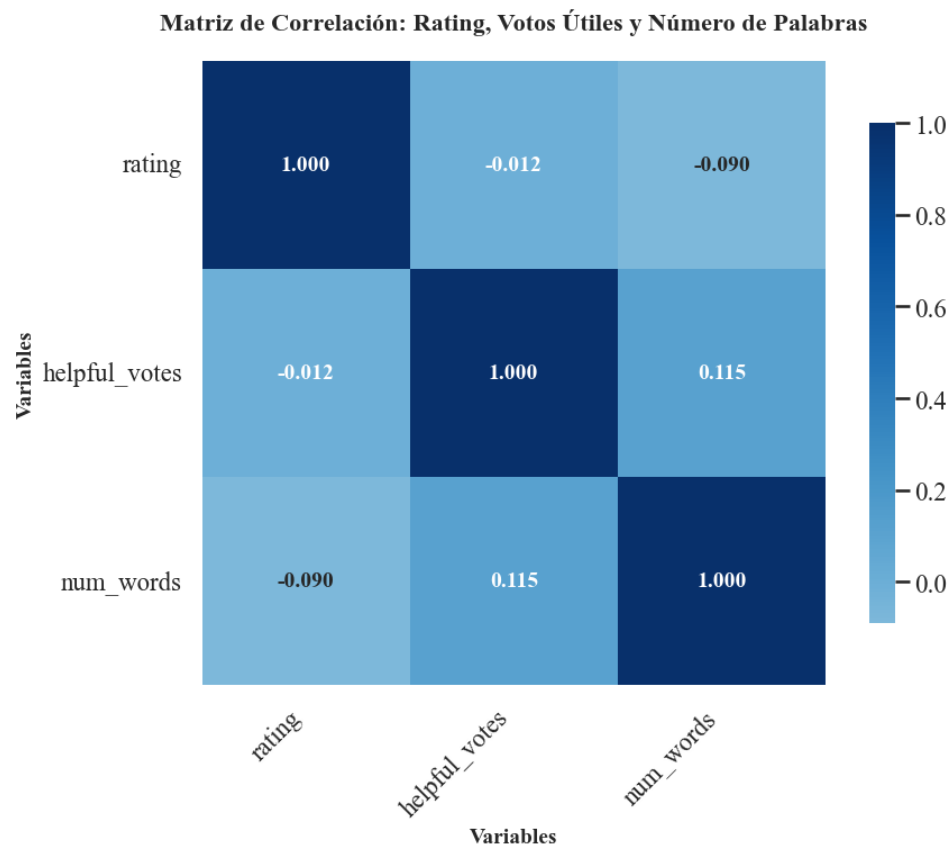


Figura 3-16. Correlación de variables: Calificación, votos útiles y número de palabras.

3.5 Selección de variables relevantes

Tras la realización de un análisis exploratorio sobre los datos, se encuentra que algunas variables no pueden ser tenidas en cuenta dentro del proceso de modelado debido a la

carencia de valor debido tanto a una baja presencia de valores no vacíos en el conjunto de datos, como la poca relevancia en relación al propósito de este trabajo.

Para el caso de información de productos, y como se puede evidenciar en la Figura 3-2, el precio del ítem no puede ser tenido en cuenta debido a la carencia del 71% de datos con respecto al tamaño general del conjunto.

Así mismo, existen ciertas columnas cuya información no resulta relevante para el estudio, como la tienda que vende el producto. La Tabla 3-3 relata la indagación realizada sobre cada uno de los campos del conjunto de datos original.

Conjunto	Columna	Incluida	Justificación
Producto	Título	Sí	El título de la publicación ciertamente aporta información breve relevante para el estudio, como se discutió en la sección de análisis de textos.
Producto	Categoría principal	Sí	Se ha analizado este campo bajo el alias de <i>subcategoría</i> . En este sentido, se considera relevante dado que permite caracterizar el dominio del producto y reflejar los criterios de selección previamente concebidos.
Producto	Características	Sí	La lista de funcionalidades del producto permite caracterizar de manera precisa las capacidades que se tienen para la resolución de una necesidad. Este campo es vital en la medida en la cual hace parte de la identidad del ítem. Así mismo, será transformado para condensarse en un componente textual.
Producto	Descripción	Sí	Al igual que la sección de características, presenta información necesaria para caracterizar al producto dentro de un contexto determinado.
Producto	Calificación promedio	Sí	A pesar de ser posible el cálculo de este campo en función de las reseñas dadas, se considera útil su inclusión como un atributo adicional que puede contribuir a la vista general y atómica del producto.
Producto	Cantidad de calificaciones	Sí	Al igual que el campo anterior, la cantidad de calificaciones previamente consolidada puede contribuir a la estimación de relevancia de publicaciones en función de la cantidad de reseñas que se poseen.
Producto	Precio	No	Como se discutió anteriormente, el precio no será considerado debido a la alta tasa de valores nulos presentes en el conjunto de datos.
Producto	Tienda	No	No se considera relevante la tienda que comercia el producto, ya que se busca generalizar la información del mismo dentro de la interacción del comercio en línea.
Producto	Identificador de producto padre	Sí	Se requiere el identificador del producto padre para persistir la relación entre reseñas y productos.

Producto	Categorías asociadas	No	Las categorías asociadas en el tercer nivel de jerarquía de clasificación del producto extienden el alcance de la primera categoría, sin embargo, se consideran un elemento redundante dado que la descripción y características del producto ya hacen referencia a sus especificaciones y dominio.
Producto	Detalles	No	Los detalles del producto pueden transformarse en atributos generales que contribuyen a la caracterización del producto. Sin embargo, se decide dentro de un primer alcance no considerar su inclusión dada la complejidad de extracción de atributos numéricos y la existencia de características textuales que ya reflejan atributos específicos del producto.
Producto	Imágenes	No	No se procesará información de imágenes o contenido multimedia en un primer alcance del presente trabajo.
Reseña	Datos textuales unificados	Sí	Dentro del flujo de preprocesamiento, se condensa la información textual del producto, específicamente de los campos: Título, descripción y características.
Reseña	Calificación	Sí	La calificación de la reseña representa información relevante para el análisis que se busca llevar a cabo. Permite la asociación en una primera instancia de la satisfacción del usuario hacia el producto en cuestión.
Reseña	Título	Sí	El título de la reseña aporta una cantidad de información útil para la identificación preliminar de su propósito. Cabe resaltar que su información se extiende de manera amplia en el campo de texto asociado a su descripción.
Reseña	Texto	Sí	Presenta información adicional de la reseña de manera análoga a la descripción en la información de productos.
Reseña	Fecha	No	No se requiere de conocimiento sobre la información de la fecha de publicación de la reseña dada la naturaleza del trabajo.
Reseña	Calificación de utilidad de la reseña	Sí	La utilidad de la reseña puede permitir la asignación de un peso superior sobre el resultado final a aquellas reseñas que representan más valía para los usuarios del producto analizado.
Reseña	Identificador de producto padre	Sí	Se requiere la identificación del producto asociado con fines de conservación de la relación de las entidades desde el conjunto de datos original.
Reseña	Imágenes	No	No se procesará información multimedia asociada a la reseña. Además, en el caso de reseñas no se cuenta con imágenes en la mayoría de los registros.

Tabla 3-3. Resumen de inclusión y exclusión de campos de conjunto original según análisis realizado.

3.6 Limpieza y preprocesamiento

Anteriormente se discutió la necesidad de tratar campos faltantes y listas de componentes textuales sin elementos. A continuación, se detalla el flujo de preprocesamiento propuesto para tratar apropiadamente cada una de las anotaciones realizadas. La Figura 3-17 presenta de manera gráfica este flujo.

Inicialmente, se considerarán únicamente las subcategorías de producto analizadas en una sección anterior del presente documento, seleccionando así las subcategorías presentadas en la Tabla 3-4.

Subcategorías seleccionadas		
Cell Phones & Accessories	All Electronics	Home Audio & Theater
Computers	Camera & Photo	Car Electronics
Amazon Home	Industrial & Scientific	Tools & home improvement
Office Products	Sports & outdoors	

Tabla 3-4. Subcategorías seleccionadas para análisis.

Posteriormente, se descartaron los productos cuya información textual no incluye descripciones de producto o información de características, ya que la información proveída por títulos únicamente no representa una cantidad significativa de contenido. Para esto, se remueven registros con menos de 50 tokens tras la consolidación de los tres elementos anteriores.

De manera similar, se descartan productos con menos de 5 reseñas asociadas, ya que la carencia de información de reseñas para un producto puede llegar a inutilizar su presencia en el conjunto de datos al no existir un punto de comparación que permita encontrar relaciones entre sus características y la percepción del usuario.

Finalizando con el procesamiento de productos, se segmentan las oraciones pertenecientes a los componentes textuales incluidos dentro de los campos seleccionados para esta entidad. En este caso, se usa una expresión regular ([.!?]) para facilitar este proceso y proporcionando balance al desempeño de procesamiento con los recursos disponibles.

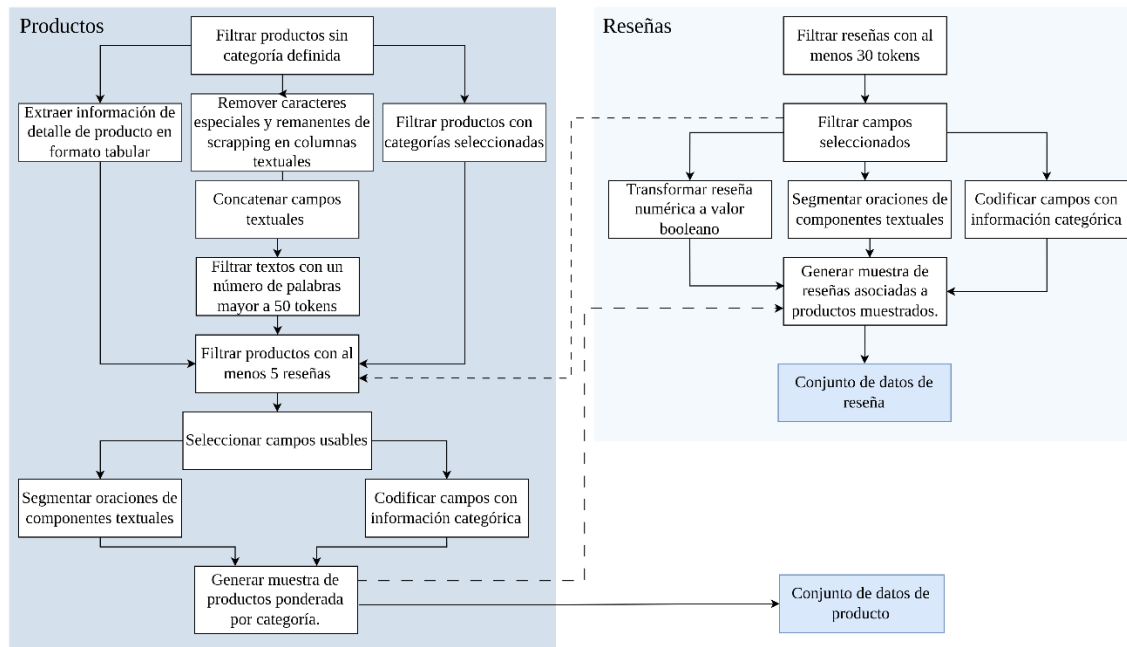


Figura 3-17. Descripción gráfica de flujo de preprocesamiento.

Para el caso de reseñas, se realiza una operación similar, se descartarán reseñas con una cantidad de tokens poco significativa, en este caso, menor a 30 tokens dada la distribución general de la longitud de textos en este conjunto de datos visualizada en la Figura 3-10. Se realiza además un proceso de segmentación de oraciones en descripciones de reseñas usando la misma metodología previamente descrita para productos. Dados los resultados de la Figura 3-13, se agrega además una columna de información para presentar la reseña numérica transformada a valores booleanos codificados como valores 0 y 1.

Posteriormente, para reseñas y productos, se realiza un proceso de selección de columnas, siguiendo la justificación dada en la Tabla 3-3, a fin de reducir la cantidad de información a procesar y simplificar pasos posteriores del presente trabajo. Se realiza además un proceso de codificación de columnas categóricas mediante vectores one-hot. En este caso, estas columnas se reducen a subcategorías de segundo y tercer nivel para información de productos.

Finalmente, y a fin de facilitar el proceso de entrenamiento y pruebas sobre un conjunto de datos de tamaño manejable por los recursos computacionales disponibles, se generará una muestra de productos y reseñas a usar como conjunto de entrenamiento para el próximo entrenamiento de modelos de aprendizaje de máquina. El muestreo se generará mediante un método aleatorio estratificado ponderado por el inverso de la frecuencia de ocurrencia de cada categoría presentada en la Figura 3-15.

3.7 Conclusiones de capítulo

Se evidenció que, pese a la popularidad y madurez del comercio electrónico, la información de productos publicada en la plataforma presenta una cantidad considerable de campos

vacíos o incompletos, incluso en atributos críticos como el precio. Tras una revisión de fuentes de información alternativas, incluyendo la validación en la plataforma misma, se encontró que esta ausencia puede atribuirse a la naturaleza dinámica de los precios y la ausencia del campo para productos discontinuados.

Adicionalmente, se observó que la actividad del comercio en línea se concentra de manera significativa en categorías específicas, destacándose *celulares* y *accesorios* como la principal, al representar más de un tercio del total de productos tecnológicos analizados. Esta concentración no solo refleja una alta demanda por parte de los usuarios, sino que también convierte a esta categoría en un foco prioritario para el análisis, al disponer de un mayor volumen de información y, por ende, un mayor potencial explicativo.

Asimismo, el análisis permitió identificar una relación directa entre la extensión del contenido textual de las reseñas y su utilidad percibida por otros usuarios. En particular, una mayor riqueza léxica se asocia con niveles más altos de utilidad, entendida como el grado de acuerdo colectivo de otros usuarios respecto a una percepción u observación sobre un producto.

Finalmente, estos hallazgos orientan los pasos posteriores del trabajo, especialmente en la definición de métricas derivadas del conjunto de datos. En este sentido, variables como la cantidad de votos de utilidad de una reseña resultan fundamentales, ya que representan un consenso implícito entre múltiples usuarios y aportan una señal relevante para inferir necesidades y niveles de satisfacción. De manera complementaria, los productos con mayor completitud y detalle de información presentan una mayor relevancia analítica, al facilitar la asociación con nuevas entradas y mejorar la calidad de los procesos de planteamiento de métricas y modelamiento.

4. Métricas para medir la satisfacción del consumidor

En este capítulo se recopilan métricas candidatas derivadas de trabajos previos y se plantean adaptaciones de estas en función de la estructura de la información documentada en el capítulo anterior. Asimismo, se presentan interpretaciones numéricas formuladas que permiten la definición de variables objetivo que servirán como base para la construcción del indicador final de viabilidad del producto. Se propusieron 2 métricas basadas en la calificación numérica de reseñas con diferentes especificaciones respecto a estructura de los datos de entrada.

4.1 Recopilación de métricas candidato

4.1.1 Métrica basada en metodología SHELL

Como se describe en [1], SHELL es una metodología basada en la revisión de causas y riesgos de fracasos en empresas emergentes, basada principalmente en apreciaciones expertas. Dichas apreciaciones se estructuran en cinco componentes fundamentales: *Planteamiento de modelo de negocio, contexto de la organización, entorno o competencia, características del producto y características de los consumidores.*

Cada uno de estos componentes recibe una valoración porcentual que posteriormente es agregada para obtener un resultado numérico final. Adicionalmente, el trabajo citado emplea técnicas de agrupamiento para analizar la correlación entre las diferentes causas de fracaso asociadas a los distintos componentes del modelo.

Cabe destacar que esta métrica depende de la generación de valoraciones subjetivas por parte de un público experto. No obstante, es posible plantear una estrategia alternativa que aproveche la estructura de la información disponible en el presente trabajo, particularmente en lo referente a la evaluación cuantitativa del entorno competitivo, mediante el análisis comparativo de métricas observables en productos similares.

4.1.2 Métrica basada en metodología JTBD

El trabajo realizado por [5] presenta una metodología más rigurosa respecto al análisis de viabilidad de productos, centrada no en la prevención del fracaso, sino en la identificación y satisfacción de necesidades reales de clientes potenciales.

La metodología Jobs-To-Be-Done (JTBD) analiza las *funcionalidades* ofrecidas por un producto en relación con las acciones o “jobs” que los usuarios esperan realizar,

incorporando además un marco contextual asociado a los segmentos de mercado objetivo y a sus características particulares.

El proceso inicia con la identificación de los segmentos de mercado y de los trabajos que los usuarios buscan ejecutar dentro de un contexto de uso específico. En etapas posteriores, se evalúa el grado de cumplimiento de cada trabajo a partir de las características de productos concretos. Estas evaluaciones se realizan desde la perspectiva de un grupo especializado, que asigna puntuaciones enteras en una escala de 1 a 5. Finalmente, como se formaliza en las ecuaciones (1) y (2), la información recolectada se consolida mediante un promedio ponderado que considera tanto el tamaño del segmento de mercado como la puntuación obtenida para cada trabajo identificado.

4.1.3 Métrica basada en puntuación de usuario

Como describe [19], la satisfacción del consumidor puede estimarse a partir de la calificación numérica expresada en términos de estrellas, en una escala discreta de 0 a 5, asociada a las reseñas de producto. La construcción del indicador propuesto se fundamenta principalmente en las ecuaciones (3), (4) y (5), las cuales incorporan factores adicionales relacionados con la relevancia y efectividad de las reseñas.

Si bien el estudio original se basa exclusivamente en atributos numéricos derivados de las valoraciones de usuarios, en el presente trabajo se propone una adaptación que aproveche la estructura del conjunto de datos analizado, incorporando de manera explícita los elementos textuales asociados tanto a los productos como a reseñas.

En síntesis, se plantea el uso de la puntuación numérica ponderada del conjunto de reseñas asociadas a un producto como una métrica directa para estimar la satisfacción del consumidor, sirviendo como uno de los componentes fundamentales en la definición del indicador final de viabilidad del producto.

4.2 Construcción y definición de métricas candidato

Los modelos de representación semántica, como el Universal Sentence Encoder (USE) [10], permiten codificar información semántica presente en oraciones textuales expresadas en lenguaje natural. Estas codificaciones facilitan la comparación y asociación entre textos de manera computacional a partir de métodos numéricos basados en el procesamiento de vectores de dimensión fija.

Como se describe en la Figura 2-3, la codificación de elementos textuales mediante USE produce vectores numéricos de tamaño constante. A partir de estas representaciones, se define la similitud semántica entre dos elementos textuales, en este caso oraciones completas codificadas, como $s(a, b)$.

De manera habitual, la similitud del coseno se emplea como métrica para cuantificar la similitud entre representaciones vectoriales obtenidas a partir de codificaciones semánticas. La definición formal de esta medida se presenta en la ecuación (7).

$$s(a, b) = \frac{a \cdot b}{\|a\| \|b\|} \quad (7)$$

Las métricas propuestas en la sección anterior tienen como objetivo condensar información proveniente de productos o lanzamientos de productos, con el fin de generar indicadores que faciliten su análisis comparativo. En el caso de la metodología SHELL, dichas métricas se orientan a la estimación de factores de riesgo y posibles causas de fracaso; mientras que, en el caso de la metodología JTBD, se enfocan en la estimación de la viabilidad del producto a partir de la solvencia de necesidades identificadas en uno o varios sectores de mercado.

Desde la perspectiva de la metodología SHELL, resulta viable el análisis de dos componentes específicos, correspondientes a la competencia (Entorno) y características del producto, dado que no se dispone de información contextual relacionada con el entorno financiero de la categoría del producto, la situación interna de la organización fabricante o las estrategias de mercadeo más allá de la publicación en la plataforma de comercio seleccionada.

En relación con la formulación numérica de esta métrica, su estimación puede realizarse mediante la combinación de la definición de similitud semántica y la información disponible tanto de productos como de reseñas, como se describe a continuación.

Como punto de partida común para todas las métricas propuestas, se genera una representación numérica de cada entidad mediante una codificación basada en USE, complementada con un esquema de ponderación de los distintos componentes textuales presentes en la estructura de datos. En particular, y de acuerdo con las definiciones estudiadas en [5], se establece una ponderación específica para los componentes textuales asociados a la entidad de producto. La Tabla 4-1 presenta la definición de estas ponderaciones, asignando un peso ligeramente mayor al componente de características, dada su relación directa con las actividades descritas en la metodología JTBD.

Componente textual	Ponderación
Título	30%
Características	40%
Descripción	30%

Tabla 4-1. Ponderación de componentes textuales de producto para generación de representación final.

La Figura 4-1 describe la segunda fase de la propuesta para el cálculo de viabilidad basada en el análisis de competencia y la información proveniente de reseñas de producto. En esta etapa, los productos se agrupan a partir de sus representaciones vectoriales previamente generadas, lo que facilita la identificación de productos similares mediante la selección de vectores de referencia y el cálculo de similitudes semánticas entre un nuevo producto y aquellos presentes en una base de información definida.

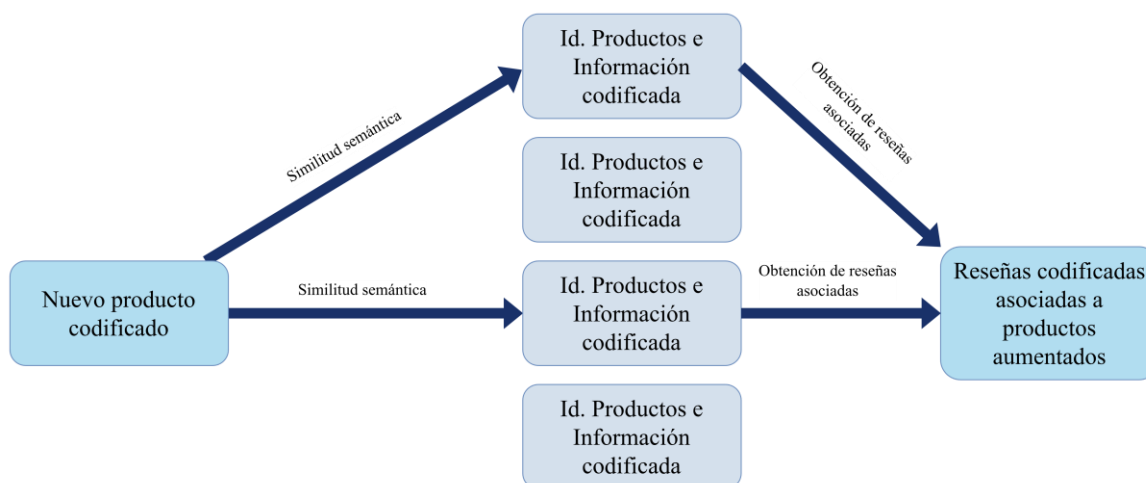


Figura 4-1. Proceso de búsqueda de productos con propósitos de asociación de productos similares.

Por otra parte, la Figura 4-2 presenta la estrategia de caracterización de la métrica de viabilidad en función de las calificaciones otorgadas por los usuarios en las reseñas correspondientes a un grupo de productos seleccionados como referencia para el nuevo ítem bajo análisis.

En la Figura 4-2 (a) y la Figura 4-2 (b) se ilustran dos enfoques ligeramente diferentes para la estimación de la viabilidad a partir de la asociación entre las codificaciones semánticas de las reseñas y la representación del producto. El enfoque presentado en la Figura 4-2 (a) propone el uso directo de la calificación numérica original de la reseña, expresada en una escala de 0 a 5. Por otro lado, la Figura 4-2 (b) plantea una alternativa orientada a mitigar los hallazgos previamente discutidos en la Figura 3-3, transformando las calificaciones numéricas en valores booleanos. En este esquema, las reseñas con puntuaciones entre 0 y 3 se consideran no satisfactorias y se asocian a un valor de 0, mientras que aquellas con calificaciones de 4 o 5 se consideran satisfactorias y se asocian un valor de 1. Finalmente, la Figura 4-2 (c) presenta el esquema de inferencia para un nuevo producto, a partir de su descripción codificada y las reseñas de los productos pertenecientes al grupo seleccionado en la etapa anterior.

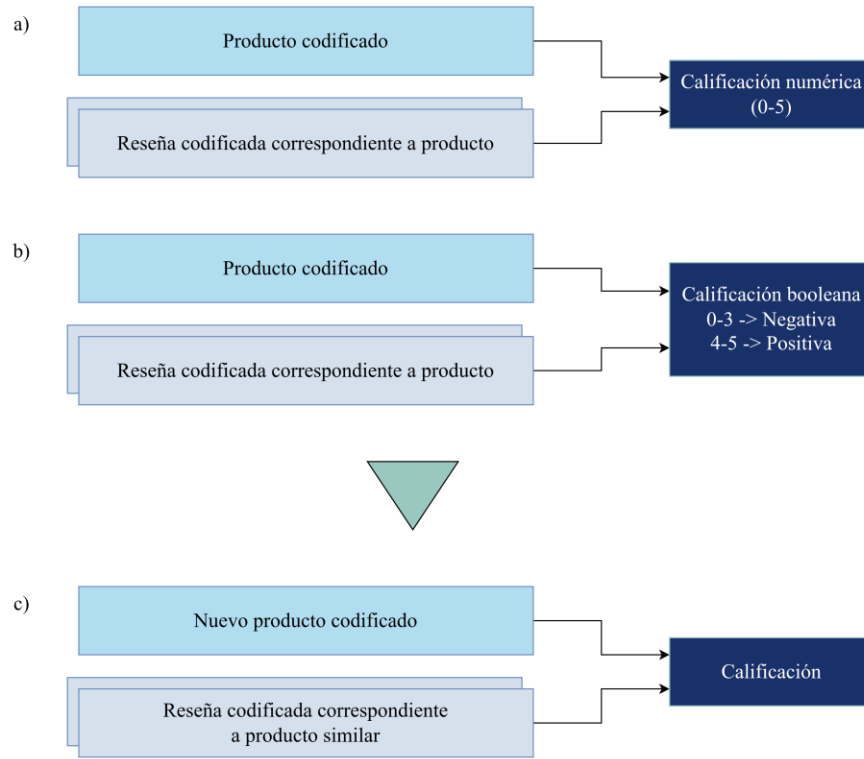


Figura 4-2. Proceso de cálculo de viabilidad en función de la calificación dada a un producto en reseñas individuales. (a) y (b) proponen formulaciones distintas para el tratamiento de la variable objetivo. (c) presenta un esquema del proceso de inferencia para un nuevo producto.

Adicionalmente, en la Figura 4-3 se plantea una estrategia alternativa que no requiere necesariamente el uso de métodos supervisados para el cálculo de la viabilidad final. En su lugar, se propone un enfoque matemático basado en la similitud semántica entre codificaciones textuales, que permite comparar las características de un producto con reseñas positivas y negativas pertenecientes a un conjunto de productos seleccionados por su similitud con la entidad bajo predicción.

En este contexto, las ecuaciones (8) y (9) definen una aproximación que integra la similitud semántica con las consideraciones previamente introducidas en la ecuación (4), incorporando además un mecanismo de ponderación basado en la relevancia de las reseñas y su calificación asociada.

$$E_{ij} = 1 + \varphi \frac{H_i}{\sqrt{T_j}} + \frac{1}{Z} \sum_{k=1}^Z S_{CZ_{kj}} C_{R_{ij}} \quad (8)$$

$$C = \frac{1}{M} \sum_{j=1}^M \frac{\sum_{i=1}^{N_j} E_{ij} R_{ij}}{5 \cdot \sum_{i=1}^{N_j} E_{ij}} \quad (9)$$

Donde H_i representa el número total de votos de utilidad de una reseña específica, y T_j es el total de votos para todas las reseñas de un producto j . De esta forma, C es la ponderación final basada en todas las N_j reseñas de un producto j de un total de M ítems asociados a la predicción realizada. Por otra parte, se agrega al componente de ponderación la relación de la similitud semántica S entre cada una de las características codificadas CZ_{kj} del producto j y la reseña codificada CR_{ij} . Este enfoque permite contextualizar la ponderación final en función del grado en que un producto se asemeja, desde el punto de vista semántico, a comentarios positivos que reflejan valor o satisfacción por parte del usuario final.

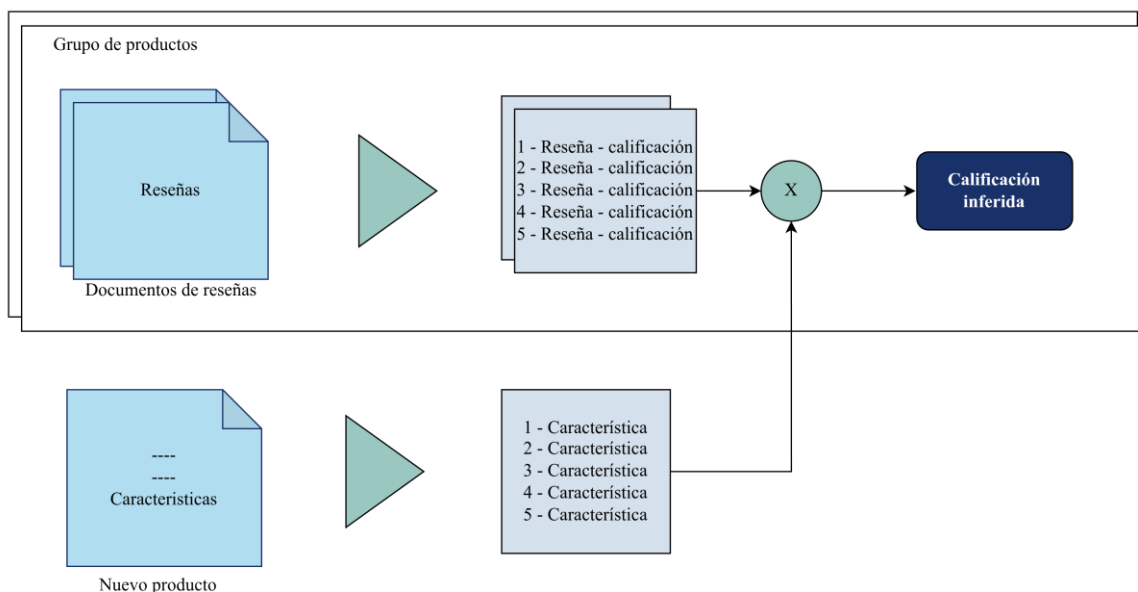


Figura 4-3. Proceso de cálculo de viabilidad a partir de las funcionalidades conocidas para un determinado grupo de productos y las definiciones dadas por JTBD [5].

4.3 Conclusiones de capítulo

En este capítulo se presentó una metodología para la estimación de la viabilidad de productos a partir del análisis de información en plataformas de comercio electrónico. A partir de la revisión de metodologías existentes, se identificó que enfoques como SHELL y Jobs-To-Be-Done (JTBD), aunque originalmente basados en apreciaciones expertas y contexto organizacional, pueden adaptarse a un entorno computacional mediante el uso de datos observables.

La adaptación de la metodología SHELL permitió abordar el análisis del entorno competitivo a partir de la comparación semántica entre productos similares, aun en ausencia de información financiera u organizacional. Por su parte, JTBD aportó un marco conceptual para relacionar las características de los productos con las necesidades de los usuarios, el cual fue reinterpretado mediante representaciones semánticas extraídas de descripciones y reseñas.

El uso de Universal Sentence Encoder posibilitó la generación de representaciones vectoriales de los componentes textuales y la definición formal de similitud semántica

mediante la similitud coseno. Esto facilitó tanto la identificación de productos comparables como la integración de información proveniente de reseñas de usuarios. Se propusieron enfoques supervisados y formulados para la estimación de la viabilidad, destacándose una formulación matemática que incorpora la relevancia de las reseñas, su calificación y la similitud semántica entre características y comentarios.

En conjunto, el capítulo demuestra que es posible construir indicadores cuantitativos de viabilidad de productos a partir de datos públicos, proporcionando una base metodológica sólida para el análisis experimental y la validación de resultados e interpretaciones presentados en los capítulos posteriores.

5. Implementación de modelo predictor de viabilidad numérica

En este capítulo se describen las arquitecturas, herramientas y procedimientos mediante los cuales los resultados obtenidos en la fase de preprocesamiento, junto con las métricas definidas en el capítulo anterior, son transformados en un indicador numérico final de viabilidad. Se detallan los procesos de codificación semántica de productos y reseñas mediante modelos pre-entrenados, las estrategias de agrupamiento y búsqueda de productos similares, y las arquitecturas de modelado empleadas para la estimación del indicador, junto con las especificaciones de sus hiperparámetros.

5.1 Metodología de modelamiento

A partir del conjunto de datos resultante del preprocesamiento, se ejecutó una secuencia de transformaciones orientadas a la construcción de representaciones numéricas de elementos textuales que permitan a su vez realizar una caracterización coherente de las descripciones de entidades. Estas transformaciones permitieron la preparación de los datos para su uso en arquitecturas supervisadas y aplicación de aproximaciones formuladas.

Las categorías de producto fueron codificadas mediante esquemas one-hot, al tratarse de una clasificación única por producto. De manera paralela, los componentes textuales de productos y reseñas fueron codificados y ponderados utilizando el modelo *Universal Sentence Encoder*.

Posteriormente, a partir de las codificaciones semánticas, se generaron conjuntos de datos listos para entrenamiento y validación a partir del emparejamiento de productos similares junto con reseñas asociadas, siguiendo la definición dada en la Figura 4-1.

Los conjuntos de datos fueron divididos de la siguiente manera: 80 % para entrenamiento, 10 % para validación y 10 % para prueba. En los casos pertinentes, se realizó versionamiento de modelos y experimentos, registrando arquitectura, pesos y métricas de desempeño. La cantidad de épocas de entrenamiento para modelos supervisados fue definida de forma empírica con base en el comportamiento observado sobre la evolución de métricas de funciones de pérdida y precisión.

Una vista general del proceso explicado para la construcción del conjunto de datos usado para las arquitecturas de modelo planteadas es presentada en la Figura 5-1.

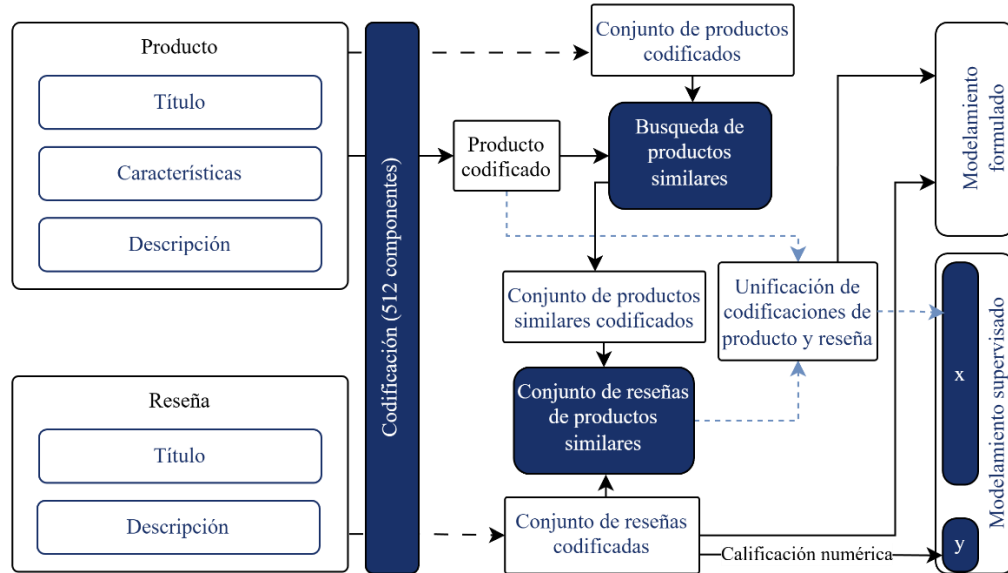


Figura 5-1. Resumen de flujo de procesamiento y construcción de modelo de predicción de viabilidad numérica.

5.2 Identificación de algoritmos y herramientas

Se utilizó PySpark como herramienta principal para la orquestación de flujos de datos, aprovechando sus capacidades de paralelización y su integración con bibliotecas de aprendizaje automático. Esta elección permitió unificar los procesos de transformación, entrenamiento y almacenamiento de modelos dentro de un mismo entorno de datos.

Para el entrenamiento de modelos de aprendizaje profundo se empleó TensorFlow [22], dada su flexibilidad para la construcción de arquitecturas personalizadas y su integración con hardware de aceleración. Adicionalmente, se utilizó MLflow como herramienta de gestión de experimentos y versionamiento, permitiendo la trazabilidad de resultados y la reutilización de configuraciones previas.

5.2.1 Módulos pre-entrenados

Se utilizó la implementación de Universal Sentence Encoder (USE) provista por TensorFlow Hub [23], seleccionando la variante basada en transformadores debido a su capacidad de capturar relaciones semánticas complejas en lenguaje natural.

5.3 Codificación de entidades

Las oraciones previamente segmentadas durante el preprocesamiento fueron codificadas individualmente mediante USE, sin requerir tokenización adicional. Este proceso se optimizó mediante el uso de paralelización distribuida en PySpark y la aceleración por GPU ofrecida por TensorFlow.

Las codificaciones obtenidas fueron agregadas y ponderadas para generar representaciones finales de productos y reseñas según las ecuaciones (10), (11) y (12). En este esquema, se

asignó mayor peso al componente de características del producto, dada su relación directa con las actividades definidas en el marco JTBD, como se mencionó anteriormente.

$$SE_{i,<comp>} = \frac{1}{N_{i,<comp>}} \sum_{j=1}^{N_{i,<comp>}} S_{j,i,<comp>} \quad (10)$$

$$SP_i = 0.3 \cdot SE_{i,titulo} + 0.4 \cdot SE_{i,caracteristicas} + 0.3 \cdot SE_{i,descripcion} \quad (11)$$

$$SR_i = SE_{i,descripcion} \quad (12)$$

$S_{j,i,<comp>}$ representa la codificación de una oración resultante tras aplicar el modelo USE sobre la oración textual j de la entidad i en el componente $<comp>$, por ejemplo, la *segunda* oración del producto número *3* en el componente *descripción*. $N_{i,<comp>}$ es la cantidad de oraciones de una entidad i en el componente $<comp>$.

Así, dado que $SE_{i,<comp>}$ señala la codificación de un componente específico, se tiene finalmente a SP_i y SR_i como las codificaciones representativas de un producto o reseña, respectivamente.

Las representaciones resultantes, originalmente compuestas por 512 valores, fueron sometidas a una reducción de dimensionalidad mediante Análisis de Componentes Principales (PCA) [24]. De esta manera, se entrenaron codificadores PCA independientes para productos y reseñas, buscando preservar al menos el 90 % de la varianza explicada. Como resultado, las codificaciones de productos se redujeron a 150 dimensiones representando una varianza explicada del 91.66 %, y las codificaciones de reseñas se redujeron a 250 dimensiones con una varianza explicada de 91,79 %. Estos resultados son visibles en las Figuras 5-2 y 5-3. El proceso de codificación realizado hasta este punto se encuentra resumido en la Figura 5-4.

Finalmente, los vectores fueron normalizados mediante norma L2, garantizando magnitud unitaria y preservando las propiedades de la similitud coseno. Este paso contribuyó además a la estabilidad numérica y a la compatibilidad con algoritmos basados en distancias euclidianas, como se discutirá más adelante.

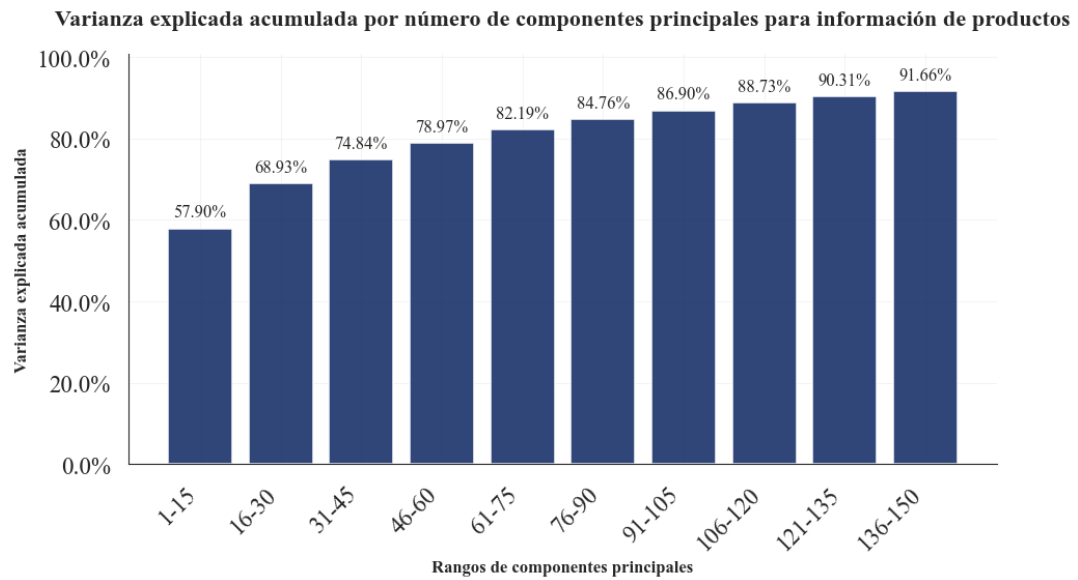


Figura 5-2. Generación de codificaciones mediante PCA para información de productos

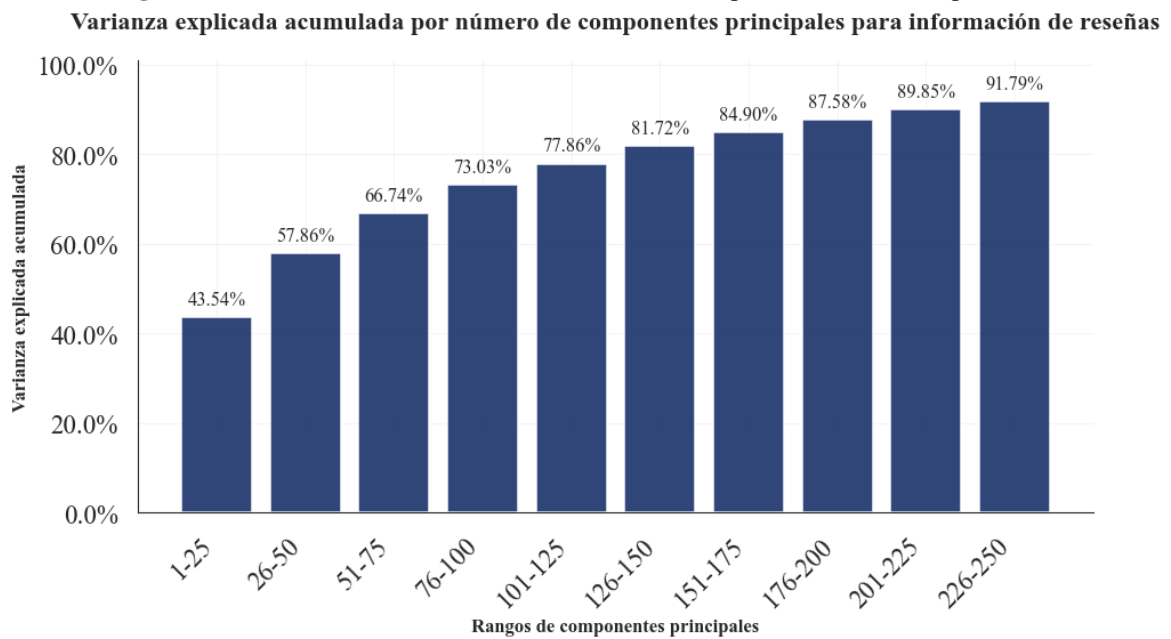


Figura 5-3. Generación de codificaciones mediante PCA para información de reseñas.

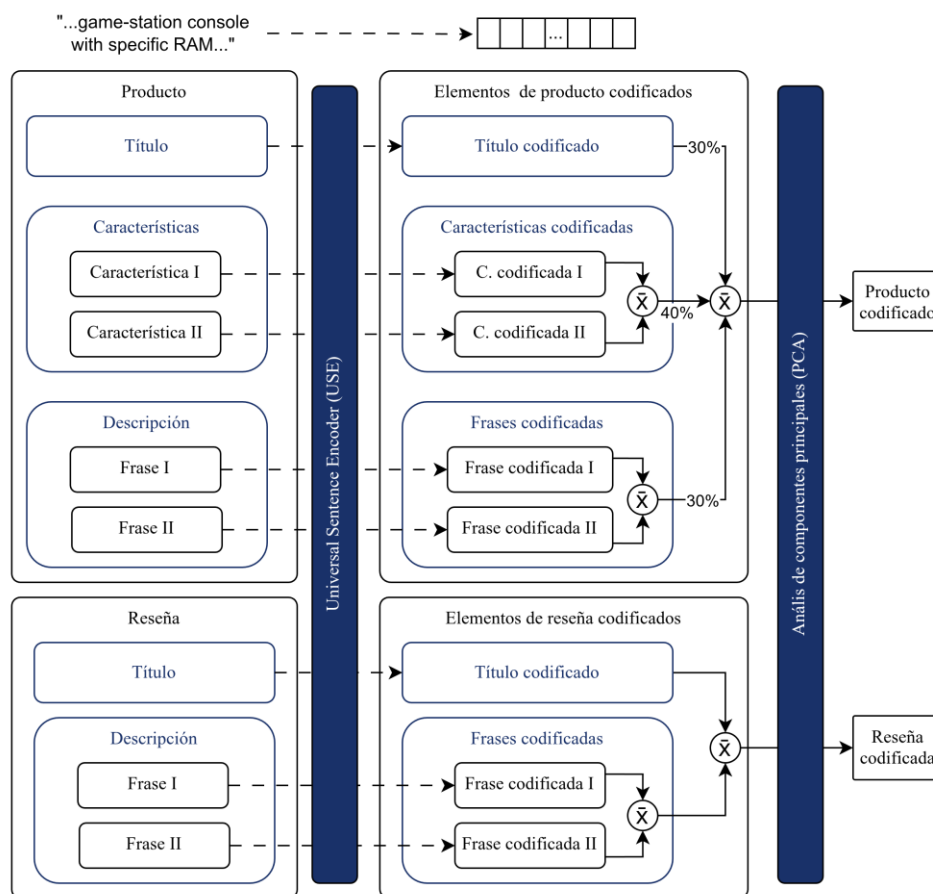


Figura 5-4. Resumen de flujo de generación de codificaciones de entidades de producto y reseña.

5.4 Búsqueda de productos similares

Durante el desarrollo del trabajo se evaluaron diferentes alternativas para la generación de agrupaciones semánticamente coherentes que permitieran la identificación eficiente de productos similares dentro del conjunto de datos. Este paso resulta fundamental, ya que define el subconjunto de productos y reseñas relevantes que serán utilizados para la estimación de viabilidad de un nuevo ítem. Todos los métodos considerados se basan en la medición de distancias entre representaciones vectoriales normalizadas, utilizando la similitud coseno como métrica principal, definida previamente en la ecuación (7).

5.4.1 LSH (Locality Sensitive Hashing)

Se seleccionó el algoritmo Locality Sensitive Hashing (LSH) [25], el cual permite la identificación eficiente de ítems similares sin necesidad de definir un número fijo de grupos. LSH se basa en la proyección de vectores en funciones hash sensibles a la localidad o cercanía según distancia euclidiana, lo que facilita la recuperación de vecinos cercanos en espacios de alta dimensionalidad.

Para ello, y haciendo provecho de la normalización L2 aplicada sobre las codificaciones de representaciones vectoriales, se puede plantear la relación entre la similitud coseno y la distancia euclidiana como se muestra en la ecuación (13).

$$d_{euc} = \sqrt{2 \cdot (1 - d_{cos})} \quad (13)$$

Se definió como criterio un valor mínimo de similitud coseno de 0.9. Este umbral permitió garantizar que los productos asociados presentaran una alta coherencia semántica, al tiempo que se mantuvo un balance adecuado entre precisión y cobertura.

A partir de la generación de pares de productos similares, se construyeron grupos dinámicos sin imponer restricciones explícitas sobre su tamaño o cantidad. Esta estrategia demostró ser particularmente eficiente para el conjunto de datos analizado, compuesto por aproximadamente 65.000 productos, evitando el costo computacional asociado a enfoques de comparación exhaustiva.

5.5 Plan de entrenamiento y selección de hiperparámetros

Los modelos de aprendizaje profundo fueron implementados utilizando TensorFlow, permitiendo la construcción de arquitecturas flexibles adaptadas a las características del problema. En todos los casos se empleó el optimizador Adam [26], seleccionado por su capacidad de convergencia eficiente y su robustez frente a variaciones en la escala de los gradientes.

Se definió una tasa de aprendizaje base de $1 \cdot 10^{-3}$ para todos los modelos, manteniendo consistencia entre las diferentes arquitecturas evaluadas. Esta tasa fue ajustada dinámicamente mediante callbacks de reducción automática con base en el comportamiento de la función de pérdida sobre el conjunto de validación, generando principalmente disminuciones graduales ante reducciones en pérdida poco significativas.

Cada una de las arquitecturas de modelo planteadas propondrá una función de pérdida apropiada según las características de la salida de cada modelo, así como métricas de evaluación que se discutirán próximamente. De manera similar, la estructura de la red, así como las propiedades de cada una de las capas de la misma, serán presentados en la definición de arquitectura por modelo.

Cada modelo fue entrenado utilizando callbacks para monitoreo de métricas de validación, reducción adaptativa de la tasa de aprendizaje, restauración de pesos ante degradación del desempeño y guardado persistente de los mejores estados del modelo.

Todas las ejecuciones de entrenamiento y sus resultados fueron versionadas mediante MLflow, registrando arquitecturas, hiperparámetros, métricas y artefactos de salida. El almacenamiento de los modelos se realizó localmente en formato de archivos planos, facilitando la reproducibilidad de los experimentos.

5.6 Arquitecturas de modelado para predicción de viabilidad numérica

Una vez definidos los grupos de productos similares, se procedió con el diseño de arquitecturas de modelado para la predicción de viabilidad numérica. Estas arquitecturas utilizan como entrada representaciones vectoriales concatenadas de productos y reseñas, derivadas de las etapas de codificación y reducción de dimensionalidad previamente descritas. Inicialmente, se plantearon modelos supervisados para la utilización de codificaciones de reseñas y productos asociados mediante la revisión de pertenencia a un mismo grupo. Sin embargo, también se propusieron aproximaciones matemáticas que no requerirán de modelamiento adicional supervisado al ser resultado de la asociación de codificaciones numéricas únicamente.

Las representaciones finales consisten en vectores de 400 componentes, producto de la concatenación de 150 componentes representativos de información de producto y 250 componentes representativos de información de reseñas.

Dado que cada producto puede asociarse a múltiples reseñas dentro de su grupo semántico, las predicciones generadas por los modelos supervisados se consolidan mediante un promedio ponderado. El peso asignado a cada predicción individual se calcula utilizando el factor definido en la ecuación (14), donde H_i representa los votos de utilidad de la reseña i y T_j el total de votos del producto j .

$$\frac{H_i}{\sqrt{T_j}} \quad (14)$$

5.6.1 Modelo de regresión mediante valoración binaria

El primer modelo propuesto aborda la estimación de viabilidad como un problema de clasificación binaria. Para ello, las calificaciones originales de reseñas fueron transformadas a valores booleanos, asignando el valor 1 a reseñas con 4 o 5 estrellas y 0 a los demás casos. La arquitectura del modelo, presentada en la Figura 5-5, consiste de una red neuronal densa con un total de 145,921 parámetros entrenables. Se utilizó el optimizador Adam con tasa de aprendizaje $1 \cdot 10^{-3}$ y la función de pérdida de entropía cruzada binaria. Este enfoque permitió mitigar el desbalance observado en la distribución de la variable objetivo y facilitó una interpretación directa del indicador de viabilidad, simplificando su interpretación a un valor de 1 como producto viable y 0 como producto no viable.

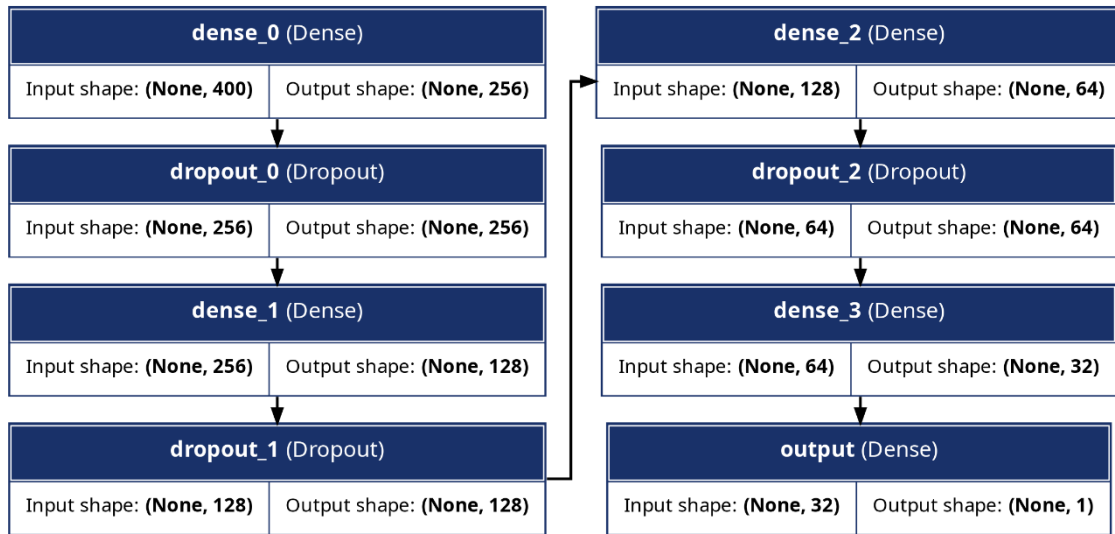


Figura 5-5. Modelo para la predicción de viabilidad en función de calificaciones dadas en reseñas de usuario

5.6.2 Modelo de clasificación mediante valoración categórica

Como segunda aproximación, se planteó un modelo de clasificación en el cual la calificación de la reseña fue considerada como un valor categórico no ordinal donde cada uno de los 6 posibles niveles (valores enteros en el rango de 0 a 5) representa un nivel de satisfacción con implicaciones específicas. Este modelo, cuya arquitectura se muestra en la Figura 5-6, busca evaluar la posibilidad de que las apreciaciones numéricas elegidas por usuarios reflejen por sí mismas un nivel de satisfacción independiente de escalas superiores o inferiores al sistema de estrellas propuesto.

El modelo cuenta con 146,086 parámetros entrenables, emplea el optimizador Adam con tasa de aprendizaje $1 \cdot 10^{-3}$ y la función de pérdida de entropía cruzada categórica.

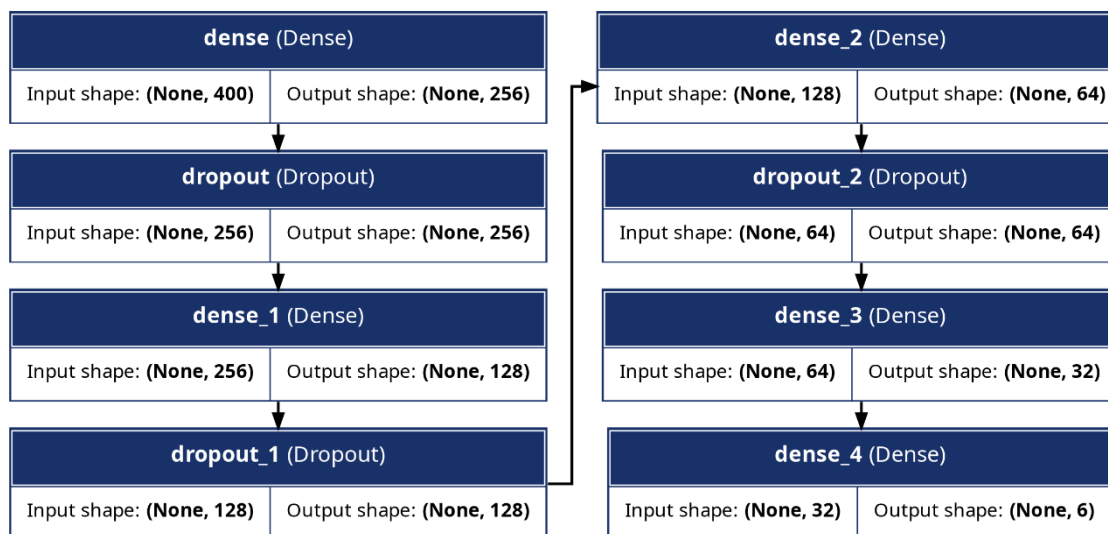


Figura 5-6. Modelo para la predicción de viabilidad en función de regresión real sobre información de calificación de reseña.

5.6.3 Modelo de regresión mediante valoración categórica (incluyendo información de categoría)

Retomando la consideración de la puntuación de reseñas como variables booleanas, se creó una modificación del primer modelo para incluir información de la categoría de producto que se asocia al objeto de predicción actual.

Como se observa en la Figura 5-7, se plantea un modelo adicional destinado a generar predicciones de la categoría de producto dada la codificación de un ítem. Nótese que se cuenta con las 9 categorías previamente seleccionadas y listadas en la Tabla 3-4. Así mismo, la Figura 5-8 presenta la arquitectura de modelo completa que toma bajo consideración la categoría de producto en una fase cercana al final del cálculo de viabilidad numérica, a fin de proveer relevancia a esta variable, conformando así una arquitectura de modelo en cascada.

Se cuenta con 861,961 parámetros para el modelo predictor de categorías y 146,705 para el predictor de viabilidad numérica. Para la construcción de conjuntos de entrada se considera la adición de un total de 9 variables, representativas de las categorías de producto posibles, previamente codificadas como vectores *one-hot*, resultando así en conjuntos de entrada compuestos por dos entradas de 400 y 9 componentes.

Para ambos modelos se usó como base el mismo optimizador Adam para la propagación de la función de pérdida de entropía cruzada; multiclase para el caso de la predicción de categoría y binaria para el caso de la predicción de viabilidad. Se mantuvo la misma tasa de aprendizaje de los modelos anteriores, correspondiente a $1 \cdot 10^{-3}$.

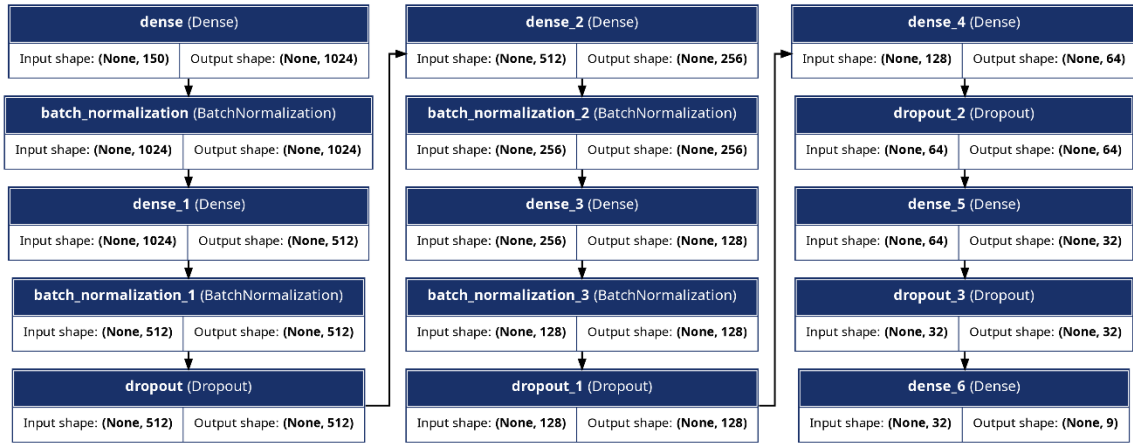


Figura 5-7. Modelo para la predicción de categoría de producto.

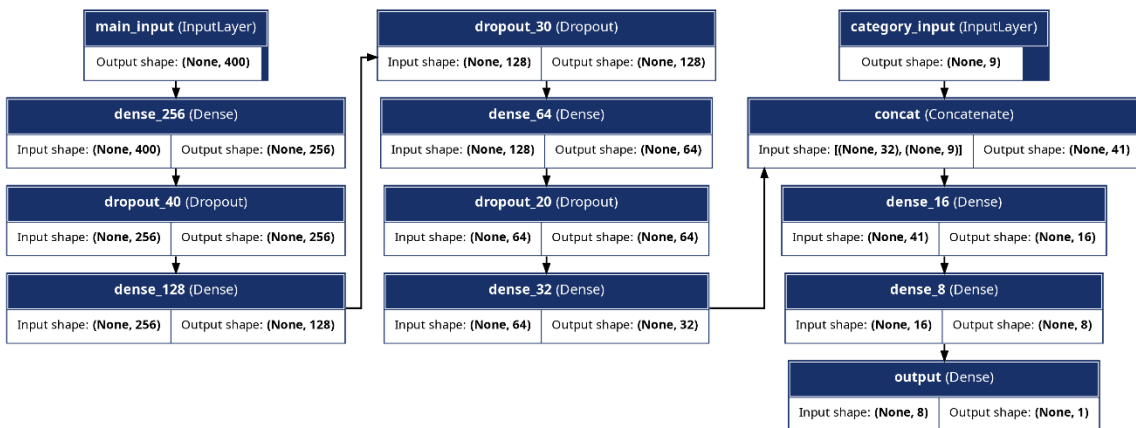


Figura 5-8. Modelo para la predicción de viabilidad booleana de producto con contexto de categoría de producto elegida. La segunda entrada del modelo, proveniente de la salida del modelo en la Figura 5-7, corresponde a la capa de entrada a la derecha, cerca de la salida final.

5.6.4 Valoración mediante modelado formulado

Como alternativa a los modelos supervisados, se implementó una métrica formulada definida en las ecuaciones (8) y (9). Esta aproximación utiliza directamente la similitud semántica entre características de producto y reseñas, ponderada por la utilidad de las reseñas, sin requerir entrenamiento adicional. Al igual que los modelos supervisados previamente descritos, se usará la misma metodología de agrupación de productos con reseñas asociadas para calcular el resultado de la ponderación matemática sobre un objeto de predicción nuevo.

5.7 Conclusiones de capítulo

El planteamiento de métodos tanto supervisados como formulados para la generación del indicador de viabilidad destino no solo ofrece un mayor rango de estudio de posibilidades para alcanzar el objetivo del trabajo, sino además presenta la posibilidad de aprovechar las intuiciones matemáticas concebidas para el tratamiento de vectores y enfrentar sus capacidades y limitaciones a aproximaciones comunes en el contexto actual, como lo son las redes neuronales densas y sus variantes.

La cantidad de parámetros requeridos para los modelos supervisados basados en redes neuronales reflejan una complejidad de red relativamente baja, principalmente gracias al papel de la codificación y posterior reducción de dimensionalidad a través de un modelo pre-entrenado, lo cual permitió la síntesis de información semántica de manera previa a la incorporación de una estructura con menos parámetros, siendo este planteamiento similar a la estrategia Transfer Learning.

La aplicación de LSH como método de búsqueda de productos similares ofreció una estrategia eficiente para llevar a cabo esta tarea sobre un conjunto de datos real de alrededor de 65.000 productos, lo cual bajo una estrategia de fuerza bruta conllevaría a la realización de hasta 4'225.000.000 comparaciones para lograr obtener un conjunto de datos de entrenamiento consistente.

6. Evaluación de modelo predictor de viabilidad numérica

Este capítulo define, evalúa y analiza las métricas de desempeño asociadas a cada una de las arquitecturas de modelos planteadas anteriormente. Se revisan tanto métricas evaluadas sobre conjuntos de datos de prueba, como la evolución de aprendizaje en conjuntos de datos de validación. Adicionalmente, se realizan pruebas integrales sobre el flujo de codificación, agrupamiento e inferencia para entradas de usuario generadas a partir de la estructura del conjunto de datos original.

6.1 Planteamiento de métricas de desempeño

Para la evaluación de los modelos supervisados implementados para la predicción de la calificación numérica de usuarios, se emplearon métricas basadas en el comportamiento de la función de pérdida seleccionada para cada arquitectura. En particular, se utilizó la entropía cruzada, según lo descrito en [27].

Adicionalmente, se calcularon métricas clásicas de evaluación para clasificación, tales como precisión, recall y puntuación F1, con el fin de analizar el desempeño del modelo ante posibles desbalances de clase y evaluar su capacidad de generalización sobre conjuntos de validación.

Por otra parte, la evaluación del modelo formulado, derivado de las expresiones matemáticas presentadas en las ecuaciones (8) y (9), se planteó mediante la comparación entre el valor de viabilidad predicho y la calificación promedio de las reseñas asociadas a cada producto que contara con al menos un ítem similar dentro del conjunto identificado mediante similitud semántica. Dado que los valores de salida del modelo formulado se encuentran normalizados en el rango [0,1], se seleccionaron métricas basadas en la medición de error continuo, específicamente el error absoluto medio (MAE) y el error cuadrático medio (RMSE), definidos en las ecuaciones (15) y (16), respectivamente.

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i| \quad (15)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (16)$$

6.2 Evaluación de desempeño de modelos

6.2.1 Modelo de clasificación mediante valoración binaria

La Figura 6-1 presenta los resultados en términos de evolución de función de pérdida y precisión tanto en conjunto de datos de entrenamiento como conjunto de validación para el modelo de clasificación basado en la transformación de la valoración de reseñas a formato binario. Se observa una precisión de alrededor del 94% para el conjunto de datos de validación, considerándose una métrica prometedora para el proceso de inferencia realizado.

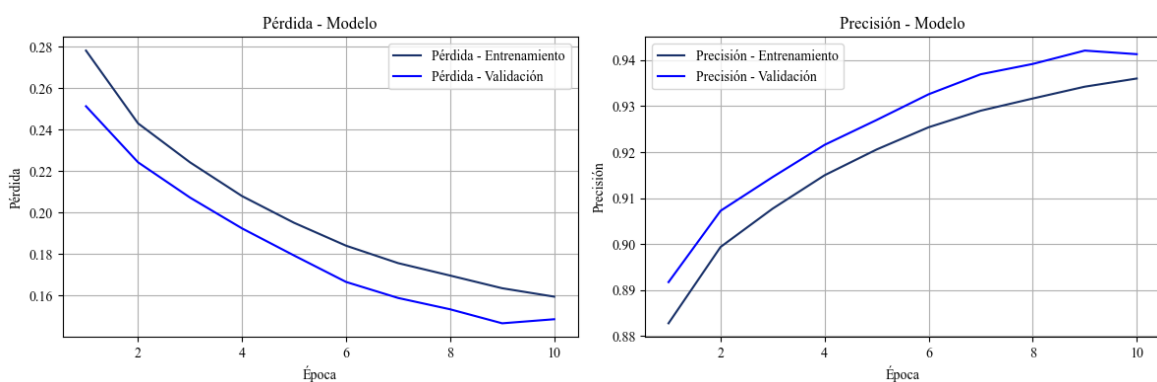


Figura 6-1. Métricas de evaluación para modelo de predicción de viabilidad en función de clasificación binaria sobre información de calificación de reseña.

La Figura 6-2 presenta, por otra parte, la matriz de confusión asociada al modelo categórico implementado. Pueden intuirse algunas formulaciones que permiten contemplar de manera numérica el desempeño de la clasificación binaria, interpretando la etiqueta 1 como producto “viable” y la etiqueta 0 como producto no viable, según la información de la reseña y el producto unificados, como se explicó anteriormente con relación a la construcción de los conjuntos de datos de entrenamiento.

Una precisión de 96.63 %, así como en conjunto de validación y una tasa de verdaderos positivos o recall del 94.74 %, reflejan una capacidad certera para la tarea de predecir la calificación de un producto dada una reseña, lo anterior se ve reforzado por un valor F1 del 95.67 %, aportando veracidad ante el desbalance de clases previamente suavizado en etapas de preprocesamiento.

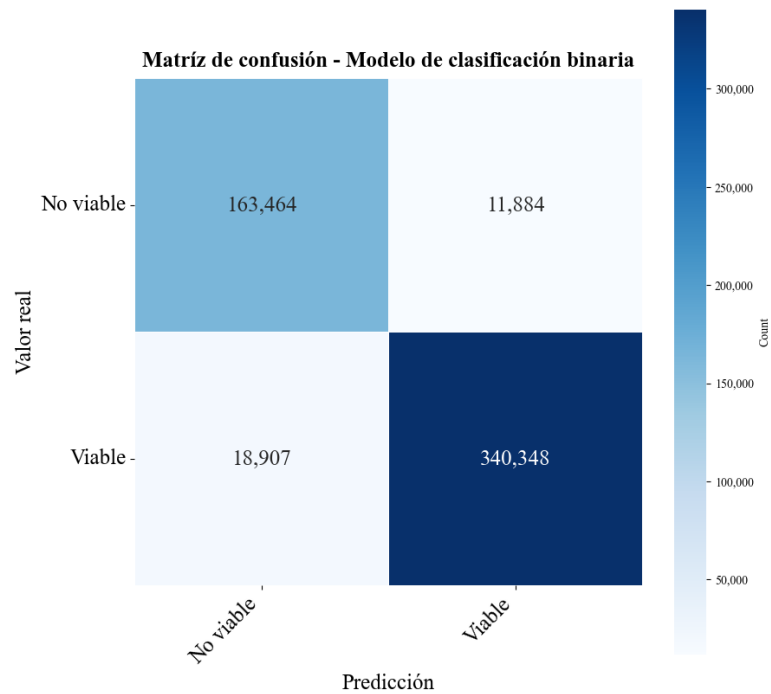


Figura 6-2. Matriz de confusión para modelo de predicción de viabilidad en función de clasificación binaria sobre información de calificación de reseña.

6.2.2 Modelo de clasificación mediante valoración categórica

La Figura 6-3 presenta los resultados en términos de evolución de función de pérdida y precisión en conjuntos de entrenamiento y validación para el caso del modelo de clasificación basado en la calificación original conservada en el rango discreto 0-5. Se observan resultados inferiores a los alcanzados por el modelo anterior. Este comportamiento podría explicarse dado el desbalance de clases discutido anteriormente en la Figura 3-13, el cual, a pesar de ser tratado mediante el muestreo de clases, reduce la capacidad de generalización del modelo. Sin embargo, se considera que los resultados sustentan una precisión moderada del 78 % que puede resultar útil para la predicción bajo la suposición de una variable no ordinal.

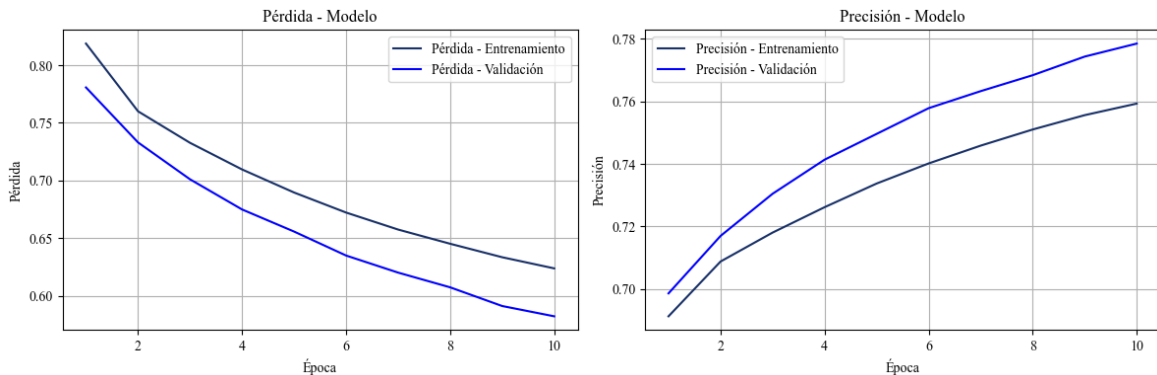


Figura 6-3. Métricas de evaluación para modelo de predicción de viabilidad en función de clasificación multiclase sobre información de calificación de reseña.

La Figura 6-4 no refleja resultados precisos a comparación con la matriz de confusión presentada en la Figura 6-2. Es evidente que se cuenta con una precisión razonable para los casos de la calificación perfecta y la calificación más baja en el conjunto de datos de prueba, sin embargo, existen sesgos significativos para las clases intermedias, lo cual se ve reflejado en métricas macro como una precisión de 65.50 %, puntuación F1 de 61.60 % y recall del 61.08 %.

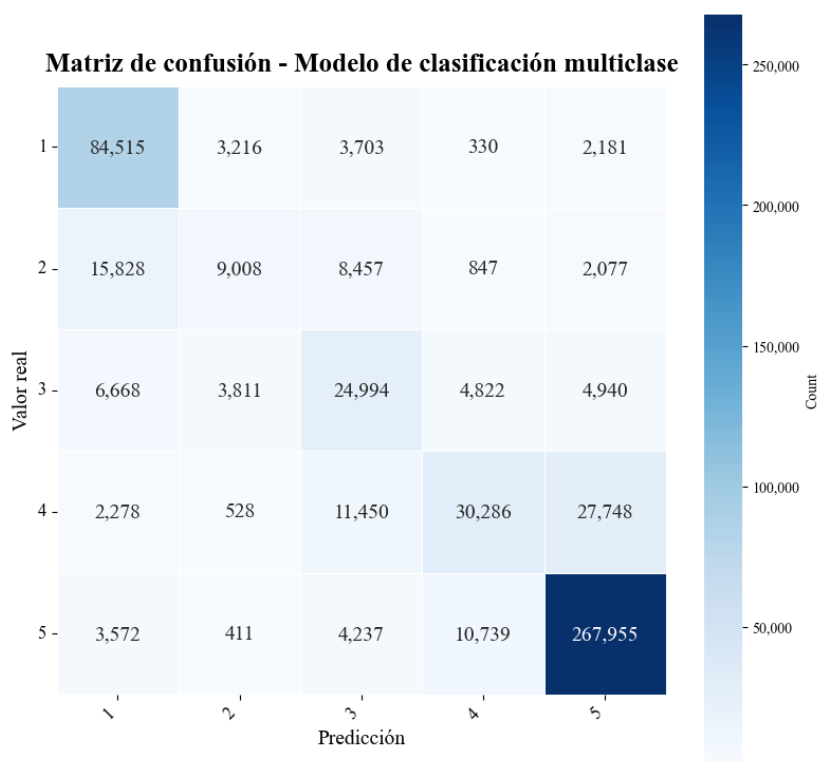


Figura 6-4. Matriz de confusión para modelo de predicción de viabilidad en función de clasificación multiclase sobre información de calificación de reseña.

6.2.3 Modelo de clasificación mediante valoración binaria (incluyendo información de categoría)

La Figura 6-5 presenta los resultados en términos de evolución de función de pérdida y precisión tanto en conjunto de datos de entrenamiento como conjunto de validación para el modelo de clasificación binaria modificado para incluir información de categorías inferidas. Nuevamente se alcanzaron precisiones en conjuntos de validación cercanas al 94%, presentando una relativa baja mejoría con respecto a la alternativa sin consideración de información categórica.

En la Figura 6-6 se sustenta el hallazgo de métricas prometedoras para su variante modificada, llegando esta vez a resultados ligeramente superiores respecto a precisión, con un valor de 96.29 % y un valor F1 de 95.69 %, así como un acierto de casos positivos del 95.11 %. Así, es posible inferir que el contexto adicional de categoría introducida por el modelo llega a resultar de utilidad para la predicción de su viabilidad, sin embargo, el mejoramiento observado al incluir esta variable se concluye como poco significativa dada su diferencia con la aproximación binaria sin inclusión de categoría.

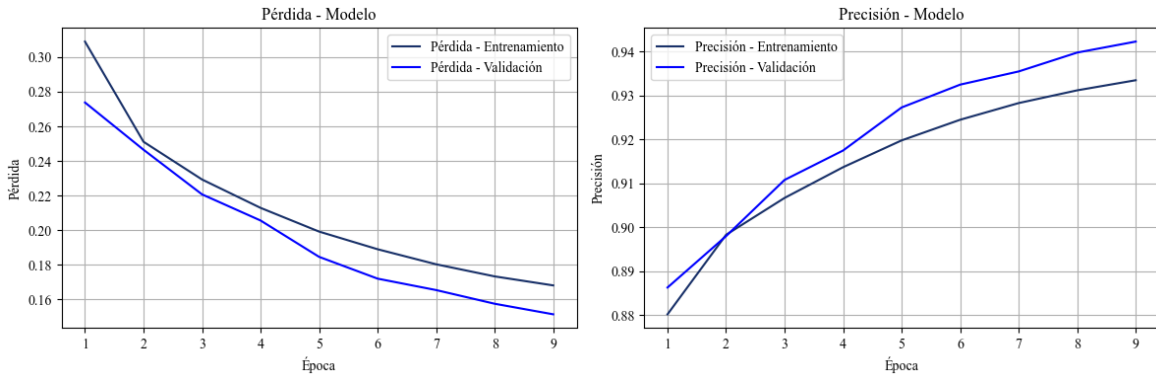


Figura 6-5. Métricas de evaluación para modelo de predicción de viabilidad basado en etiqueta booleana modificado para inclusión de información de categoría de producto.

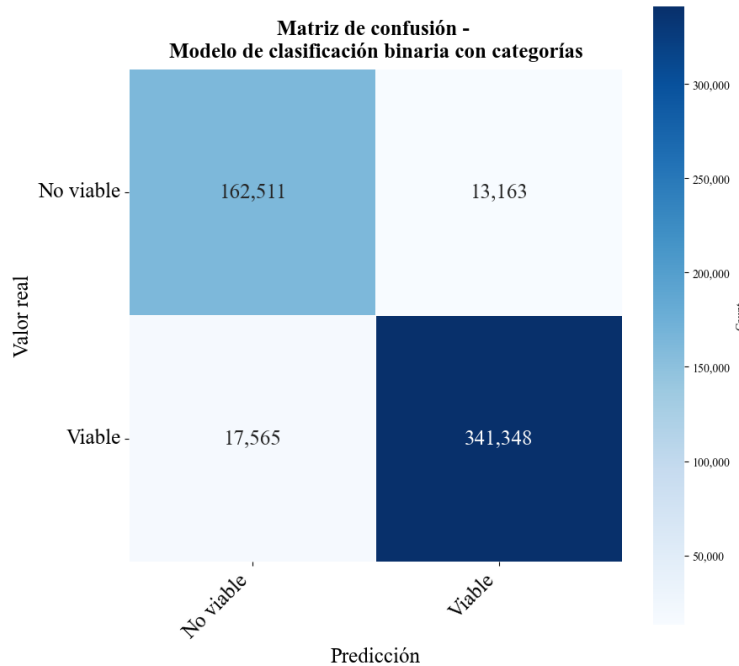


Figura 6-6. Matriz de confusión para modelo de predicción de viabilidad basado en etiqueta booleana modificado para inclusión de información de categoría de producto.

6.2.4 Modelo de regresión formulado

Se medirá el desempeño del modelo de regresión para los ítems con al menos un elemento dentro del conjunto de ítems similares bajo el mínimo de similitud coseno de 0.9, como se describió anteriormente. Se realizó de esta manera una predicción de la viabilidad numérica de los productos con el condicionamiento mencionado. Posteriormente, se calcularon las métricas de error absoluto medio (MAE) y error cuadrático medio (RMSE), como se propuso en la definición de las métricas a usar para medir el desempeño de esta etapa. Se obtuvo un valor de 0.1359 para la métrica MAE, así como un valor de 0.1781 para la métrica RMSE.

6.3 Ejemplificación de caso de uso

Se construyó un flujo de procesamiento y entrenamiento que consolida la secuencia de pasos retratada en el presente documento para datos de entrada nuevos. Como se muestra en la Figura 6-7, se ejecuta el flujo de cada una de las fases descritas en el presente trabajo, incluyendo etapas de preprocesamiento, codificación, búsqueda basada en similitud, modelamiento y consolidación de la métrica final a calcular.

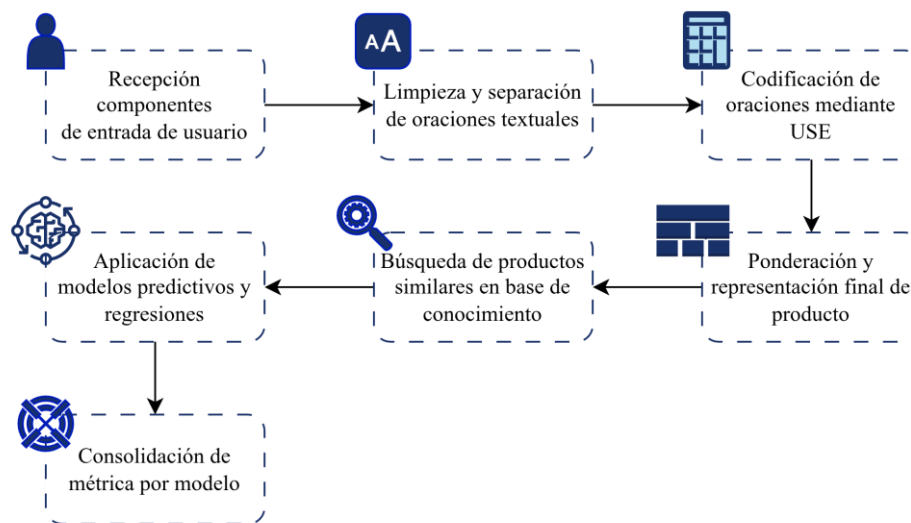


Figura 6-7. Representación de flujo completo de procesamiento de un nuevo ejemplo entrante al proceso de inferencia construido.

Se generó una entidad de predicción con las propiedades especificadas en la Tabla 6-1. Posterior a la limpieza y separación de oraciones textuales, se generó la codificación para cada una de las oraciones segmentadas, generado así una codificación de 512 elementos para cada una, dadas las características de USE y la dimensionalidad de la codificación generada. Posteriormente, y siguiendo con la ponderación especificada en la Tabla 4-1. Se genera una codificación consolidada que representa los datos de entrada en un único componente vectorial. Se condensa además el vector resultante mediante la estrategia de reducción de componentes previamente relacionada en la Figura 5-2, esto es, la reducción a un total de 150 elementos.

Componente	Texto
Título	Canon PowerShot SD1000 7.1MP Digital Elph Camera with 3x Optical Zoom (Black) (OLD MODEL)
Características	<ul style="list-style-type: none"> - 7.1-megapixel CCD captures enough detail for photo-quality 15 x 20-inch prints - 3x optical zoom; ISO 1600 and High ISO Auto - DIGIC III Image Processor; Face Detection AF/AE
Descripciones	<ul style="list-style-type: none"> - Canon PowerShot SD1000 7.1MP Digital Elph Camera with 3x Optical Zoom (Black) - PowerShot SD1000 Highlights - Slim, stylish 7.1-megapixel digital Elph with 3x optical zoom

Tabla 6-1. Estructura de ejemplo generado como entrada para flujo de modelo.

Dada la representación vectorial previa, se procede a realizar la búsqueda de productos similares, usando como métrica la similitud coseno definida en la ecuación (7) y un límite de al menos 0.9 como resultado de esta para filtrar los ítems relacionados. De esta manera se encontró 1 ítem cuyos atributos se presentan en la Tabla 6-2.

#	Componente	Texto
1	Título	Canon PowerShot SD1000 7.1MP Digital Elph Camera with 3x Optical Zoom (Black) (OLD MODEL)
	Características	- 8.0-megapixel CCD captures enough detail for photo-quality 16 x 22-inch prints

Descripciones	- 4x Optical Image Stabilized zoom for steady, long zoom shooting
	- High-resolution 2.5-inch PureColor LCD with scratch-resistant, anti-reflection coating
	- Find multiple faces with Canon's improved Face Detection technology
	- Vivid, high-resolution 2.5-inch PureColor LCD
	- Sensitivity range expanded to ISO 1600

Tabla 6-2. Estructura de productos similares asociados a ejemplo de la Tabla 6-1. Se generaron los conjuntos de predicción para aplicación en cada una de las arquitecturas de modelos dadas, resultando así en la consolidación de métricas presentada en la Tabla 6-3.

Modelo	Resultado		
Clasificación binaria	0.8069		
Clasificación categórica	0 => 6.946e-10	1 => 6.223e-02	2 => 4.460e-02
	3 => 9.136e-02	4 => 2.236e-01	5 => 5.782e-01 (Clase seleccionada)
Categoría predicha	Computers:	All Electronics:	Home Audio & Theater:
	3.637e-10	1.845e-06	1.345e-12
	Amazon Home:	Industrial & Scientific:	Cell Phones & Accessories:
	7.899e-09	6.012e-09	1.253e-08
	Car Electronics:	Office Products:	Camera & Photo:
	5.805e-17	4.95e-12	9.999e-01

	(Categoría seleccionada)
Clasificación binaria con información de categoría	0.8024
Clasificación basada en modelo formulado	0.8715

Tabla 6-3. Resultados de predicción para ejemplo de flujo completo planteado.

6.4 Análisis de resultados y selección de mejores modelos

Tras el análisis de las métricas de evaluación aplicadas al conjunto de modelos implementados, se observa inicialmente un desempeño limitado bajo la suposición de una variable objetivo categórica. En este escenario, la mejor variante alcanzó una precisión máxima del 78%, lo que evidencia una capacidad predictiva insuficiente. Esta aproximación fue adoptada con el objetivo de analizar el comportamiento del sistema al asumir un nivel de significado independiente para cada posible calificación, descartando deliberadamente el carácter ordinal inherente a dichas valoraciones. Sin embargo, los resultados sugieren que la viabilidad del producto no puede representarse de manera adecuada mediante un conjunto discreto de valores. Si bien el conjunto de datos original presenta esta propiedad por definición, una apreciación continua aporta mayor valor al permitir una mayor flexibilidad en la arquitectura de parámetros utilizada para generar dicha evaluación.

En contraste, las aproximaciones basadas en clasificación binaria, tanto aquellas que consideran la categoría del producto como las que se fundamentan exclusivamente en codificaciones textuales, presentan un desempeño notablemente superior y, además, resultados comparables entre sí. Cabe destacar que se optó por una formulación de clasificación binaria debido a su equivalencia conceptual con un modelo de regresión, al generar apreciaciones dentro de un intervalo continuo $[0,1]$. En ambos casos se alcanzaron niveles de precisión de hasta el 96% para conjuntos de datos de pruebas, lo cual refleja las ventajas de una arquitectura menos compleja y de una variable objetivo definida como un valor escalar. Este enfoque habilita una reinterpretación de la viabilidad del producto, transformando valoraciones discretas de usuarios en un resultado derivado del análisis de información contextual, la relación con productos previamente comercializados y apreciaciones que incorporan retroalimentación en lenguaje natural.

Adicionalmente, se observa que la aproximación formulada a partir de trabajos previos y métodos predominantemente cualitativos guarda una estrecha relación conceptual con los modelos de redes neuronales implementados. En ambos casos, el proceso se fundamenta en la identificación de productos similares y en la comparación de elementos textuales representados como vectores numéricos. Por un lado, esta comparación puede realizarse mediante métricas que cuantifican el grado de similitud semántica entre textos, como la similitud coseno. Por otro, las redes neuronales aprenden de manera implícita un conjunto de transformaciones sobre la combinación de dichos vectores codificados.

Ambas estrategias utilizan como base la calificación numérica proveniente de reseñas de productos similares, obtenida tras el análisis de similitud de sus componentes textuales. No obstante, la implementación basada en redes neuronales ofrece resultados dinámicos que permiten capturar abstracciones derivadas de transformaciones intermedias propias de su arquitectura. Esto permite la generación de interpretaciones que trascienden las limitaciones de una formulación matemática rígida, lo cual puede explicar el desempeño superior observado en los modelos neuronales frente a soluciones derivadas de enfoques previos para la estimación de viabilidad de productos.

Es importante señalar que el subconjunto de información utilizado en la etapa de modelamiento de este trabajo representa únicamente una fracción de las consideraciones abordadas por métodos clásicos, lo que constituye una limitación para su formulación en función de la totalidad de los aspectos documentados desde dichas perspectivas.

La ejemplificación de casos de uso refuerza la relación entre ambos enfoques, dado que las diferencias obtenidas en los resultados se mantienen dentro del rango estimado a partir de la evaluación del error absoluto medio del modelo propuesto.

Finalmente, cabe mencionar que la variable correspondiente a la categoría del producto no alcanza un nivel de significancia relevante, a juzgar por las diferencias marginales observadas entre las arquitecturas que incluyen o excluyen dicha variable. Este comportamiento puede atribuirse a la selección inicial de categorías primarias asociadas al contexto tecnológico, en el cual es posible que no exista una relación fuerte entre la categoría del producto y la información textual analizada. En este contexto, dicha información se encuentra implícita en la descripción y en las características recopiladas del ítem, lo que hace redundante la incorporación explícita de una variable categórica adicional.

Así, ambas variantes del modelo binario, con y sin inclusión de la categoría de producto asociada como entrada del modelo, son seleccionados como alternativas viables para estimación de viabilidad de productos tecnológicos.

6.5 Conclusiones de capítulo

La representación de elementos textuales mediante codificaciones vectoriales demostró ser una estrategia adecuada para la identificación y comparación de productos similares, validando la construcción de conjuntos de entrenamiento y validación para las arquitecturas propuestas.

Los modelos supervisados evaluados no presentaron indicios de sobreajuste ni estancamiento durante el entrenamiento, lo que respalda la validez de las arquitecturas densas construidas a partir de representaciones semánticas generadas por modelos basados en transformadores.

Se identificó que el modelo de clasificación binaria, con y sin inclusión de información de categoría, ofrece el mejor desempeño global, beneficiándose de la reducción del desbalance de clases y del enriquecimiento contextual proporcionado por la categoría del producto.

Finalmente, la aproximación formulada inspirada en JTBD presentó resultados coherentes y complementarios, destacando el valor de integrar información de relevancia de reseñas en la estimación de viabilidad, especialmente en escenarios con disponibilidad limitada de etiquetas.

7. Conclusiones y trabajo futuro

7.1 Conclusiones

- La viabilidad de un producto en línea puede medirse tomando como punto de partida la estimación de la calificación numérica dada por usuarios en plataformas de comercio electrónico siendo esta aproximación soportada tanto por trabajos previos como por el marco de metodologías cuantitativas ampliamente aceptadas en el contexto de la planeación estratégica.
- Una mayoría significativa del comercio electrónico en línea en el contexto tecnológico recae sobre ciertas categorías específicas, principalmente aquellas relacionadas con productos celulares y accesorios de celulares, reflejando así la priorización de la practicidad y movilidad sobre la complejidad en el contexto tecnológico diario.
- Los modelos de aprendizaje automático basados en redes neuronales demostraron una capacidad consistente para estimar la viabilidad de productos a partir de representaciones vectoriales, confirmando la pertinencia de enfoques neuronales en la interpretación de información no estructurada proveniente de plataformas de comercio electrónico.
- La evaluación de arquitecturas con y sin consideración de categorías de producto reflejan una relativa poca relevancia de esta propiedad, ya que se encontraron resultados similares tanto en desempeño durante entrenamiento como en la realización de pruebas sobre información de un producto de ejemplo.
- Las métricas de evaluación obtenidas en los conjuntos de prueba evidencian una superioridad clara de los modelos de clasificación binaria frente a otras formulaciones del problema, posicionándolos como la alternativa más adecuada para la tarea de predicción de viabilidad abordada en este trabajo.
- La aproximación formulada inspirada en el marco conceptual *Jobs To Be Done* (JTBD) generó resultados coherentes con los modelos supervisados entrenados. Sin embargo, se vio superada por estos últimos posiblemente debido a la limitación en la consideración de variables y la rigidez de una aproximación matemática ante la consideración de un contexto proveniente de expresiones en lenguaje natural.
- La incorporación de un método de ponderación, como la relevancia de reseñas, logró añadir un nivel adicional de abstracción semántica, enriqueciendo la predicción final y aportando información de valor para la interpretación de resultados desde una perspectiva de negocio.
- La codificación de componentes textuales facilita la interpretación de entidades como variables numéricas procesables por modelos computacionales. Sin embargo,

resulta útil la incorporación de deducciones basadas en el contexto de negocio, en este caso, la relevancia de atributos de producto diferentes, para la construcción y comparación de codificaciones numéricas.

- La combinación de interpretaciones matemáticas con la robustez de los modelos de aprendizaje de máquina permite la conversión de entidades semánticas y la inferencia de propiedades interpretables a nivel de negocio.

7.2 Recomendaciones

A partir de la base de conocimiento construida, los modelos derivados de esta, y el flujo de procesamiento e inferencia desarrollado, se espera que tanto personas como organizaciones puedan sustentar su toma de decisiones en información proveniente del comercio electrónico real. En particular, se recomienda considerar la existencia de productos similares a los propuestos y la capacidad de las reseñas para representar de manera efectiva las necesidades y percepciones reales de los consumidores, con el fin de estimar de forma más precisa el nivel de satisfacción asociado a un producto específico dentro del mercado.

Durante el desarrollo de este proyecto se aplicaron diversas aproximaciones, tanto en el ámbito del aprendizaje de máquina como en el uso de estándares para el procesamiento de grandes volúmenes de información como etapa previa al entrenamiento y despliegue de modelos de inferencia. En este sentido, se recomienda que el proceso descrito en este documento sea utilizado como referencia y, en determinados contextos, como guía metodológica, para la implementación de flujos completos de investigación y pruebas de concepto orientadas al procesamiento de lenguaje natural, la construcción de modelos predictivos y la estructuración de procesos ETL aplicados.

7.3 Trabajo Futuro

El presente trabajo se desarrolló sobre un conjunto de datos acotado a productos pertenecientes al sector tecnológico. No obstante, el repositorio original recolectado en [6] contiene información adicional correspondiente a múltiples categorías fuera de dicho sector. En este contexto, una línea natural de trabajo futuro consiste en extender los procedimientos aquí descritos al procesamiento de estas categorías adicionales, ampliando así el alcance del análisis y contribuir a la resolución de problemas análogos en distintos dominios. Esta extensión podría derivar en la construcción de herramientas equivalentes, adaptables a modelos de negocio con características y dinámicas diferentes.

Por otro lado, los modelos presentados se fundamentan en una línea base basada en redes neuronales profundas y en esquemas de codificación semántica derivados de estas arquitecturas. Sin embargo, se espera que este trabajo pueda evolucionar en consonancia con los avances del aprendizaje de máquina y, en particular, con los desarrollos emergentes en procesamiento de lenguaje natural y su aplicabilidad a la representación semántica de entidades, incorporando nuevos modelos, arquitecturas y enfoques que permitan mejorar el desempeño y la interpretabilidad de los resultados.

A. Anexo repositorio de código fuente

En este anexo se hace referencia al repositorio digital que contiene el código fuente desarrollado y utilizado durante la realización del trabajo de grado. Dicho repositorio reúne los scripts, funciones y notebooks empleados en las etapas de procesamiento, análisis y experimentación descritas en este documento.

El repositorio se encuentra disponible en la plataforma GitHub en la dirección https://github.com/UNAL-Midas/numeric_availability_predictor.

El código fuente se organiza dentro de la carpeta src/, siguiendo la arquitectura Medallion definida para el desarrollo del proyecto. En esta estructura, las capas bronze, silver y gold representan las distintas etapas del flujo de datos.

B. Anexo repositorio de conjuntos de datos principales

Este anexo hace referencia al repositorio digital que almacena los conjuntos de datos principales utilizados en el desarrollo del trabajo. En este repositorio se incluyen los resultados generados principalmente a partir de las etapas de codificación y pre-modelamiento.

El repositorio de datos se encuentra disponible en la plataforma GitHub, en la siguiente dirección: https://github.com/UNAL-Midas/numeric_viability_predictor_data. Para la correcta descarga de los archivos es necesario el uso de Git LFS, debido al tamaño de los conjuntos de datos almacenados.

Los datos se encuentran almacenados en formato .parquet y organizados mediante un esquema de particionado. Esta estrategia fue utilizada como mecanismo de optimización a lo largo del desarrollo del trabajo.

C. Publicación académica derivada

Como resultado del desarrollo de este trabajo de grado, se elaboró un artículo académico que resume los principales objetivos, metodología y resultados obtenidos. Dicho artículo se encuentra actualmente en proceso de publicación en un medio académico bajo formato IEEE.

8. Referencias

- [1] M. Cantamessa, V. Gatteschi, G. Perboli, y M. Rosano, “Startups’ Roads to Failure”, *Sustainability*, vol. 10, núm. 7, p. 2346, jul. 2018, doi: 10.3390/su10072346.
- [2] Deena, D. P. y Gupta, D. M., “A study on factors that contribute to the failure of startups”, *International Journal of Aquatic Science*, pp. 2634–2640, 2021.
- [3] U. Iyyaz Billah, “Reasons for failure of new products in the consumer goods industry”, *Business Review*, vol. 7, núm. 2, pp. 119–129, jul. 2012, doi: 10.54784/1990-6587.1209.
- [4] A. Murray y V. Scuotto, “The Business Model Canvas”, *Symphonya. Emerging Issues in Management*, pp. 94–109, nov. 2016, doi: 10.4468/2015.3.13murray.scuotto.
- [5] P. Moessner, R. Haegle, L. Eiler, y K. Kloepfer, “Evaluate Market Potential of an Initial Product Concept With Jobs-to-Be-Done”, *IEEE Eng. Manag. Rev.*, vol. 52, núm. 2, pp. 165–173, abr. 2024, doi: 10.1109/EMR.2024.3353618.
- [6] Y. Hou, J. Li, Z. He, A. Yan, X. Chen, y J. McAuley, “Bridging Language and Items for Retrieval and Recommendation”, el 6 de marzo de 2024, *arXiv*: arXiv:2403.03952. doi: 10.48550/arXiv.2403.03952.
- [7] Eggert, A., Thiesbrummel, C., y Deutscher, C., “Differential effects of product and service innovations on the financial performance of industrial firms”, *Journal of Business Market Management*, pp. 380–405, 2014.
- [8] S. A. Rijdsdijk, E. J. Hultink, y A. Diamantopoulos, “Product intelligence: its conceptualization, measurement and impact on consumer satisfaction”, *J. of the Acad. Mark. Sci.*, vol. 35, núm. 3, pp. 340–356, sep. 2007, doi: 10.1007/s11747-007-0040-6.
- [9] Wirth, R. y Hipp, J., “CRISP-DM: Towards a standard process model for data mining”, en *Proceedings of the 4th International Conference on Practical Applications of Knowledge Discovery and Data Mining*, 2000.
- [10] D. Cer *et al.*, “Universal Sentence Encoder”, el 12 de abril de 2018, *arXiv*: arXiv:1803.11175. doi: 10.48550/arXiv.1803.11175.
- [11] S. Karita *et al.*, “A Comparative Study on Transformer vs RNN in Speech Applications”, en *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, dic. 2019, pp. 449–456. doi: 10.1109/ASRU46091.2019.9003750.
- [12] Chaudhary, A., “Universal sentence encoder visually explained”, Amit Chaudhary Blog. Consultado: el 13 de septiembre de 2025. [En línea]. Disponible en: <https://amitness.com/posts/universal-sentence-encoder/>
- [13] Linková, M. y Gurský, P., “Attributes extraction from product descriptions on e-shops”, en *Proceedings of ITAT 2017*, Hlaváčová, J., 2017, pp. 23–26.

- [14]S. Gallin y A. Portes, “Online shopping: How can algorithm performance expectancy enhance impulse buying?”, *Journal of Retailing and Consumer Services*, vol. 81, p. 103988, nov. 2024, doi: 10.1016/j.jretconser.2024.103988.
- [15]A. Chiche y B. Yitagesu, “Part of speech tagging: a systematic review of deep learning and machine learning approaches”, *J Big Data*, vol. 9, núm. 1, p. 10, ene. 2022, doi: 10.1186/s40537-022-00561-y.
- [16]S. V. Praveen, P. Gajjar, R. K. Ray, y A. Dutt, “Crafting clarity: Leveraging large language models to decode consumer reviews”, *Journal of Retailing and Consumer Services*, vol. 81, p. 103975, nov. 2024, doi: 10.1016/j.jretconser.2024.103975.
- [17]Y. Wang, D. Y. Mo, y M. M. Tseng, “Mapping customer needs to design parameters in the front end of product design by applying deep learning”, *CIRP Annals*, vol. 67, núm. 1, pp. 145–148, 2018, doi: 10.1016/j.cirp.2018.04.018.
- [18]Y. Jiao, Y. Yang, y H. Zhang, “Mapping High Dimensional Sparse Customer Requirements into Product Configurations”, *IOP Conf. Ser.: Mater. Sci. Eng.*, vol. 261, p. 012022, oct. 2017, doi: 10.1088/1757-899X/261/1/012022.
- [19]Y. Liu, Y. Wan, X. Shen, Z. Ye, y J. Wen, “Product Customer Satisfaction Measurement Based on Multiple Online Consumer Review Features”, *Information*, vol. 12, núm. 6, p. 234, may 2021, doi: 10.3390/info12060234.
- [20]Emilio, N., “What is Medallion architecture in a Data Lakehouse context?”
- [21]R. Bandi, J. Amudhavel, y R. Karthik, “Machine Learning with PySpark - Review”, *IJEECS*, vol. 12, núm. 1, p. 102, oct. 2018, doi: 10.11591/ijeecs.v12.i1.pp102-106.
- [22]M. Abadi *et al.*, “TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems”, el 16 de marzo de 2016, *arXiv*: arXiv:1603.04467. doi: 10.48550/arXiv.1603.04467.
- [23]TensorFlow, “TensorFlow Hub”, TensorFlow. Consultado: el 3 de noviembre de 2025. [En línea]. Disponible en: <https://www.tensorflow.org/hub>
- [24]S. Mishra *et al.*, “Principal Component Analysis”, *Int. J. Livest. Res.*, p. 1, 2017, doi: 10.5455/ijlr.20170415115235.
- [25]O. Jafari, P. Maurya, P. Nagarkar, K. M. Islam, y C. Crushev, “A Survey on Locality Sensitive Hashing Algorithms and their Applications”, el 17 de febrero de 2021, *arXiv*: arXiv:2102.08942. doi: 10.48550/arXiv.2102.08942.
- [26]D. P. Kingma y J. Ba, “Adam: A Method for Stochastic Optimization”, el 30 de enero de 2017, *arXiv*: arXiv:1412.6980. doi: 10.48550/arXiv.1412.6980.
- [27]A. Mao, M. Mohri, y Y. Zhong, “Cross-Entropy Loss Functions: Theoretical Analysis and Applications”, el 20 de junio de 2023, *arXiv*: arXiv:2304.07288. doi: 10.48550/arXiv.2304.07288.