



**UNIVERSIDAD ESTATAL PENÍNSULA
DE SANTA ELENA
FACULTAD DE SISTEMAS Y TELECOMUNICACIONES**

TITULO DEL TRABAJO DE TITULACIÓN

SISTEMA AUTOMATIZADO PARA EL ANÁLISIS DE SENTIMIENTOS Y
MINERÍA DE OPINIONES EN RESEÑAS DE PRODUCTOS DE ALIEXPRESS
MEDIANTE ALGORITMOS DE APRENDIZAJE SUPERVISADO Y NO
SUPERVISADO

AUTOR

CRESPO ZAMBRANO ABEL ALEXANDER

EXAMEN COMPLEXIVO

Previo a la obtención del grado académico en
INGENIERO EN TECNOLOGÍAS DE LA INFORMACIÓN

TUTOR

ING. ROSERO VASQUEZ SHENDRY BALMORE, MSC.

Santa Elena, Ecuador

Año 2026



**UNIVERSIDAD ESTATAL PENÍNSULA
DE SANTA ELENA
FACULTAD DE SISTEMAS Y TELECOMUNICACIONES**

TRIBUNAL DE SUSTENTACIÓN



Validar únicamente en FirmaEC.
Firmado electrónicamente por:
**JOSE MIGUEL SANCHEZ
AQUINO**

**Shendry
Rosero**

Firmado digitalmente
por Shendry Rosero
DN: cn=Shendry Rosero
gn=Shendry Rosero c=ES Spain
f=ES Spain
e=shendry.rosero.v@gmail.com
Motivo: Soy el autor de este
documento
Ubicación:
Fecha: 2026-08-19 16:01:05:00

**Ing. José Sánchez Aquino, Mgtr.
DIRECTOR DE LA CARRERA**

**Ing. Rosero Vasquez Shendry, Msc.
TUTOR**



Validar únicamente en FirmaEC.
Firmado electrónicamente por:
**JAIME BENJAMIN
OROZCO IGUASNIA**

**Ing. Jaime Orozco Iguasnia, Mgtr.
DOCENTE ESPECIALISTA**



Validar únicamente en FirmaEC.
Firmado electrónicamente por:
**MARJORIE ALEXANDRA
CORONEL SUAREZ**

**Ing. Marjorie Coronel Suárez, Mgtr.
DOCENTE GUÍA UIC**



**UNIVERSIDAD ESTATAL PENÍNSULA
DE SANTA ELENA
FACULTAD DE SISTEMAS Y TELECOMUNICACIONES**

CERTIFICACIÓN

Certifico que luego de haber dirigido científica y técnicamente el desarrollo y estructura final del trabajo, este cumple y se ajusta a los estándares académicos, razón por el cual apruebo en todas sus partes el presente trabajo de titulación que fue realizado en su totalidad por CRESPO ZAMBRANO ABEL ALEXANDER, como requerimiento para la obtención del título de Ingeniero en Tecnologías de la Información.

La Libertad, a los 19 días del mes de agosto del año 2026

TUTOR

**Shendry
Rosero**

Firmado digitalmente
por Shendry Rosero
DN: cn=Shendry Rosero
gn=Shendry Rosero c=ES
Spain l=ES Spain
e=shendry.rosero.v@gmail.co
m
Motivo: Soy el autor de este
documento
Ubicación:
Fecha: 2026-08-19 16:01:05:00

ING. SHENDRY BALMORE ROSERO VASQUEZ, MSC.



**UNIVERSIDAD ESTATAL PENÍNSULA
DE SANTA ELENA
FACULTAD DE SISTEMAS Y TELECOMUNICACIONES
DECLARACIÓN DE RESPONSABILIDAD**

Yo, Crespo Zambrano Abel Alexander

DECLARO QUE:

El trabajo de Titulación, Sistema Automatizado para el Análisis de Sentimientos y Minería de Opiniones en Reseñas de Productos de AliExpress Mediante Algoritmos de Aprendizaje Supervisado y no Supervisado previo a la obtención del título en Ingeniero en Tecnologías de la Información, ha sido desarrollado respetando derechos intelectuales de terceros conforme las citas que constan en el documento, cuyas fuentes se incorporan en las referencias o bibliografías. Consecuentemente este trabajo es de mi total autoría.

En virtud de esta declaración, me responsabilizo del contenido, veracidad y alcance del Trabajo de Titulación referido.

La Libertad, a los 19 días del mes de agosto del año 2026

EL AUTOR

Abel Alexander Crespo Zambrano



**UNIVERSIDAD ESTATAL PENÍNSULA
DE SANTA ELENA**

FACULTAD DE SISTEMAS Y TELECOMUNICACIONES

CERTIFICACIÓN DE ANTIPLAGIO

Certifico que después de revisar el documento final del trabajo de titulación denominado Sistema Automatizado para el Análisis de Sentimientos y Minería de Opiniones en Reseñas de Productos de AliExpress Mediante Algoritmos de Aprendizaje Supervisado y no Supervisado, presentado por el estudiante, Crespo Zambrano Abel Alexander fue enviado al Sistema Antiplagio, presentando un porcentaje de similitud correspondiente al 6%, por lo que se aprueba el trabajo para que continúe con el proceso de titulación.



Certificado de análisis
Compilatio Magister+ | UPSE-ECU

CRESPO ZAMBRANO ABEL ALEXANDER_compilatio
ID : 47279b670aee5eaf797f31fa6b68a2a7def81ce9



6%
Textos sospechosos

Nombre del fichero : CRESPO ZAMBRANO ABEL ALEXANDER_compilatio.txt
Tamaño del archivo original : 11,52 MB
Número de palabras : 21.931

Depositante : SHENDRY BALMORE ROSERO VASQUEZ
Fecha de depósito : 19 de agosto de 2026
Tipo de carga : interface
fecha de fin de análisis : 19 de agosto de 2026

Resumen (sección 1/2)

Localización de los textos sospechosos en el documento :



TUTOR

Shendry
Rosero

Firmado digitalmente por Shendry
Rosero
DN: cn=Shendry Rosero
gn=Shendry Rosero c=ES, Spain
n=ES, Spain
e=shendry.rosero.v@gmail.com
Motivo Soy el autor de este
documento
Ubicación:
Fecha: 2026-08-19 16:01:05:00

ING. SHENDRY BALMORE ROSERO VASQUEZ, MSC.



**UNIVERSIDAD ESTATAL PENÍNSULA
DE SANTA ELENA
FACULTAD DE SISTEMAS Y TELECOMUNICACIONES**

AUTORIZACIÓN

Yo, CRESPO ZAMBRANO ABEL ALEXANDER

Autorizo a la Universidad Estatal Península de Santa Elena, para que haga de este trabajo de titulación o parte de él, un documento disponible para su lectura consulta y procesos de investigación, según las normas de la Institución.

Cedo los derechos en línea patrimoniales del presente trabajo de titulación con fines de difusión pública, además apruebo la reproducción dentro de las regulaciones de la Universidad, siempre y cuando esta reproducción no suponga una ganancia económica y se realice respetando mis derechos de autor

Santa Elena, a los 19 días del mes de agosto del año 2026

EL AUTOR

Abel Alexander Crespo Zambrano

AGRADECIMIENTO

Agradezco, en primer lugar, a Dios, por brindarme la fortaleza, la salud y la sabiduría necesarias para culminar esta etapa tan importante de mi vida personal y profesional.

A mis padres, por ser el pilar fundamental de mi formación, por su amor incondicional y por apoyarme en cada momento difícil a lo largo de esta carrera. Gracias por enseñarme el valor del esfuerzo y la perseverancia.

A mi tutor, el Ing. Shendry Balmore Rosero, por su orientación, paciencia y acompañamiento profesional durante el desarrollo de este proyecto de titulación. Su guía fue fundamental para alcanzar los resultados obtenidos.

A la Universidad Estatal Península de Santa Elena y a todos los docentes que formaron parte de mi formación académica, por compartir sus conocimientos y experiencias a lo largo de estos años.

Abel Alexander, Crespo Zambrano

DEDICATORIA

Dedico este trabajo de titulación, en primer lugar, a Dios, por guiar cada uno de mis pasos y permitirme llegar hasta este momento.

A mis padres, por cada sacrificio realizado, por creer en mí incluso cuando yo dudé y por ser siempre mi mayor motivación para seguir adelante. Este logro es tan suyo como mío.

Abel Alexander, Crespo Zambrano

ÍNDICE GENERAL

TITULO DEL TRABAJO DE TITULACIÓN	I
TRIBUNAL DE SUSTENTACIÓN	II
CERTIFICACIÓN	III
DECLARACIÓN DE RESPONSABILIDAD	IV
CERTIFICACIÓN DE ANTIPLAGIO	V
AUTORIZACIÓN	VI
AGRADECIMIENTO	VII
DEDICATORIA	VIII
ÍNDICE GENERAL	IX
ÍNDICE DE TABLAS	XIII
ÍNDICE DE FIGURAS	XIV
ÍNDICE DE ANEXOS	XVI
RESUMEN	XVII
ABSTRACT	XVIII
INTRODUCCIÓN	1
CAPÍTULO 1. FUNDAMENTACIÓN	3
1.1 ANTECEDENTES	3
1.2 DESCRIPCIÓN DEL PROYECTO	5
1.3 OBJETIVOS DEL PROYECTO	7
1.3.1 OBJETIVO GENERAL	7
1.3.2 OBJETIVOS ESPECÍFICOS	7
1.4 JUSTIFICACIÓN DEL PROYECTO	8
1.5 ALCANCE DEL PROYECTO	9

CAPÍTULO 2. MARCO TEÓRICO Y METODOLOGÍA DEL PROYECTO	10
2.1 MARCO CONCEPTUAL	10
2.1.1 INTELIGENCIA ARTIFICIAL (IA)	10
2.1.2 PROCESAMIENTO DEL LENGUAJE NATURAL (PLN)	15
2.1.3 EXTRACCIÓN Y GESTIÓN DE DATOS	20
2.1.4 HERRAMIENTAS DE DESARROLLO	23
2.2 MARCO LEGAL	27
2.2.1 LEY ORGÁNICA DE PROTECCIÓN DE DATOS PERSONALES (LOPD)	27
2.2.2 ANONIMIZACIÓN Y PRIVACIDAD DE LOS USUARIOS	27
2.2.3 LEY DE COMERCIO ELECTRÓNICO Y MENSAJES DE DATOS	28
2.2.4 LEY ORGÁNICA DE TELECOMUNICACIONES (LOT)	28
2.2.5 TÉRMINOS DE SERVICIO Y CARÁCTER PÚBLICO DE LOS DATOS	28
2.2.6 USO ACADÉMICO Y SIN FINES COMERCIALES	29
2.3 MARCO TEÓRICO	29
2.3.1 TEORÍA DE LA MINERÍA DE OPINIONES Y SENTIMIENTOS	29
2.3.2 TEORÍA DE LA VALORACIÓN (APPRAISAL THEORY)	30
2.3.3 FUNDAMENTOS DEL PROCESAMIENTO DEL LENGUAJE NATURAL	30
2.3.4 TEORÍA DISTRIBUCIONAL DEL SIGNIFICADO (SEMÁNTICA DISTRIBUCIONAL)	31
2.3.5 TEORÍA DEL APRENDIZAJE ESTADÍSTICO	31
2.3.6 TEORÍA DEL APRENDIZAJE NO SUPERVISADO Y CLUSTERING	32
2.4 METODOLOGÍA DEL PROYECTO	33
2.4.1 METODOLOGÍA DE INVESTIGACIÓN	33
2.4.2 TÉCNICAS E INSTRUMENTOS DE RECOLECCIÓN DE DATOS	33
2.4.3 METODOLOGÍA DE DESARROLLO	34
CAPÍTULO 3. PROPUESTA	36

3.1	REQUERIMIENTOS	36
3.1.1	REQUERIMIENTOS FUNCIONALES	36
3.1.2	REQUERIMIENTOS NO FUNCIONALES	42
3.2	DISEÑO DE LA PROPUESTA	44
3.2.1	ARQUITECTURA GENERAL DEL SISTEMA	45
3.2.2	DIAGRAMA DE FLUJO DEL PIPELINE	46
3.2.3	DIAGRAMAS DE CASOS DE USO	47
3.2.4	DISEÑO DE INTERFACES	48
3.3	DESARROLLO DEL SISTEMA	49
3.3.1	FASE 1 - EXTRACCIÓN DE COMENTARIOS	49
3.3.2	FASE 2 - PREPROCESAMIENTO Y LIMPIEZA DE DATOS	50
3.3.3	FASE 3 - VECTORIZACIÓN DE TEXTO	50
3.3.4	FASE 4 - ENTRENAMIENTO DE MODELOS	51
3.3.5	FASE 5 - CONFIGURACIÓN DE MÉTRICAS DE EVALUACIÓN	52
3.4	DESARROLLO DEL APLICATIVO Y DASHBOARD	54
3.4.1	INTEGRACIÓN DEL MODELO CON LA INTERFAZ PRINCIPAL	54
3.4.2	VISUALIZACIÓN DE RESULTADOS EN EL DASHBOARD	55
3.5	PRUEBAS DEL SISTEMA	60
3.5.1	PRUEBAS FUNCIONALES	60
3.5.2	PRUEBAS DE LOS MODELOS	63
3.6	ANÁLISIS DE RESULTADOS	65
3.6.1	RESULTADOS DE CLASIFICACIÓN SVM	65
3.6.2	RESULTADOS DE AGRUPAMIENTO K-MEANS	67
3.6.3	COMPARACIÓN DE EXPERIMENTOS ORIGINAL VS MEJORADO CON SMOTE	69
3.7	ESTUDIO DE FACTIBILIDAD	71
3.7.1	FACTIBILIDAD TÉCNICA	71
3.7.2	FACTIBILIDAD OPERATIVA	73
	CONCLUSIONES	75

RECOMENDACIONES	77
REFERENCIAS	78
ANEXOS	87

ÍNDICE DE TABLAS

Tabla 1. Requerimientos funcionales – Extracción de comentarios. Elaboración propia.	37
Tabla 2. Requerimientos funcionales - Limpieza de datos. Elaboración propia.	38
Tabla 3. Requerimientos funcionales - Vectorización de texto. Elaboración propia.	39
Tabla 4. Requerimientos funcionales – Entrenamiento de modelos. Elaboración propia.	40
Tabla 5. Requerimientos funcionales - Evaluación y visualización de resultados. Elaboración propia.	42
Tabla 6. Requerimientos funcionales - Pipeline automatizado. Elaboración propia.	42
Tabla 7. Requerimientos no funcionales. Elaboración propia.	44
Tabla 8. Pruebas funcionales ejecutadas sobre el sistema.	63
Tabla 9. Tabla de resultados en las pruebas de los modelos SVM.	64
Tabla 10. Tabla de resultados en las pruebas de los modelos K-Means	65
Tabla 11. Tabla de diferencias entre los dos experimentos.	70
Tabla 12. Herramientas tecnológicas utilizadas en el desarrollo del sistema	73

ÍNDICE DE FIGURAS

Figura 1. Estructura de las Fases del Proyecto.	6
Figura 2. Arquitectura general del sistema de análisis de sentimientos en cinco capas. Elaboración propia.	45
Figura 3. Diagrama de flujo del procesamiento del sistema de análisis de sentimientos. Elaboración propia.	46
Figura 4. Diagrama de caso de uso del sistema de análisis de sentimientos. Elaboración propia.	47
Figura 5. Diseño de la interfaz del sistema de análisis de sentimientos.	48
Figura 6. Fase 1 - Extracción de Comentarios.	49
Figura 7. Fase 2 - Preprocesamiento y Limpieza de Datos.	50
Figura 8. Fase 3 - Vectorización de Texto.	51
Figura 9. Fase 4 - Entrenamiento de Modelos.	52
Figura 10. Fase 5 - Visualización de resultados.	53
Figura 11. Interfaz Principal	55
Figura 12. Pestaña de sentimiento	56
Figura 13. Pestaña de palabras Clave	56
Figura 14. Nube de palabras	57
Figura 15. Pestaña de evaluación SVM	57
Figura 16. Pestaña de k-means	58
Figura 17. Palabras más frecuentes por clúster	58
Figura 18. Pestaña de comparación con gráfico de barras y tabla de métricas acumulativas.	59
Figura 19. Pestaña exportar	59
Figura 20. Comparación de Experimentos Original vs Mejorado con SMOTE	67

Figura 21. Palabras más frecuentes por clúster.	68
Figura 22. PCA 2D	68
Figura 23. Anexo 1. Producto 1: Cámara digital.	87
Figura 24. Anexo 1. Producto 2: Reloj inteligente.	88
Figura 25. Anexo 1. Producto 3: Gafa de sol.	88
Figura 26. Anexo 1. Producto 4: Llave de trinquete inalámbrica	89
Figura 27. Anexo 2. Producto 1: Auriculares inalámbricos.	90
Figura 28. Anexo 2. Producto 2. Robot aspirador.	90
Figura 29. Anexo 2. Producto 3. Cámara deportiva.	91
Figura 30. Anexo 3. Producto 1. Cámara vigilancia wifi	92
Figura 31. Anexo 3. Producto 2. Humidificador de aire.	92

ÍNDICE DE ANEXOS

Anexo 1. Observación de comentarios Negativos relacionados con la baja calidad del producto	87
Anexo 2. Observación de reseñas contradictorias en productos con alto volumen de comentarios	89
Anexo 3. Observación de patrones en comentarios de productos con grandes volúmenes de reseñas	91

RESUMEN

El presente trabajo desarrolla un sistema automatizado de análisis de sentimientos y minería de opiniones en reseñas de productos de AliExpress, implementado como aplicación de escritorio en Python. El sistema integra cinco fases secuenciales: extracción automatizada con Selenium WebDriver, preprocesamiento con NLTK y spaCy, vectorización con Word2Vec y FastText, clasificación supervisada con SVM y agrupamiento no supervisado con K-Means. Se extrajeron 1.358 reseñas del producto seleccionado, de las cuales 544 resultaron válidas tras el preprocesamiento, distribuidas en 509 positivas, 22 negativas y 13 neutras. Se aplicó SMOTE para mitigar el desbalance de clases. Los resultados muestran que FastText con SVM y balanceo alcanzó una Precision de 0.9239 y F1-score de 0.8260, mostrando capacidad de discriminación entre las tres categorías de sentimiento, aunque con un desempeño limitado en las clases minoritarias. Se concluye que las técnicas de balanceo son esenciales para clasificadores entrenados sobre corpus naturalmente desbalanceados como las reseñas de comercio electrónico.

Palabras claves: Análisis de Sentimientos, Aprendizaje Automático, Minería de Opiniones.

ABSTRACT

This work develops an automated sentiment analysis and opinion mining system for AliExpress product reviews, implemented as a Python desktop application. The system integrates five sequential phases: automated extraction with Selenium WebDriver, preprocessing with NLTK and spaCy, vectorization with Word2Vec and FastText, supervised classification with SVM, and unsupervised clustering with K-Means. A total of 1,358 reviews were extracted from the selected product, of which 544 remained valid after preprocessing, distributed into 509 positive, 22 negative, and 13 neutral reviews. SMOTE was applied to mitigate class imbalance. The results show that FastText with SVM and class balancing achieved a Precision of 0.9239 and an F1-score of 0.8260, demonstrating the ability to discriminate among the three sentiment categories, although with limited performance in the minority classes. It is concluded that balancing techniques are essential for classifiers trained on naturally imbalanced corpora such as e-commerce reviews.

Keywords: Sentiment Analysis, Machine Learning, Opinion Mining.

INTRODUCCIÓN

El análisis de sentimientos constituye una técnica fundamental en el procesamiento de lenguaje natural (PLN), orientada a interpretar automáticamente las emociones en grandes volúmenes de texto no estructurado [1]. Actualmente, plataformas de comercio electrónico como AliExpress permiten a millones de usuarios compartir sus experiencias de compra, generando una fuente masiva de datos valiosa para comprender el comportamiento del consumidor. El uso de algoritmos de Machine Learning facilita la automatización de este proceso, permitiendo clasificar sentimientos de forma precisa y eficiente en tres categorías de polaridad: Positivo, Neutro y Negativo [2]. En este sentido, el análisis de las reseñas se convierte en una herramienta estratégica para mejorar la transparencia y la toma de decisiones en el entorno del comercio electrónico.

El problema central de esta investigación radica en la carencia de un análisis sistemático de las reseñas publicadas en AliExpress, lo que dificulta la extracción de información útil tanto para compradores como para vendedores. El inmenso volumen de comentarios no estructurados impide una evaluación objetiva de la calidad de los productos, ocultando patrones críticos de satisfacción o insatisfacción que podrían orientar decisiones de compra [3]. Mediante la aplicación combinada de técnicas de clustering y clasificación supervisada, es posible transformar estos datos textuales en información procesable. Por ello, el objetivo general de este trabajo es desarrollar un sistema automatizado que utilice SVM y K-Means, en combinación con modelos de representación vectorial Word2Vec y FastText, para analizar y agrupar las opiniones de los usuarios de manera efectiva.

Este proyecto se justifica al emplear un enfoque dual que combina algoritmos supervisados y no supervisados para una minería de opiniones integral sobre reseñas en español. El uso de Máquinas de Vectores de Soporte (SVM) permite clasificar cada reseña como positiva, neutra o negativa según la calificación otorgada por el usuario [4]. Por su parte, K-Means facilita el descubrimiento de agrupaciones temáticas latentes relacionadas con aspectos como la calidad o el envío de los productos [5]. La integración de representaciones vectoriales densas como Word2Vec y FastText asegura que el sistema comprenda las relaciones semánticas y morfológicas de los textos, superando las limitaciones de los análisis basados en palabras clave simple [6]. Este conjunto de herramientas permite procesar grandes flujos de información de forma confiable.

La estructura de este trabajo se organiza en tres capítulos que detallan el desarrollo de la solución propuesta. El Capítulo I presenta los antecedentes y el marco referencial sobre el crecimiento del comercio electrónico, además de definir los objetivos, la justificación y el alcance técnico del proyecto. En esta sección se establece la base del problema y la importancia de utilizar técnicas de inteligencia computacional para mitigar la desconfianza en las compras en línea, destacando el papel de las reseñas como factor determinante en la percepción de calidad y en la reputación de los vendedores en plataformas de escala internacional.

El Capítulo II desarrolla el marco conceptual, el marco legal, el marco teórico y la metodología estructurada en cinco fases sistemáticas para la implementación del sistema. Se abordan conceptos clave como el preprocesamiento de texto con NLTK y spaCy, la extracción automatizada de datos con Selenium, la vectorización con Word2Vec y FastText, y la arquitectura de los modelos implementados con Scikit-learn. El Capítulo III presenta la propuesta completa del sistema, incluyendo el diseño arquitectónico, el desarrollo de las cinco fases del pipeline, las pruebas funcionales y de modelos, el análisis de resultados experimentales y el estudio de factibilidad, ofreciendo evidencia concreta del funcionamiento del sistema desarrollado sobre datos reales de AliExpress.

CAPÍTULO 1. FUNDAMENTACIÓN

1.1 ANTECEDENTES

En el comercio electrónico actual, plataformas como AliExpress permiten acceder a una amplia variedad de productos a precios competitivos. En la observación realizada para esta investigación se identificaron comentarios de usuarios que reportan diferencias entre las características descritas en las publicaciones y los productos recibidos, así como experiencias relacionadas con productos defectuosos o que no cumplen con las expectativas. Estos casos se presentan en el Anexo 1.

En la muestra de reseñas analizada también se identificó una considerable variabilidad en las opiniones de los usuarios. Para un mismo producto se encontraron comentarios positivos y negativos, situación que puede dificultar la interpretación de la experiencia general de los compradores. Algunos ejemplos de esta variabilidad se presentan en el Anexo 2.

Cuando un producto concentra un volumen elevado de reseñas, la identificación manual de aspectos recurrentes como calidad, durabilidad o funcionalidad puede resultar más compleja. En el Anexo 3 se presentan ejemplos que evidencian la necesidad de organizar y analizar las opiniones para identificar patrones dentro del conjunto de reseñas.

El trabajo realizado en la Universidad de Bogotá Jorge Tadeo Lozano de Colombia, con el tema “Análisis de sentimientos desde la influencia de Twitter en los procesos de pólizas de seguros del ámbito colombiano”, se centró en analizar cómo las opiniones expresadas en Twitter pueden influir en los procesos de pólizas de seguros en Colombia. Se desarrolló una visualización de datos para analizar el sentimiento de las opiniones en Twitter, utilizando técnicas de aprendizaje automático y procesamiento del lenguaje natural. En un mundo donde las redes sociales desempeñan un papel fundamental en la comunicación cotidiana, se ha vuelto crucial entender cómo las opiniones expresadas en plataformas como Twitter pueden impactar en la toma de decisiones en la industria de seguros [7].

El artículo de la Universidad Nacional de Córdoba de Argentina, con el tema “Análisis de sentimientos en Twitter: el bueno, el malo y el >:(“), propuso un enfoque que combinó

el uso de un clasificador de sentimientos con algoritmos de clustering para analizar la opinión pública en Twitter sobre diversos temas. Se entrenó un clasificador de sentimientos utilizando una base de datos de reseñas de películas y se aplicó el algoritmo de clustering K-Means para identificar temas de discusión y el tono de las opiniones asociadas. Esta metodología permitió una comprensión más profunda de la opinión pública en Twitter sobre diferentes temas de interés [8].

El trabajo realizado en la Universidad César Vallejo de Perú, con el tema “Uso del análisis de sentimientos de reseñas en restaurantes de comidas peruanas”, se centró en el análisis de sentimientos de reseñas. Con el aumento del uso de plataformas web y redes sociales, se ha generado una gran cantidad de datos que pueden ser aprovechados para la toma de decisiones. La investigación adoptó un enfoque cuantitativo y descriptivo, utilizando un diseño no experimental transeccional para recolectar datos en un momento específico. Los resultados revelaron que la mayoría de las reseñas tenían una polaridad positiva, con un porcentaje de 95%. Las emociones predominantes fueron la confianza y la alegría, y las palabras más utilizadas en las reseñas fueron "buena", "ambiente", "lugar", "platos" y "sazón" [9].

En los trabajos citados, se destaca la aplicación de técnicas de análisis de sentimientos en diferentes contextos, desde la influencia de las redes sociales en la industria de seguros hasta la evaluación de opiniones en Twitter sobre diversos temas y el análisis de reseñas de restaurantes.

En este trabajo, se empleará un enfoque integral utilizando Python, combinando web scraping, procesamiento de lenguaje natural (NLP) y modelos de machine learning (SVM y K-Means) con el fin de analizar las reseñas de un producto específico en la plataforma AliExpress y otras plataformas web. La metodología incluye cinco fases: extracción de datos, preprocesamiento, vectorización, aplicación de algoritmos y evaluación con visualizaciones interactivas. Los resultados permitirán identificar patrones ocultos en las opiniones y generar dashboards gráficos para una interpretación intuitiva, contribuyendo a optimizar la experiencia de compra en línea. El proyecto se implementará mediante una aplicación de escritorio desarrollada con Tkinter.

1.2 DESCRIPCIÓN DEL PROYECTO

El presente proyecto consiste en el desarrollo de una aplicación de escritorio en Python orientada al análisis automatizado de reseñas de productos en la plataforma AliExpress, con el propósito de extraer, procesar y analizar las opiniones de los usuarios mediante técnicas de procesamiento de lenguaje natural (NLP) y aprendizaje automático. La solución propuesta permite transformar datos no estructurados en información útil para la toma de decisiones, tanto para compradores como para vendedores.

La aplicación integra un flujo completo de procesamiento que abarca desde la recolección de datos hasta la visualización de resultados, garantizando un análisis eficiente y automatizado. En este contexto, el desarrollo de este sistema se estructura en cinco fases principales:

Primera Fase: Recolección de Información.

- Recolectar una muestra representativa de reseñas de un producto específico en AliExpress, a partir de una URL ingresada, usando técnicas de web scraping con Selenium para capturar el texto y la calificación (estrellas).
- Explorar la estructura de los datos, identificando tipos de datos (texto y emojis).
- Almacenar los datos recolectados en un formato estructurado (CSV).

Segunda Fase: Preprocesamiento de Datos.

- Limpieza de datos para eliminar ruido (caracteres especiales, etiquetas HTML).
- Normalización del texto: conversión a minúsculas, eliminación de stopwords y stemming o lematización para optimizar la calidad del análisis.
- Tratamiento o transformación de emojis y símbolos presentes en las reseñas.

Tercera Fase: Vectorización de Palabras.

- Aplicación de técnicas de Word2Vec y FastText para la transformación de textos en representaciones vectoriales adecuadas para el análisis de sentimientos.

- Se calcula el vector promedio de las palabras en cada comentario para representar cada uno de ellos como un único vector de características, que servirá como entrada para los modelos.

Cuarta Fase: Entrenamiento.

- Implementación del modelo supervisado Máquina de Soporte Vectorial (SVM) para clasificar la polaridad del sentimiento (Positivo/Negativo) usando las estrellas como etiquetas.
- Aplicación del modelo no supervisado Clustering (K-Means) para el agrupamiento de opiniones por similitud semántica, utilizando K=3 como una decisión experimental para explorar la estructura de los datos, sin asumir una correspondencia directa entre los clústeres obtenidos y las categorías Positivo, Neutro y Negativo.

Quinta Fase: Evaluación.

- Validación técnica de los modelos mediante métricas de rendimiento estadístico para el algoritmo SVM y análisis de cohesión para el agrupamiento con K-Means.
- Generación de reportes gráficos interactivos y nubes de palabras integrados en una interfaz de escritorio con Tkinter.

La siguiente figura muestra de forma general las cinco fases del desarrollo del proyecto. Cada una de estas fases está diseñada para abordar un aspecto crítico en el proceso del análisis:

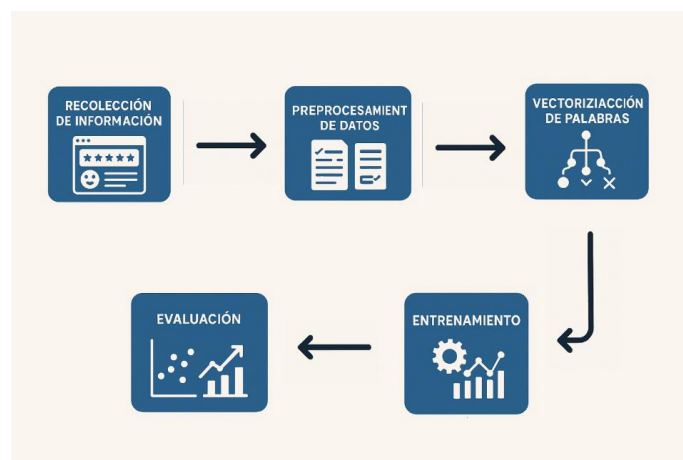


Figura 1. Estructura de las Fases del Proyecto.

El presente trabajo tiene como línea de investigación principal Tecnologías y Sistemas de la Información (TSI) [10] y como Sub-Línea de investigación se detalla en Inteligencia Computacional debido a que se va a aplicar algoritmos de sentimientos para el análisis de los datos.

1.3 OBJETIVOS DEL PROYECTO

1.3.1 OBJETIVO GENERAL

- Desarrollar una aplicación de escritorio que integre técnicas de Machine Learning para el análisis de sentimientos y la agrupación temática de reseñas de productos en AliExpress, permitiendo una interpretación visual de las opiniones de los usuarios para facilitar la toma de decisiones.

1.3.2 OBJETIVOS ESPECÍFICOS

- Recolectar mediante Web Scraping una muestra representativa de reseñas de productos de AliExpress a partir de una URL, extrayendo tanto el texto del comentario como la calificación (estrellas) para su posterior procesamiento.
- Procesar los datos recolectados mediante técnicas de procesamiento de lenguaje natural (NLP) y modelos de representación vectorial (Word2Vec y FastText), para obtener vectores promedio que representen semánticamente cada reseña.
- Implementar el algoritmo SVM para la clasificación de la polaridad de los sentimientos (Positivo/Negativo) y el algoritmo K-Means para el agrupamiento de reseñas por tópicos, con el fin de identificar qué aspectos específicos del producto generan mayor impacto en los usuarios.
- Evaluar el rendimiento de los modelos de aprendizaje automático mediante métricas de validación estadística, garantizando la precisión y confiabilidad de los resultados obtenidos en el análisis de sentimientos y agrupamiento temático.
- Visualizar los resultados a través de una interfaz gráfica en Tkinter que incluya herramientas interactivas y dashboards, proporcionando una experiencia intuitiva para usuarios y vendedores.

1.4 JUSTIFICACIÓN DEL PROYECTO

La expansión del comercio electrónico ha convertido a los marketplaces digitales en canales relevantes para el acceso de los consumidores a una amplia variedad de productos y servicios. Sin embargo, estos entornos también presentan riesgos relacionados con la información proporcionada por los vendedores y la calidad de los productos, así como prácticas comerciales que pueden afectar la confianza del consumidor [11].

Asimismo, las calificaciones y reseñas de otros consumidores constituyen una fuente importante de información para la toma de decisiones de compra, aunque su interpretación puede verse afectada por información incompleta o poco fiable [12]. En la muestra analizada en esta investigación se identificaron comentarios relacionados con problemas de calidad y diferencias entre las características descritas y los productos recibidos (ver Anexo 1). Por ello, surge la necesidad de disponer de herramientas que permitan procesar y analizar automáticamente grandes volúmenes de reseñas para identificar tendencias y patrones de opinión.

Desde una perspectiva técnica, este proyecto se justifica al aplicar un enfoque dual de Inteligencia Computacional. El uso de la Máquina de Soporte Vectorial (SVM) permite clasificar la polaridad de las reseñas, mientras que el algoritmo K-Means facilita el descubrimiento de tópicos latentes en grandes volúmenes de texto. Esta combinación es fundamental para la minería de opiniones, ya que no solo identifica si un sentimiento es Positivo o Negativo, sino que permite interpretar el contexto temático (calidad, envío, funcionalidad) que origina dicha percepción (ver Anexo 2). La implementación de modelos de representación semántica como Word2Vec y FastText asegura que el sistema comprenda las relaciones contextuales de las palabras, superando las limitaciones de los análisis basados en palabras clave tradicionales.

Desde el punto de vista práctico, esta investigación proporciona una solución automatizada que beneficia tanto a compradores como a vendedores. Para los consumidores, optimiza la toma de decisiones al resumir miles de opiniones en información visual digerible; para los vendedores, actúa como una herramienta de auditoría de calidad para identificar fallas en sus productos o logística (ver Anexo 3). En conclusión, el proyecto promueve un entorno de comercio electrónico más transparente

y confiable mediante la integración de técnicas avanzadas de procesamiento de lenguaje natural en una herramienta accesible para usuarios no técnicos.

1.5 ALCANCE DEL PROYECTO

El presente trabajo abarca el diseño, desarrollo e implementación de una aplicación de escritorio funcional basada en Python y la biblioteca Tkinter, diseñada para automatizar el análisis sistemático de reseñas en la plataforma AliExpress. El alcance técnico comprende desde el desarrollo de un motor de extracción de datos en tiempo real mediante Selenium hasta la generación de Dashboards interactivos que faciliten la interpretación de resultados. El sistema está diseñado para procesar reseñas de productos específicas a través de ingreso de una URL directa, garantizando la obtención de datos actualizados que incluyen tanto el texto del comentario como la calificación cuantitativa proporcionada por el usuario.

En cuanto al procesamiento de datos, el alcance incluye un pipeline completo de procesamiento de lenguaje natural que abarca la limpieza de etiquetas HTML, normalización de texto y el tratamiento especializado de emojis para preservar la carga emocional de los mensajes. La transformación de los datos se limita al uso de modelos de representación vectorial densa, específicamente Word2Vec y FastText, calculando promedios vectoriales para representar cada reseña. El modelado analítico se restringe a la aplicación de algoritmo supervisado SVM para la clasificación de polaridad y del algoritmo no supervisado K-Means para la agrupación temática con K=3 grupos, evaluado mediante el índice de Silhouette.

Finalmente, el proyecto culmina con la implementación de visualizaciones dinámicas desarrolladas con Matplotlib. El software permitirá la evaluación del desempeño de los modelos mediante métricas de validación estadística, garantizando la fiabilidad tanto en la clasificación de sentimientos como en la distribución temática de las reseñas. El alcance no incluye la predicción de ventas ni el análisis de perfiles de usuarios, enfocándose estrictamente en la minería de opiniones y el análisis de sentimientos para mejorar la comprensión de la experiencia del cliente en el entorno de AliExpress.

CAPÍTULO 2. MARCO TEÓRICO Y METODOLOGÍA DEL PROYECTO

2.1 MARCO CONCEPTUAL

Este estudio proporciona un contexto integral para comprender los fundamentos y conceptos clave relacionados con el análisis de sentimientos utilizando técnicas de machine learning en textos en español.

2.1.1 INTELIGENCIA ARTIFICIAL (IA)

La inteligencia artificial se define como la rama de la ciencia de la computación encargada de diseñar sistemas capaces de realizar tareas que requieren del intelecto humano, tales como el razonamiento abstracto, la generalización y el aprendizaje autónomo [13]. A lo largo de su evolución tecnológica, esta disciplina se ha consolidado como un pilar fundamental para la automatización analítica y el tratamiento de grandes volúmenes de datos. Su desarrollo abarca múltiples áreas del conocimiento, incluyendo la estadística, la lingüística, la neurociencia y la ingeniería de software, lo que le permite adaptarse a una amplia variedad de dominios de aplicación. Esta tecnología posibilita la interpretación de flujos masivos de datos, transformando registros digitales aislados en conocimiento útil para la toma de decisiones.

2.1.1.1 APRENDIZAJE AUTOMÁTICO (MACHINE LEARNING)

El aprendizaje automático es una subdisciplina de la inteligencia artificial que permite a los sistemas informáticos identificar patrones estadísticos dentro de un conjunto de datos sin ser programados explícitamente para ello [14]. Esta metodología se sustenta en el desarrollo de algoritmos matemáticos que refinan su precisión de forma progresiva a medida que se exponen a mayor volumen de información. Los modelos generados a través de este enfoque son capaces de realizar predicciones o tomar decisiones sobre datos nuevos no observados anteriormente, lo que los convierte en herramientas de gran utilidad para el análisis automatizado de información.

APRENDIZAJE SUPERVISADO

El aprendizaje supervisado es un enfoque de entrenamiento en el que un algoritmo informático procesa muestras de datos previamente etiquetadas, es decir, donde cada

vector de entrada posee una respuesta correcta conocida [15]. El objetivo de este paradigma consiste en aprender una función de correspondencia matemática capaz de predecir la etiqueta de salida para nuevos registros no observados. Este mecanismo es ampliamente utilizado en tareas de clasificación y regresión, donde la disponibilidad de datos etiquetados permite guiar el proceso de aprendizaje de manera estructurada y controlada.

ALGORITMOS DE CLASIFICACIÓN

Los algoritmos de clasificación constituyen un subconjunto de técnicas supervisadas cuyo propósito esencial consiste en asignar de manera automática etiquetas categóricas a elementos de datos [14]. Estos modelos matemáticos evalúan las características numéricas extraídas de un fenómeno observado para determinar la probabilidad de pertenencia a una clase predefinida. Entre los algoritmos más representativos se encuentran las Máquinas de Vectores de Soporte, los árboles de decisión, el clasificador K-NN y la regresión logística, cada uno con características propias que los hacen adecuados para distintos tipos de problemas.

MÁQUINAS DE VECTORES DE SOPORTE (SVM)

Las Máquinas de Vectores de Soporte es un algoritmo supervisado diseñado para resolver problemas de clasificación lineal y no lineal mediante la proyección de datos a espacios multidimensionales [16]. El principio operativo de esta técnica radica en la identificación de un hiperplano de separación óptimo que maximice el margen geométrico entre las muestras de clases distintas en el conjunto de entrenamiento. Los puntos más cercanos a este hiperplano, denominados vectores de soporte son los que determinan su posición exacta. Debido a su robustez matemática ante espacios vectoriales de alta dimensionalidad, este algoritmo destaca en la categorización de documentos de texto, ofreciendo una notable resistencia al fenómeno del sobreajuste.

DESBALANCE DE CLASES

El fenómeno del desbalance de clases se manifiesta cuando una de las categorías objetivo supera de forma desproporcionada en cantidad de muestras a las demás dentro de un

conjunto de datos [17]. Esta asimetría estadística induce un sesgo en el entrenamiento de los modelos predictivos, provocando que los algoritmos optimicen su rendimiento favoreciendo a la clase mayoritaria y descuidando las clases minoritarias. Como consecuencia, las métricas de evaluación globales pueden proyectar un rendimiento aparentemente alto, ocultando un desempeño deficiente sobre las categorías menos representadas.

SMOTE (SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE)

La técnica de sobremuestreo minoritario sintético es un método algorítmico diseñado para mitigar el desbalance de clases mediante la creación artificial de muestras pertenecientes a la categoría menos representada [18]. A diferencia del sobremuestreo tradicional por duplicación, este método opera en el espacio de características numéricas interconectando los vectores de la clase minoritaria mediante el cálculo de sus vecinos más cercanos para generar nuevos ejemplos sintéticos. La aplicación de este procedimiento incrementa la diversidad del conjunto de datos sin incurrir en problemas severos de sobreajuste, permitiendo que los algoritmos de clasificación aprendan fronteras de decisión más equitativas.

APRENDIZAJE NO SUPERVISADO

El aprendizaje no supervisado es un enfoque computacional orientado al análisis de conjuntos de datos que carecen de etiquetas previas o anotaciones humanas explícitas [15]. Bajo este paradigma, los algoritmos asumen la tarea de explorar la estructura intrínseca de la información para descubrir patrones, correlaciones o distribuciones ocultas de manera autónoma. Este enfoque resulta especialmente valioso en escenarios donde se desconoce de antemano la naturaleza exacta de los grupos o categorías presentes en los datos, permitiendo estructurar la información en función de sus similitudes internas sin intervención manual.

ALGORITMOS DE CLUSTERING (AGRUPAMIENTO)

Los algoritmos de agrupamiento o clustering son técnicas del aprendizaje no supervisado orientadas a dividir un conjunto de datos en subgrupos cuyos elementos compartan rasgos

comunes de alta similitud [5]. El criterio de optimización de estas técnicas se basa en maximizar la cohesión interna de cada grupo, al mismo tiempo que se maximiza la separación geométrica respecto a los demás conjuntos identificados. Este tipo de análisis es especialmente útil en la exploración de datos, ya que facilita la identificación de tendencias, segmentos y comportamientos comunes dentro de grandes volúmenes de información.

ALGORITMO K-MEANS

El algoritmo K-Means es un método de agrupamiento particional que divide un conjunto de datos en un número predefinido de agrupaciones denotado por la variable matemática K [19]. El proceso opera de manera iterativa calculando la distancia euclidiana entre cada vector de datos y los centroides asignados provisionalmente, reasignando cada elemento al grupo cuyo centro geométrico resulte más cercano. El algoritmo repite este ciclo de optimización hasta alcanzar la convergencia, garantizando una configuración donde la suma de los errores cuadráticos internos dentro de cada clúster se reduzca al mínimo posible.

MÉTRICAS DE EVALUACIÓN DE MODELOS

Las métricas de evaluación son indicadores estadísticos empleados para cuantificar con precisión el rendimiento, la robustez y la capacidad de generalización de los modelos predictivos entrenados [14]. El uso riguroso de estas herramientas permite comparar de manera objetiva diferentes arquitecturas algorítmicas bajo condiciones experimentales controladas. La validación estadística de los modelos constituye una fase fundamental en cualquier proceso de machine learning, ya que garantiza que los resultados generados posean niveles de confiabilidad adecuados antes de ser utilizados en entornos de aplicación real.

MATRIZ DE CONFUSIÓN

La matriz de confusión es una herramienta bidimensional de validación que expone visualmente la concordancia entre las clasificaciones reales de un conjunto de prueba y las predicciones emitidas por el algoritmo supervisado [15]. A través de su estructura

matricial, permite identificar la frecuencia de verdaderos Positivos, falsos Positivos, Verdaderos Negativos y Falsos Negativos generados durante la evaluación. Esta disposición analítica resulta superior a la evaluación por precisión global, dado que desglosa la naturaleza exacta de los errores cometidos por el modelo ante cada categoría del sistema.

PRECISIÓN

La precisión se define matemáticamente como la proporción de predicciones Positivas emitidas por el modelo de aprendizaje que corresponden efectivamente a casos clasificados de manera correcta en la realidad [14]. Esta métrica actúa como un indicador de la calidad del algoritmo, midiendo el impacto de los errores por falsos Positivos dentro del sistema. Un valor alto de precisión indica que cuando el modelo predice una categoría determinada, la probabilidad de que esa predicción sea correcta es elevada.

EXHAUSTIVIDAD (RECALL)

La exhaustividad, técnicamente denominada sensibilidad o recall, expresa la capacidad de un modelo predictivo para identificar la totalidad de los casos Positivos reales presentes dentro de un conjunto de datos analizado [15]. Este indicador estadístico evalúa la tasa de omisión del algoritmo, midiendo el impacto Negativo de los errores clasificados como Falsos Negativos. Un valor elevado de exhaustividad garantiza que el sistema no pase por alto una proporción significativa de instancias pertenecientes a una clase objetivo.

F1-SCORE

El indicador F1-Score representa la media armónica ponderada entre las métricas de precisión y exhaustividad, consolidándose como un estándar de evaluación integral para medir el desempeño de algoritmos de clasificación [14]. Esta métrica resulta particularmente idónea en escenarios caracterizados por la presencia de un marcado desbalance de clases, donde la precisión global tiende a proyectar un rendimiento engañoso. Al exigir un equilibrio estricto entre la calidad de la predicción y el nivel de

cobertura del modelo, proporciona una aproximación fidedigna de la efectividad general del algoritmo evaluado.

COEFICIENTE DE SILHOUETTE

El coeficiente de Silhouette es una métrica de validación empleada en el aprendizaje no supervisado para medir la calidad geométrica y la consistencia de los agrupamientos generados por algoritmos de clustering [20]. Este indicador computa de forma simultánea la distancia media intra-clúster, denominada cohesión, y la distancia media hacia el clúster más cercano, denominada separación, devolviendo un valor acotado entre menos uno y uno. Un coeficiente próximo a uno certifica que los elementos analizados se encuentran correctamente asignados a grupos compactos y claramente diferenciados entre sí.

2.1.2 PROCESAMIENTO DEL LENGUAJE NATURAL (PLN)

El procesamiento del lenguaje natural es una disciplina interdisciplinaria situada entre la ciencia de la computación, la inteligencia artificial y la lingüística formal, cuyo propósito es permitir que las máquinas comprendan, interpreten y generen el lenguaje humano [21]. Esta tecnología aborda la complejidad intrínseca de las lenguas naturales, caracterizadas por la ambigüedad conceptual, el uso de modismos y las variaciones sintácticas propias de cada idioma. Su funcionamiento implica etapas clave como el preprocesamiento del texto, la representación vectorial y la aplicación de algoritmos capaces de identificar patrones y significados dentro del lenguaje escrito.

2.1.2.1 FUNDAMENTOS LINGÜÍSTICOS

ANÁLISIS SEMÁNTICO

El análisis semántico constituye la fase avanzada del procesamiento lingüístico enfocada en descifrar el significado, el contexto y la intención comunicativa de las estructuras gramaticales, más allá de la simple inspección sintáctica [21]. Esta disciplina evalúa la interacción de los componentes léxicos dentro de un enunciado para inferir la carga

conceptual transmitida por el emisor. A través del análisis semántico, los sistemas de procesamiento de lenguaje natural logran transitar desde la lectura mecánica de caracteres hacia la comprensión contextual del contenido expresado.

LINGÜÍSTICA SEMÁNTICA

La lingüística semántica representa el estudio formal de las relaciones existentes entre las expresiones lingüísticas y los significados abstractos que los hablantes atribuyen a dichos componentes de comunicación [22]. Esta rama del conocimiento examina cómo el orden estructural y la combinación de las unidades léxicas determinan la interpretación final de un mensaje. Su comprensión teórica resulta esencial para el desarrollo de algoritmos informáticos orientados a la minería de textos, proveyendo los modelos lógicos necesarios para representar el sentido de las palabras en formatos computables.

ANÁLISIS DE SENTIMIENTOS

El análisis de sentimientos es una metodología computacional enmarcada en el PLN dedicada a la identificación, extracción y categorización automatizada de las opiniones y estados afectivos expresados en un texto digital [1]. Esta técnica procesa el lenguaje natural para discernir la actitud subjetiva de un autor respecto a un tópico, servicio o producto específico. Mediante su implementación, es posible auditar grandes volúmenes de comentarios de manera automática, transformando las percepciones humanas cualitativas en indicadores cuantitativos accionables.

POLARIDAD

La polaridad es un concepto fundamental dentro del análisis de sentimientos que permite identificar la orientación emocional de un texto, clasificándolo como Positivo, Negativo o Neutro [1]. Este enfoque no solo ayuda a interpretar opiniones, sino también a medir la intensidad con la que se expresan, asignando valores que facilitan su análisis cuantitativo. La determinación precisa de la polaridad requiere que los modelos algorítmicos interpreten correctamente modificadores léxicos como adverbios de negación y adjetivos intensificadores, los cuales alteran el sentido general del mensaje.

MINERÍA DE OPINIONES

La minería de opiniones es un campo de estudio tecnológico estrechamente vinculado al análisis de sentimientos, cuyo alcance se extiende hacia la extracción exhaustiva de juicios de valor, atributos específicos y entidades mencionadas en textos [1]. Mientras el análisis de sentimientos determina la carga emocional general de un texto, la minería de opiniones busca desagregar qué características específicas de un objeto generan dicha reacción en el usuario. Esta metodología permite estructurar resúmenes informativos detallados que facilitan la evaluación de aspectos concretos de un producto o servicio.

DETECCIÓN DE TÓPICOS

La detección de tópicos es una técnica automatizada de la minería de textos orientada a identificar de forma estadística los temas latentes o hilos conductores de discusión recurrentes dentro de un corpus documental extenso [5]. Esta aproximación asume que cada documento se compone de una mezcla de palabras clave cuyas frecuencias de coocurrencia revelan un concepto subyacente. Su integración con técnicas de agrupamiento permite organizar grandes volúmenes de textos en categorías temáticas estables, facilitando la comprensión del contenido sin necesidad de leer cada documento de forma individual.

PREPROCESAMIENTO DE TEXTO

El preprocesamiento de texto es la etapa inicial y fundamental del pipeline de procesamiento del lenguaje natural, enfocada en la limpieza, estandarización y acondicionamiento de las cadenas de caracteres antes de ser introducidas en modelos de aprendizaje [23]. Los textos provenientes de medios digitales suelen carecer de estructura uniforme y contener elementos irrelevantes que dificultan el análisis. La correcta ejecución de esta fase depura la información textual, reduce la dimensionalidad del espacio de características y potencia la eficiencia computacional de los algoritmos de clasificación y agrupamiento posteriores.

TOKENIZACIÓN

La tokenización es el proceso de división de una cadena continua de texto en unidades discretas denominadas tokens, las cuales corresponden generalmente a palabras

individuales o términos compuestos [21]. Este procedimiento descompone la estructura sintáctica del texto eliminando signos de puntuación y espacios redundantes, obteniendo una lista ordenada de unidades léxicas procesables. La obtención de una lista limpia de tokens constituye la base obligatoria sobre la cual operan los análisis morfológicos y las conversiones vectoriales en cualquier arquitectura de procesamiento de lenguaje natural.

NORMALIZACIÓN DE TEXTO

La normalización de texto engloba el conjunto de operaciones orientadas a estandarizar los tokens obtenidos para eliminar variaciones ortográficas superficiales que introducen dispersión artificial en los datos [23]. Este proceso incluye la conversión de todos los caracteres a minúsculas, la remoción de acentos gráficos, la eliminación de caracteres especiales y la corrección de errores de escritura comunes. Al unificar los términos bajo una única representación formal, se evita que un mismo concepto sea interpretado por el algoritmo como múltiples variables independientes debido a discrepancias ortográficas.

ELIMINACIÓN DE STOPWORDS

La eliminación de stopwords consiste en el filtrado y descarte de aquellas palabras de uso extremadamente frecuente en un idioma que carecen de carga semántica sustancial para tareas de clasificación o agrupamiento [21]. Este grupo abarca principalmente artículos, preposiciones, conjunciones y ciertos verbos auxiliares que actúan como conectores gramaticales necesarios en el lenguaje cotidiano, pero que no aportan información relevante al análisis de contenido. La remoción de estos términos permite que los algoritmos concentren sus capacidades de cómputo en las palabras con mayor valor semántico, donde reside el núcleo del mensaje.

LEMATIZACIÓN

La lematización es un proceso morfológico avanzado que consiste en transformar cada palabra analizada a su forma base canónica o raíz lingüística formal, denominada técnicamente lema [23]. A diferencia del stemming, que corta los sufijos de manera heurística sin considerar el idioma, la lematización utiliza diccionarios léxicos

estructurados para garantizar que las inflexiones verbales regresen al infinitivo y los sustantivos a su forma singular. Esta sofisticación lingüística preserva el contexto semántico de los términos, agrupando diferentes variantes morfológicas bajo un único identificador conceptual.

TRANSFORMACIÓN DE EMOJIS

La transformación de emojis es un procedimiento de adaptación técnica enfocado en traducir los ideogramas gráficos presentes en textos digitales en cadenas textuales de carácter semántico equivalente [24]. En la comunicación digital contemporánea, los usuarios emplean estos símbolos visuales de forma masiva para condensar o amplificar estados de ánimo que complementan o sustituyen al texto alfabético. Al mapear cada emoji hacia su descripción textual precisa, se preserva la carga emocional expresada visualmente, garantizando que esta información sea procesable por los modelos de análisis de lenguaje natural.

REPRESENTACIÓN VECTORIAL (WORD EMBEDDING)

La representación vectorial de palabras, técnicamente denominada word embedding, comprende el conjunto de metodologías en PLN destinadas a mapear términos lingüísticos hacia vectores de números reales dentro de un espacio continuo multidimensional [25]. La característica fundamental de estas aproximaciones radica en que palabras que comparten contextos semánticos similares se ubican en posiciones geométricamente cercanas dentro del espacio vectorial resultante. Esta codificación numérica densa permite superar las limitaciones de los métodos tradicionales basados en matrices dispersas de alta dimensionalidad, otorgando a los modelos la capacidad de operar algebraicamente sobre conceptos abstractos.

WORD2VEC

Word2Vec es un modelo de representación vectorial basado en redes neuronales de baja profundidad, diseñado para aprender representaciones de palabras a partir de grandes corpus textuales sin supervisión [25]. Desarrollado mediante las arquitecturas de bolsa de

palabras continua y skip-gram, este algoritmo predice la aparición de términos basándose en las palabras del contexto circundante de cada oración. Su implementación captura relaciones semánticas complejas y analogías lingüísticas, generando vectores que reflejan la similitud de significado entre los términos del vocabulario analizado.

FASTTEXT

FastText es una arquitectura avanzada para el aprendizaje de embeddings de palabras desarrollada por Facebook AI Research, cuya innovación radica en considerar la estructura interna de los términos dividiéndolos en subpalabras o n-gramas de caracteres [6]. A diferencia de modelos como Word2Vec, FastText genera vectores de alta calidad incluso para términos poco comunes o con errores tipográficos, heredando las características semánticas de las raíces compartidas entre palabras relacionadas. Esta propiedad resulta especialmente útil para procesar textos informales provenientes de plataformas digitales, donde la presencia de neologismos y variaciones ortográficas es frecuente.

VECTOR PROMEDIO DE PALABRAS

El vector promedio de palabras es una técnica de agregación matemática empleada para condensar las representaciones vectoriales individuales de todos los tokens de un documento en un único vector de características [25]. Este cálculo computa la media aritmética de las coordenadas de cada embedding correspondiente a las palabras válidas presentes en un texto. El vector resultante sintetiza de manera homogénea la semántica global del documento analizado, generando una estructura numérica compacta y estandarizada apta para ser procesada por algoritmos de clasificación y agrupamiento.

2.1.3 EXTRACCIÓN Y GESTIÓN DE DATOS

2.1.3.1 WEB SCRAPING

El web scraping comprende el conjunto de técnicas informáticas orientadas a automatizar la extracción de datos e información alojada en páginas web mediante el uso de scripts de software [26]. Esta metodología simula el comportamiento de un usuario humano al

navegar por la red, permitiendo recopilar contenidos no estructurados en formato HTML para convertirlos en registros organizados y analizables. Su aplicación es frecuente en proyectos de investigación y análisis de datos, especialmente cuando las plataformas de interés no ofrecen interfaces de programación abiertas para el acceso directo a su información.

AUTOMATIZACIÓN DEL NAVEGADOR

La automatización del navegador es un enfoque avanzado dentro de la extracción de datos web, diseñado para interactuar directamente con elementos de interfaz dinámicos que se ejecutan del lado del cliente mediante lenguajes de scripting [26]. A diferencia de los métodos de petición estáticos, esta aproximación simula acciones de navegación complejas como clics, desplazamientos y esperas asíncronas de carga. Este mecanismo resulta necesario en plataformas cuyas arquitecturas de renderizado dinámico presentan el contenido únicamente después de la ejecución de scripts del navegador.

SELENIUM WEBDRIVER

Selenium WebDriver es un marco de automatización de código abierto diseñado para ejecutar instrucciones de control directo sobre los motores de renderizado de diversos navegadores web [27]. Esta herramienta actúa como un puente de comunicación que permite a lenguajes de programación como Python interactuar con el árbol de objetos del documento de un sitio en tiempo real. Su capacidad para gestionar eventos asíncronos y emular comportamientos de navegación humanos la convierte en una de las herramientas más utilizadas para la extracción de datos en plataformas de comercio electrónico con contenido dinámico.

MODO HEADLESS

El modo headless es una modalidad operativa de ejecución de navegadores web en la que el software realiza el renderizado completo de las páginas en segundo plano, prescindiendo de la visualización de ventanas gráficas en pantalla [27]. Esta configuración optimiza el uso de recursos de memoria y tiempo de procesamiento del sistema operativo, permitiendo acelerar los ciclos de extracción y automatización de

datos. Su uso es estándar en entornos de producción y flujos de trabajo automatizados donde no se requiere la interacción visual directa del usuario.

CHROMEDRIVER

ChromeDriver es el componente ejecutable provisto por el proyecto Chromium que permite a Selenium WebDriver dirigir los procesos internos del navegador Google Chrome [27]. Este controlador traduce las instrucciones lógicas escritas en código de programación en comandos nativos comprensibles para el motor del navegador. Para garantizar la estabilidad operativa de la automatización, es un requisito técnico estricto que la versión del ChromeDriver instalado sea compatible con la versión del navegador disponible en el sistema operativo.

ESTRUCTURAS DE DATOS

Las estructuras de datos son los formatos de organización, gestión y almacenamiento lógico de la información empleados en las ciencias de la computación para facilitar el acceso y la modificación eficiente de variables [13]. La selección adecuada de estas estructuras condiciona directamente el rendimiento de los algoritmos que operan sobre los datos almacenados. En el procesamiento de texto, la correcta transición de la información a través de distintas estructuras garantiza la integridad de los datos a lo largo de todo el flujo de análisis.

DATOS ESTRUCTURADOS Y NO ESTRUCTURADOS

Los datos estructurados corresponden a aquella información que posee un formato altamente organizado y predefinido, facilitando su almacenamiento en tablas relacionales compuestas por filas y columnas fijas [26]. En contraposición, los datos no estructurados carecen de una arquitectura interna regular, manifestándose típicamente en forma de textos libres, imágenes o código HTML. Una de las funciones principales del procesamiento del lenguaje natural es actuar como puente entre ambas categorías, transformando texto libre no estructurado en variables cuantitativas organizadas y procesables por algoritmos matemáticos.

CONJUNTO DE DATOS (DATASET)

Un conjunto de datos o dataset es la colección organizada de registros recopilados sistemáticamente a partir de un fenómeno de estudio específico para su posterior análisis computacional [14]. La calidad, diversidad léxica y tamaño representativo de un dataset condicionan directamente la precisión matemática alcanzada por los algoritmos predictivos durante su entrenamiento. En el ámbito del aprendizaje automático, contar con un conjunto de datos suficientemente amplio y representativo es un requisito fundamental para garantizar la validez de los modelos generados.

FORMATO CSV Y JSON

El formato CSV es un estándar de almacenamiento basado en archivos de texto plano donde los valores de las variables se organizan mediante la delimitación por caracteres de coma, facilitando su importación en herramientas de análisis de datos [26]. Por su parte, el formato JSON es un estándar ligero de intercambio de datos estructurado en pares de clave y valor, que ofrece mayor flexibilidad para representar colecciones de objetos complejos con atributos anidados. Ambos formatos son ampliamente adoptados en el ámbito de la ciencia de datos por su universalidad, ligereza de almacenamiento y compatibilidad con múltiples lenguajes de programación.

2.1.4 HERRAMIENTAS DE DESARROLLO

2.1.4.1 LENGUAJE DE PROGRAMACIÓN PYTHON

Python es un lenguaje de programación de alto nivel, interpretado y multiparadigma, cuya filosofía de diseño prioriza la legibilidad del código y la eficiencia del desarrollo [28]. Su arquitectura soporta de manera nativa la orientación a objetos, la programación funcional y la manipulación avanzada de estructuras de datos. Debido a su extenso ecosistema de librerías científicas y a la madurez de su comunidad global, Python se ha consolidado como el estándar tecnológico preponderante para la implementación de proyectos en inteligencia computacional, ciencia de datos y desarrollo de aplicaciones de escritorio.

LIBRERÍAS DE PROCESAMIENTO Y ANÁLISIS

NLTK

La biblioteca NLTK es una plataforma líder de código abierto desarrollada en Python para el procesamiento del lenguaje natural [21]. Proporciona una colección extensa de algoritmos lingüísticos fundamentales, módulos de normalización gramatical y corpus de palabras predefinidos para múltiples idiomas. Su enfoque didáctico y su amplia documentación la hacen adecuada tanto para tareas de investigación como para el desarrollo de soluciones de análisis de texto en entornos académicos y profesionales.

SPACY

spaCy es una librería de procesamiento del lenguaje natural diseñada bajo criterios de alta velocidad de ejecución y eficiencia en entornos de producción [23]. Implementa modelos estadísticos preentrenados de alta precisión para ejecutar tareas complejas de lematización, análisis morfosintáctico y reconocimiento de entidades nombradas. Sus modelos disponibles para el idioma español permiten aplicar procesos lingüísticos avanzados sobre textos de manera rápida y con resultados de alta calidad.

GENSIM (WORD2VEC / FASTTEXT)

Gensim es una biblioteca especializada en Python orientada al modelado de espacios vectoriales y al análisis automatizado de similitud de documentos textuales [25]. Destaca por su capacidad para procesar grandes flujos de información sin saturar la memoria del sistema, implementando de manera eficiente los algoritmos de representación vectorial Word2Vec y FastText. Su diseño permite trabajar con corpus de distintos tamaños y entrenar modelos semánticos de alta calidad de forma accesible.

NUMPY

NumPy es la librería científica fundamental de Python diseñada para el cómputo numérico de alta eficiencia mediante el uso de estructuras de arreglos multidimensionales homogéneos de memoria contigua [28]. Proporciona una colección de funciones matemáticas avanzadas para realizar operaciones de álgebra lineal, transformadas estadísticas y manipulaciones de matrices complejas. Los algoritmos de aprendizaje

automático emplean de forma nativa los arreglos de NumPy para el procesamiento matemático de los vectores de características que representan los datos.

PANDAS

Pandas es una biblioteca de Python diseñada para la manipulación, limpieza y análisis de estructuras de datos bidimensionales mediante el uso de objetos denominados DataFrames [28]. Permite gestionar de manera intuitiva conjuntos de datos tabulares, facilitando operaciones complejas de filtrado, indexación y transformación de variables. Su integración con otras librerías científicas la convierte en una herramienta fundamental dentro del flujo de trabajo del análisis de datos.

LIBRERÍAS DE APRENDIZAJE AUTOMÁTICO

SCIKIT-LEARN

Scikit-learn es la biblioteca de referencia principal para la implementación de algoritmos de aprendizaje automático tradicional en Python [29]. Unifica bajo una interfaz coherente herramientas para la clasificación, regresión, agrupamiento y validación estadística de modelos predictivos. Construida sobre la base tecnológica de NumPy y Pandas, garantiza una implementación eficiente de algoritmos ampliamente reconocidos en la literatura científica, incluyendo las Máquinas de Vectores de Soporte y el método de agrupamiento K-Means.

IMBALANCED-LEARN (SMOTE)

Imbalanced-learn es una extensión especializada del ecosistema de Scikit-learn enfocada en proveer algoritmos para el tratamiento del desbalance de clases en aprendizaje automático [18]. Ofrece implementaciones optimizadas de diversas técnicas de sobremuestreo y submuestreo estadístico, incluyendo el algoritmo SMOTE para la generación de muestras sintéticas de la clase minoritaria. Su diseño es compatible con la interfaz estándar de Scikit-learn, lo que facilita su integración en flujos de trabajo de machine learning existentes.

LIBRERÍAS DE VISUALIZACIÓN E INTERFAZ

MATPLOTLIB

Matplotlib es una librería de base fundamental en Python diseñada para la creación de visualizaciones estáticas, animadas e interactivas de alta calidad [28]. Proporciona un control detallado sobre cada componente geométrico del gráfico, permitiendo personalizar ejes, colores, leyendas y formatos de presentación. Su versatilidad permite generar desde gráficos simples de barras y dispersión hasta visualizaciones complejas embebidas en interfaces gráficas de usuario, y su compatibilidad con otras librerías científicas la consolida como un estándar en el ámbito de la ciencia de datos.

WORDCLOUD

La librería WordCloud es una herramienta gráfica especializada orientada a la generación automatizada de nubes de palabras, donde el tamaño visual de cada término refleja proporcionalmente su frecuencia estadística dentro de un corpus de texto [23]. Esta visualización permite identificar de manera intuitiva los términos más recurrentes dentro de un conjunto de documentos, facilitando la comprensión rápida del contenido predominante. Su uso es frecuente en tareas de análisis exploratorio de texto, donde se busca obtener una visión general de los temas más discutidos sin necesidad de revisar cada documento individualmente.

TKINTER

Tkinter es la biblioteca estándar de Python para el desarrollo de interfaces gráficas de usuario, integrada de forma nativa en la distribución oficial del lenguaje sin requerir instalaciones adicionales [30]. Actúa como un envoltorio sobre el kit de herramientas de ventanas Tcl/Tk, permitiendo diseñar aplicaciones de escritorio funcionales y multiplataforma compatibles con los principales sistemas operativos. Con Tkinter es posible construir ventanas, botones, menús, cuadros de texto y otros componentes visuales mediante una API accesible, lo que la convierte en una opción adecuada para el desarrollo de interfaces de escritorio en entornos académicos y profesionales.

2.2 MARCO LEGAL

El presente trabajo de investigación se desarrolla considerando el marco normativo aplicable en materia de protección de datos personales, privacidad de los usuarios y uso ético de la información en entornos digitales. Debido a la naturaleza automatizada del sistema propuesto, que incluye la extracción de reseñas mediante web scraping, la vectorización de texto y la clasificación de sentimientos, se consideran las normas jurídicas y condiciones de uso aplicables en la legislación ecuatoriana y en las plataformas comerciales utilizadas. A continuación, se presentan los principales instrumentos normativos y condiciones de uso que sirven como referencia para el diseño y desarrollo del sistema propuesto.

2.2.1 LEY ORGÁNICA DE PROTECCIÓN DE DATOS PERSONALES (LOPD)

En el año 2021, el Estado ecuatoriano promulgó la Ley Orgánica de Protección de Datos Personales, la cual establece el marco legal para la recolección, tratamiento y protección de la información personal de los ciudadanos [31]. Esta normativa establece los principios y derechos relacionados con la protección de datos personales y las obligaciones de las entidades que los tratan. En esta investigación, dichos principios se consideran durante el tratamiento automatizado de la información. El sistema fue diseñado para no almacenar información personal identificable de los usuarios de AliExpress durante ninguna etapa del procesamiento.

2.2.2 ANONIMIZACIÓN Y PRIVACIDAD DE LOS USUARIOS

En cumplimiento con las disposiciones de la LOPDP, la aplicación de escritorio desarrollada integra por diseño el principio de minimización de datos durante la fase de extracción ejecutada mediante Selenium WebDriver [31]. El módulo de web scraping está programado para capturar exclusivamente el texto de la reseña y la calificación numérica en estrellas otorgada por el comprador, descartando automáticamente cualquier dato identificable como nombres de usuario, correos electrónicos o ubicaciones geográficas. Este enfoque garantiza que las cinco fases del sistema operen únicamente sobre contenido

semántico anónimo. La aplicación no genera ni almacena ningún archivo que vincule una opinión procesada con la identidad real de su autor.

2.2.3 LEY DE COMERCIO ELECTRÓNICO Y MENSAJES DE DATOS

Las actividades desarrolladas por el sistema implican el tratamiento de información digital obtenida desde plataformas de comercio electrónico. En este contexto, la Ley de Comercio Electrónico, Firmas Electrónicas y Mensajes de Datos establece el reconocimiento jurídico de la información generada y transmitida por medios electrónicos, proporcionando el marco normativo para las operaciones realizadas sobre documentos digitales durante el desarrollo de la investigación [32].

2.2.4 LEY ORGÁNICA DE TELECOMUNICACIONES (LOT)

La ejecución del módulo de extracción automatizada requiere establecer comunicaciones mediante Internet con servidores remotos. Estas comunicaciones se realizan utilizando los servicios de telecomunicaciones regulados por la Ley Orgánica de Telecomunicaciones del Ecuador, la cual establece el marco general para la prestación y utilización de dichos servicios en el territorio nacional. Durante la ejecución del sistema se implementan mecanismos de optimización del tráfico de red, como la desactivación de la carga de imágenes y hojas de estilo, con el fin de reducir el consumo de ancho de banda durante el proceso de extracción [33].

2.2.5 TÉRMINOS DE SERVICIO Y CARÁCTER PÚBLICO DE LOS DATOS

Para operar de forma responsable sobre plataformas de terceros, el sistema considera las condiciones establecidas en los Términos de Uso de AliExpress respecto al acceso y uso de la información disponible en la plataforma [34]. La aplicación desarrollada interactúa únicamente con reseñas y valoraciones publicadas de forma pública por los usuarios y visibles sin necesidad de autenticación. El proceso de extracción no utiliza credenciales de acceso, no intenta evadir mecanismos de autenticación ni accede a secciones restringidas del sitio web. Estas características limitan la recolección de información al contenido públicamente accesible y buscan mantener el proceso dentro del alcance funcional previsto para la investigación académica.

2.2.6 USO ACADÉMICO Y SIN FINES COMERCIALES

El desarrollo y la ejecución del sistema automatizado de análisis de sentimientos tienen una finalidad exclusivamente académica y de investigación, sin fines comerciales ni intención de redistribuir la información obtenida [35]. En el ámbito internacional, la Organización Mundial de la Propiedad Intelectual (OMPI/WIPO) promueve el análisis y discusión de las limitaciones y excepciones al derecho de autor relacionadas con actividades de investigación, educación y minería de textos y datos, aspectos abordados por el Standing Committee on Copyright and Related Rights (SCCR) [35].

Los datos recopilados mediante el sistema se utilizan únicamente para entrenar y evaluar los modelos de aprendizaje automático implementados, como Word2Vec, FastText, SVM y K-Means. La información obtenida no se redistribuye ni se emplea con fines comerciales, sino exclusivamente para validar la metodología propuesta y los resultados del sistema desarrollado.

2.3 MARCO TEÓRICO

El presente marco teórico reúne los fundamentos conceptuales, científicos y metodológicos que sustentan el desarrollo del sistema automatizado de análisis de sentimientos y minería de opiniones en reseñas de AliExpress. Las teorías seleccionadas proporcionan el sustento conceptual para las decisiones técnicas implementadas durante el desarrollo de las cinco fases del sistema. Su comprensión permite entender por qué las reseñas de usuarios contienen polaridad expresable, por qué los modelos vectoriales capturan el significado de las palabras, y por qué los algoritmos SVM y K-Means son adecuados para clasificar y agrupar opiniones de forma automatizada.

2.3.1 TEORÍA DE LA MINERÍA DE OPINIONES Y SENTIMIENTOS

La minería de opiniones o análisis de sentimientos constituye un área de investigación del procesamiento del lenguaje natural dedicada a la identificación y clasificación automática de opiniones expresadas en textos. Bing Liu consolidó sus fundamentos conceptuales mediante la sistematización de los principales conceptos, componentes y niveles de análisis de esta disciplina [1].

Este marco teórico define conceptos como la polaridad del sentimiento, el titular de la opinión, el objeto evaluado y los atributos específicos sobre los cuales se emite un juicio de valor. Asimismo, distingue el análisis de sentimientos a nivel de documento, oración y aspecto, proporcionando una taxonomía estructurada para abordar la complejidad del lenguaje opinativo [1].

Esta teoría proporciona el fundamento conceptual para el análisis y clasificación automática de las reseñas empleadas en el sistema desarrollado, ya que sustenta la clasificación de las reseñas en las categorías positiva, neutra y negativa a partir de las valoraciones otorgadas por los compradores [1].

2.3.2 TEORÍA DE LA VALORACIÓN (APPRAISAL THEORY)

La teoría de la valoración, desarrollada por Martin y White, proporciona un marco lingüístico sistemático para analizar cómo los hablantes expresan actitudes, juicios y valoraciones a través del lenguaje escrito en contextos comunicativos reales [36].

Esta teoría identifica tres sistemas principales de evaluación dentro del discurso: la actitud, referida a los sentimientos y juicios del hablante; el compromiso, relacionado con las fuentes de las opiniones; y la gradación, que mide la intensidad con la que se expresan dichas valoraciones [36].

En el contexto del comercio electrónico, esta teoría proporciona un marco conceptual para analizar las expresiones lingüísticas de polaridad presentes en las reseñas de productos, las cuales pueden ser identificadas y clasificadas mediante técnicas de aprendizaje automático. La gradación descrita por Martin y White fundamenta además el uso de calificaciones numéricas como indicadores cuantitativos del nivel de satisfacción expresado en cada comentario analizado por el sistema [36].

2.3.3 FUNDAMENTOS DEL PROCESAMIENTO DEL LENGUAJE NATURAL

El procesamiento del lenguaje natural es la disciplina computacional que estudia cómo las máquinas pueden analizar, comprender y generar el lenguaje humano mediante modelos matemáticos y algoritmos estadísticos [21]. Jurafsky y Martin establecen los fundamentos teóricos y computacionales de esta disciplina.

Su marco teórico aborda desde la tokenización y el análisis morfológico hasta la representación semántica y la clasificación de textos. Asimismo, define los principios que rigen el tratamiento del lenguaje durante el flujo de procesamiento, incluyendo la eliminación de stopwords, la lematización y la vectorización de palabras [21].

Estos fundamentos proporcionan el sustento teórico para las técnicas de preprocesamiento implementadas en el sistema, donde las reseñas son sometidas a procesos de limpieza, normalización y transformación textual antes de su representación numérica para las etapas posteriores de clasificación y agrupamiento [21].

2.3.4 TEORÍA DISTRIBUCIONAL DEL SIGNIFICADO (SEMÁNTICA DISTRIBUCIONAL)

La semántica distribucional es una teoría lingüística y computacional que postula que el significado de una palabra puede inferirse a partir de los contextos en los que aparece regularmente dentro de un corpus textual [37].

Este principio, propuesto inicialmente por Zellig Harris, establece que las palabras que comparten contextos similares tienden a poseer significados relacionados, lo que permite representarlas como vectores numéricos en un espacio matemático multidimensional [37]. Posteriormente, Mikolov et al. modernizaron esta teoría mediante el desarrollo de Word2Vec, un modelo capaz de aprender representaciones vectoriales densas que capturan relaciones semánticas entre palabras [25].

Esta base teórica justifica el uso de Word2Vec y FastText en la Fase 3 del sistema, donde cada reseña es transformada en un vector promedio que resume su información semántica para ser utilizada posteriormente en las etapas de clasificación y agrupamiento [25], [37].

2.3.5 TEORÍA DEL APRENDIZAJE ESTADÍSTICO

La teoría del aprendizaje estadístico, desarrollada por Vladimir Vapnik, establece los fundamentos matemáticos que explican en qué condiciones un algoritmo de aprendizaje automático puede generalizar correctamente a partir de un conjunto finito de datos de entrenamiento [38].

Esta teoría introduce conceptos fundamentales como la dimensión de Vapnik-Chervonenkis, el riesgo empírico y el principio de minimización del riesgo estructural, los cuales sustentan la capacidad predictiva de los modelos sobre datos no observados previamente. A partir de estos principios surgieron las Máquinas de Vectores de Soporte (SVM), cuyo objetivo es maximizar el margen de separación entre clases [38].

Esta teoría fundamenta la elección del algoritmo SVM con kernel RBF implementado en la Fase 4 del sistema, donde los vectores generados por Word2Vec y FastText son clasificados automáticamente en las categorías de sentimiento Positivo, Neutro y Negativo [38].

2.3.6 TEORÍA DEL APRENDIZAJE NO SUPERVISADO Y CLUSTERING

El aprendizaje no supervisado constituye una rama del aprendizaje automático orientada a descubrir estructuras y patrones ocultos en conjuntos de datos no etiquetados, sin necesidad de conocer previamente las categorías de los datos [19].

James MacQueen propuso en 1967 el algoritmo K-Means como un método de agrupamiento particional cuyo principio consiste en dividir un conjunto de datos en K grupos minimizando iterativamente la distancia entre cada punto y el centroide de su clúster correspondiente [19]. Posteriormente, Jain destacó la vigencia de este algoritmo al señalar que continúa siendo uno de los métodos de clustering más utilizados debido a su eficiencia computacional y su capacidad para procesar grandes volúmenes de información [39].

Para evaluar la calidad de los agrupamientos obtenidos, Rousseeuw desarrolló el coeficiente de Silhouette, una métrica que mide simultáneamente la cohesión interna de cada clúster y su separación respecto a los demás, proporcionando valores que permiten determinar la calidad de la partición generada [20].

Estos aportes teóricos fundamentan el empleo del algoritmo K-Means para el agrupamiento no supervisado de las reseñas procesadas, permitiendo identificar patrones y similitudes entre los comentarios sin requerir etiquetas previas [19], [20], [39].

2.4 METODOLOGÍA DEL PROYECTO

2.4.1 METODOLOGÍA DE INVESTIGACIÓN

El presente estudio adopta un enfoque metodológico mixto con predominio cuantitativo, ya que combina el análisis de información numérica y textual para el desarrollo del sistema. Desde la perspectiva cuantitativa, se recopilan y procesan las calificaciones otorgadas por los usuarios, expresadas en una escala de una a cinco estrellas, las cuales sirven como etiquetas supervisadas para el entrenamiento del modelo SVM.

Desde la perspectiva cualitativa, se analiza el contenido textual de las opiniones mediante técnicas de procesamiento del lenguaje natural para identificar patrones semánticos y temáticos presentes en las reseñas. La investigación se clasifica como aplicada y de alcance explicativo, ya que busca desarrollar una solución tecnológica que automatice el análisis de sentimientos en reseñas de productos de AliExpress mediante una aplicación de escritorio.

2.4.2 TÉCNICAS E INSTRUMENTOS DE RECOLECCIÓN DE DATOS

La fuente de información corresponde a las reseñas públicas disponibles en la plataforma de AliExpress, obtenidas mediante la técnica de web scraping automatizado, utilizando Selenium WebDriver como instrumento principal de la extracción [40]. El proceso de la extracción se realiza únicamente sobre información pública accesible para cualquier usuario de la plataforma, respetando las condiciones de uso y sin acceder a datos personales o privados.

El usuario define el producto a analizar ingresando su URL directamente en la interfaz gráfica desarrollada en Tkinter, tras lo cual el sistema automatiza la navegación hacia la sección de reseñas del producto seleccionado. Para cada reseña disponible, el sistema extrae el texto completo del comentario y la calificación numérica otorgada por el comprador, correspondiente a un valor entre una y cinco estrellas.

El navegador opera en modo headless, con la carga de imágenes y hojas de estilo deshabilitada para optimizar el rendimiento. Además, el sistema realiza desplazamientos automáticos para recuperar la totalidad de las reseñas disponibles. Finalmente, los datos obtenidos se almacenan en un archivo con formato CSV o JSON, donde cada registro

contiene el título del producto, la calificación numérica y el texto de la reseña para su posterior procesamiento.

2.4.3 METODOLOGÍA DE DESARROLLO

La metodología de desarrollo propuesta fue diseñada específicamente para la presente investigación, respondiendo a las necesidades particulares del sistema automatizado de análisis de sentimientos desarrollado. Su estructura toma como referencia el flujo general empleado en proyectos de minería de datos, procesamiento del lenguaje natural y aprendizaje automático, organizando el proceso en etapas secuenciales que abarcan desde la obtención de los datos hasta la evaluación e interpretación de los resultados [41].

En función de ello, la metodología propuesta organiza el desarrollo del sistema en cinco fases secuenciales e interdependientes: extracción, preprocesamiento, vectorización, entrenamiento y evaluación. Cada fase genera resultados que sirven como entrada para la siguiente, garantizando un flujo coherente desde la adquisición de las reseñas hasta la presentación de los resultados mediante el dashboard interactivo.

La Fase 1, denominada Extracción, comprende la implementación del módulo de web scraping con Selenium WebDriver para recolectar el texto y la calificación de cada reseña disponible en el producto seleccionado mediante URL, almacenando los datos en formato CSV y JSON.

La Fase 2, denominada Preprocesamiento, aplica una cadena de diez procesos secuenciales de limpieza y normalización sobre los comentarios recolectados. Estos procesos incluyen la eliminación de campos vacíos y duplicados, el filtrado de comentarios en idioma distinto al español, la transformación de emojis a texto descriptivo, la eliminación de comentarios irrelevantes, la normalización del texto a minúsculas, la tokenización mediante NLTK, la eliminación de tokens de ruido, la lematización con el modelo `es_core_news_md` de spaCy y la eliminación de stopwords preservando términos con ^{carga} semántica relevante.

La Fase 3, denominada Vectorización, transforma los comentarios tokenizados en representaciones numéricas densas mediante el entrenamiento de dos modelos de word embedding. Se entrena un modelo Word2Vec mediante las arquitecturas skip-gram y

CBOW, y un modelo FastText que incorpora información de subpalabras para manejar términos poco frecuentes. Para cada comentario se calcula el vector promedio de sus tokens como representación numérica única, con una dimensionalidad que se ajusta automáticamente entre 100 y 300 dimensiones según el tamaño del corpus.

La Fase 4, denominada Entrenamiento, implementa cuatro combinaciones de modelos de manera automática: Word2Vec con SVM, FastText con SVM, Word2Vec con K-Means y FastText con K-Means. El modelo SVM utiliza kernel RBF con `class_weight` balanceado y aplica SMOTE sobre el conjunto de entrenamiento para mitigar el desbalance de clases, mientras que K-Means opera con K igual a tres clústeres para explorar agrupamientos de reseñas según su similitud semántica, sin asumir que los clústeres obtenidos representan directamente las categorías Positivo, Neutro y Negativo. La selección de K=3 constituye una decisión experimental que facilita la comparación analítica con la estructura de polaridad del modelo supervisado. Las etiquetas de clasificación para el modelo supervisado se derivan automáticamente de las calificaciones extraídas, asignando la categoría Negativo a valoraciones de una y dos estrellas, Neutro a tres estrellas y Positivo a cuatro y cinco estrellas.

La Fase 5, denominada Evaluación y Visualización, presenta los resultados a través de un dashboard integrado en la interfaz de escritorio. Se calculan y muestran métricas de exactitud, precisión, recall y F1-score ponderado para los modelos SVM, y el coeficiente de Silhouette para los modelos K-Means, permitiendo la comparación directa entre los cuatro experimentos ejecutados. Adicionalmente, se generan visualizaciones de la distribución de sentimientos, matrices de confusión comparativas entre el modelo original y el modelo mejorado con SMOTE, nubes de palabras por categoría de sentimiento, gráficos de las diez palabras más frecuentes por clase y la proyección PCA de los clústeres en dos dimensiones.

CAPÍTULO 3. PROPUESTA

3.1 REQUERIMIENTOS

3.1.1 REQUERIMIENTOS FUNCIONALES

Según las funciones y fases del proyecto, los requisitos destacados son los siguientes.

Extracción de Comentarios	
Código	Descripción
RF-01	El sistema debe permitir ingresar una URL de un producto de AliExpress a través de la interfaz gráfica para iniciar el proceso de extracción.
RF-02	El sistema debe automatizar la navegación web mediante Selenium WebDriver para acceder a la sección del producto indicado.
RF-03	El sistema debe extraer el título del producto, los comentarios de los usuarios y las calificaciones numéricas (1 a 5 estrellas) asociadas a cada reseña.
RF-04	El sistema debe desplazarse automáticamente por la ventana de comentarios para cargar y extraer la totalidad de las reseñas disponibles.
RF-05	El sistema debe permitir guardar los comentarios extraídos en formatos CSV, JSON o TXT, según la preferencia del usuario.
RF-06	El sistema debe ofrecer un modo de ejecución silencioso (headless) y un modo con ventana visible del navegador, configurable por el usuario.

RF-07	El sistema debe permitir configurar la ruta del ejecutable del navegador y del ChromeDriver desde una ventana de configuración integrada.
-------	---

Tabla 1. Requerimientos funcionales – Extracción de comentarios. Elaboración propia.

Limpieza de Datos	
Código	Descripción
RF-08	El sistema debe permitir cargar comentarios desde un archivo externo en formato CSV, JSON o TXT, o utilizar directamente los comentarios extraídos en la fase anterior.
RF-09	El sistema debe eliminar comentarios vacíos y comentarios duplicados del conjunto de datos.
RF-10	El sistema debe detectar y descartar comentarios escritos en idiomas distintos al español, basándose en la presencia de palabras clave de otros idiomas.
RF-11	El sistema debe transformar los emojis presentes en los comentarios a su representación textual equivalente en español.
RF-12	El sistema debe eliminar comentarios considerados irrelevantes, tales como respuestas de muy corta extensión o comentarios compuestos únicamente por palabras sin valor semántico.
RF-13	El sistema debe normalizar los comentarios convirtiéndolos a minúsculas, eliminando URLs, caracteres especiales y espacios redundantes.

RF-14	El sistema debe Tokenizar los comentarios utilizando NLTK, fragmentando cada texto en unidades léxicas individuales (tokens).
RF-15	El sistema debe eliminar tokens irrelevantes tales como interjecciones, muletillas y palabras sin carga semántica.
RF-16	El sistema debe eliminar Stopwords del español, preservando aquellas con valor semántico relevante para el análisis de sentimientos (por ejemplo: “no”, “ni”, “muy”).
RF-17	El sistema debe lematizar los tokens mediante el modelo de lenguaje de spaCy (es_core_news_md), reduciendo cada palabra a su forma base.
RF-18	El sistema debe permitir guardar los comentarios procesados en formato CSV o JSON, incluyendo la correspondencia con los datos originales y las calificaciones.

Tabla 2. Requerimientos funcionales - Limpieza de datos. Elaboración propia.

Vectorización de Texto	
Código	Descripción
RF-19	El sistema debe permitir cargar un archivo de comentarios tokenizados en formato CSV o JSON, o utilizar directamente los comentarios procesados en la fase de limpieza.

RF-20	El sistema debe entrenar un modelo Word2Vec sobre el corpus tokenizado, generando representaciones vectoriales densas para cada palabra.
RF-21	El sistema debe entrenar un modelo FastText sobre el corpus tokenizado, generando representaciones vectoriales que consideren subpalabras para mayor robustez ante términos poco frecuentes.
RF-22	El sistema debe calcular el vector representativo de cada comentario promediando los vectores de sus tokens, tanto para Word2Vec como para FastText.
RF-23	El sistema debe ajustar automáticamente la dimensión de los vectores en función del tamaño del corpus (100D, 150D, 200D o 300D).
RF-24	El sistema debe permitir guardar los vectores generados en formato CSV, JSON o Numpy (.npy).

Tabla 3. Requerimientos funcionales - Vectorización de texto. Elaboración propia.

Entrenamiento de Modelos	
Código	Descripción
RF-25	El sistema debe derivar automáticamente etiquetas de sentimiento (Positivo, Neutro, Negativo) a partir de las calificaciones numéricas extraídas, siguiendo la regla: 1-2 estrellas -> Negativo, 3 estrellas -> Neutro, 4-5 estrellas -> Positivo.
RF-26	El sistema debe normalizar los vectores de entrada mediante estandarización (StandardScaler) antes de entrenar los modelos.

RF-27	El sistema debe dividir el conjunto de datos en subconjuntos de entrenamiento (80%) y prueba (20%) de forma estratificada, para el entrenamiento del modelo supervisado.
RF-28	El sistema debe entrenar un modelo de Máquina de Vectores de Soporte (SVM con kernel RBF) para la clasificación supervisada de sentimientos.
RF-29	El sistema debe entrenar un modelo de agrupamiento K-Means (K=3) para el análisis no supervisado de los comentarios.
RF-30	El sistema debe ejecutar automáticamente las cuatro combinaciones posibles entre vectorización y modelo: Word2Vec+SVM, FastText+SVM, Word2Vec+K-Means y FastText+K-Means.
RF-31	El sistema debe registrar y comparar los resultados de cada combinación entrenada en una tabla comparativa acumulativa.
RF-32	El sistema debe aplicar la técnica SMOTE exclusivamente sobre el conjunto de entrenamiento para generar muestras sintéticas de las clases minoritarias y mitigar el desbalance de clases presente en las reseñas reales.
RF-33	El sistema debe ejecutar dos experimentos comparativos por cada combinación de vectorización y modelo: el Experimento 1 sin balanceo de clases y el Experimento 2 con SMOTE y <code>class_weight='balanced'</code> , permitiendo comparar el impacto del balanceo sobre las métricas obtenidas.

Tabla 4. Requerimientos funcionales – Entrenamiento de modelos. Elaboración propia.

Evaluación y visualización de Resultados	
Código	Descripción
RF-34	El sistema debe calcular y mostrar las métricas de exactitud (accuracy), precisión recall y F1-score ponderado para los modelos SVM entrenados.
RF-35	El sistema debe calcular y mostrar el índice Silhouette Score como métrica de calidad para los modelos K-Means entrenados.
RF-36	El sistema debe generar y visualizar de forma comparativa las matrices de confusión del Experimento 1 (modelo original) y del Experimento 2 (modelo mejorado con SMOTE), permitiendo al usuario identificar el impacto del balanceo sobre la clasificación.
RF-37	El sistema debe visualizar la distribución de sentimientos del corpus mediante gráficos de barras y gráficos circulares.
RF-38	El sistema debe visualizar los clusters generados por K-Means proyectados en dos dimensiones mediante Análisis de Componentes Principales (PCA).
RF-39	El sistema debe generar nubes de palabras (WordCloud) diferenciadas por categoría de sentimiento: Positivo, Neutro y Negativo.
RF-40	El sistema debe mostrar las 10 palabras más frecuentes por cada categoría de sentimiento mediante gráficos de barras horizontales.
RF-41	El sistema debe presentar un Dashboard centralizado que integre las métricas, gráficos comparativas y resumen del análisis en una única ventana.

RF-42	El sistema debe generar automáticamente una conclusión textual del análisis, indicando la distribución de sentimientos, el mejor modelo obtenido y la calidad del clasificador.
RF-43	El sistema debe permitir exportar los resultados del análisis (métricas, comparativas y conclusiones) en formato CSV, JSON o TXT.

Tabla 5. Requerimientos funcionales - Evaluación y visualización de resultados.

Elaboración propia.

Pipeline Automatizado	
Código	Descripción
RF-44	El sistema debe ofrecer un modo de ejecución automática (pipeline) que ejecute secuencialmente las cinco fases del proceso: extracción, limpieza, vectorización, entrenamiento y visualización de resultados, a partir de una única URL de producto.

Tabla 6. Requerimientos funcionales - Pipeline automatizado. Elaboración propia.

3.1.2 REQUERIMIENTOS NO FUNCIONALES

Tipo de requerimiento	Código	Especificación de requisitos
Requerimiento de Usabilidad	RNF-01	La interfaz gráfica del sistema debe estar desarrollada en tkinter, con una estructura visual organizada en fases secuenciales claramente identificadas mediante colores y etiquetas, de modo que el usuario pueda comprender el flujo del proceso sin necesidad de conocimientos técnicos avanzados.

Requerimiento de Usabilidad	RNF-02	El sistema debe mostrar retroalimentación visual del progreso en cada fase mediante barras de progreso y mensajes de estado actualizados en tiempo real.
Requerimiento de Rendimiento	RNF-03	El sistema debe ajustar automáticamente la dimensionalidad de los vectores en función del tamaño del corpus, con el fin de optimizar el tiempo de procesamiento y el uso de memoria.
Requerimiento de Rendimiento	RNF-04	Los modelos Word2Vec, FastText, SVM y K-Means deben entrenarse de forma secuencial y automática dentro de un tiempo razonable para corpus de tamaño típico (hasta varios miles de comentarios), sin requerir intervención manual entre fases.
Requerimiento de Mantenibilidad	RNF-05	El código debe estar organizado en módulos funcionales claramente separados por fase (extracción, limpieza, vectorización, entrenamiento y resultados), facilitando futuras modificaciones, ampliaciones o sustitución de componentes individuales.
Requerimiento de Mantenibilidad	RNF-06	Los parámetros clave del sistema (rutas del navegador y driver, modo headless, timeout, hiperparámetros de modelos) deben ser configurables por el usuario sin necesidad de modificar el código fuente.
Requerimiento de Portabilidad	RNF-07	El sistema debe poder ejecutarse en entornos Windows con Python 3.x instalado, sin requerir cambios en el código fuente, siempre que se cumplan las dependencias declaradas (selenium,

		scikit-learn, Gensim, spaCy, NLTK, tkinter, entre otras).
Requerimiento de interoperabilidad	RNF-08	El sistema debe soportar la importación y exportación de datos en formatos estándar (CSV, JSON, TXT, NumPy), permitiendo la integración con otras herramientas de análisis externas.
Requerimiento de Seguridad	RNF-09	El sistema únicamente debe procesar comentarios y calificaciones de acceso público disponibles en AliExpress, sin recopilar, almacenar ni transmitir datos personales o sensibles de los usuarios.
Requerimiento de Robustez	RNF-10	El sistema debe gestionar excepciones en todas las fases críticas (errores de conexión, archivos no encontrados, inconsistencias en los datos) y notificar al usuario mediante mensajes de error descriptivos, sin interrumpir abruptamente la ejecución de programa.

Tabla 7. Requerimientos no funcionales. Elaboración propia.

3.2 DISEÑO DE LA PROPUESTA

El diseño de la propuesta establece la estructura arquitectónica, el flujo de procesamiento y la organización visual del sistema automatizado de análisis de sentimientos desarrollado. Esta sección describe cómo se organizan los componentes del sistema, cómo fluye la información entre las cinco fases del pipeline y cómo interactúa el usuario con la aplicación de escritorio desarrollada en Tkinter. Las decisiones de diseño adoptadas responden a los requerimientos funcionales y no funcionales definidos en la sección anterior, garantizando coherencia entre la planificación y la implementación del sistema.

3.2.1 ARQUITECTURA GENERAL DEL SISTEMA

El sistema desarrollado sigue una arquitectura de pipeline secuencial de cinco capas, donde cada capa representa una fase del proceso de análisis y la salida de cada una constituye la entrada de la siguiente. Esta arquitectura es característica de los sistemas de procesamiento de lenguaje natural y minería de datos, ya que permite separar claramente las responsabilidades de cada módulo y facilita el mantenimiento e independencia de cada componente [41].

La primera capa corresponde al módulo de extracción, que interactúa con AliExpress mediante Selenium WebDriver para recopilar datos brutos. La segunda capa abarca el módulo de preprocesamiento, responsable de limpiar y normalizar el texto mediante diez procesos secuenciales con NLTK y spaCy. La tercera capa comprende el módulo de vectorización, que transforma los tokens en representaciones numéricas densas mediante Word2Vec y FastText de Gensim.

La cuarta capa corresponde al módulo de entrenamiento, donde los vectores generados son procesados por los algoritmos SVM con SMOTE y K-Means mediante Scikit-learn e Imbalanced-learn. La quinta capa es el módulo de visualización, que presenta los resultados a través del dashboard interactivo integrado en la interfaz gráfica desarrollada con Tkinter y Matplotlib.



Figura 2. Arquitectura general del sistema de análisis de sentimientos en cinco capas.

Elaboración propia.

3.2.2 DIAGRAMA DE FLUJO DEL PIPELINE

El pipeline del sistema representa el flujo completo de transformación de datos desde su estado original como texto no estructurado hasta la generación de los resultados analíticos presentados al usuario. Este flujo está diseñado de forma secuencial, permitiendo además que el usuario ingrese datos en cualquier fase intermedia cargando archivos CSV o JSON previamente generados, sin necesidad de ejecutar el proceso completo desde la extracción [41].

El flujo inicia cuando el usuario ingresa la URL del producto y activa la extracción. Los comentarios y calificaciones recopilados pasan al preprocesamiento, donde son sometidos a los diez procesos de limpieza. Los tokens resultantes ingresan a la vectorización, generando vectores Word2Vec y FastText de dimensionalidad variable, los cuales son normalizados mediante StandardScaler.

Estos vectores son procesados por los cuatro experimentos del módulo de entrenamiento de forma automática. Finalmente, las métricas y visualizaciones generadas son presentadas en el dashboard de la Fase 5. El sistema también ofrece un modo de ejecución automática denominado pipeline, que ejecuta las cinco fases de manera completamente secuencial a partir de una única URL ingresada.

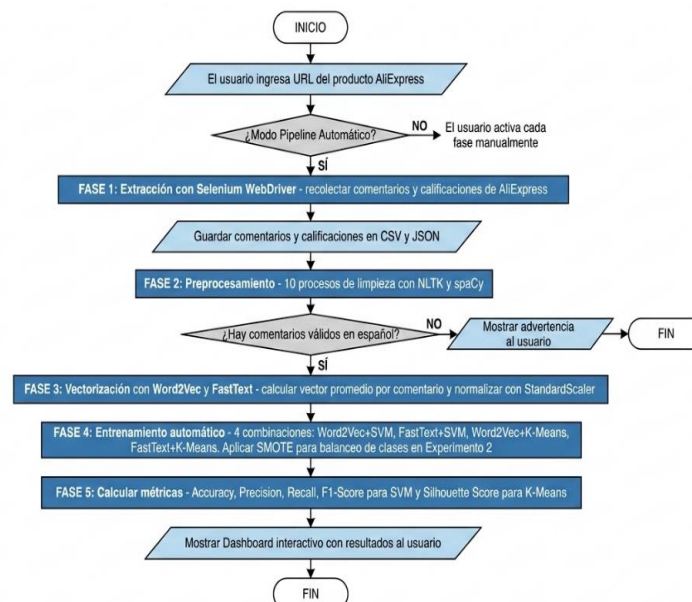


Figura 3. Diagrama de flujo del procesamiento del sistema de análisis de sentimientos.

Elaboración propia.

3.2.3 DIAGRAMAS DE CASOS DE USO

Los diagramas de casos de uso describen las interacciones posibles entre el usuario y el sistema, identificando las funcionalidades disponibles desde la perspectiva del actor principal. En el presente sistema, el actor principal es el usuario de la aplicación de escritorio, quien puede interactuar con cada fase del sistema a través de la interfaz gráfica sin necesidad de conocimientos técnicos avanzados.

En la Fase 1 el usuario puede ingresar una URL, iniciar la extracción manual o el pipeline automático, configurar el navegador y guardar los comentarios extraídos. En la Fase 2 puede cargar un archivo externo o usar los datos ya extraídos, iniciar la limpieza y guardar los comentarios procesados. En la Fase 3 puede cargar comentarios tokenizados o usar los datos limpios, iniciar la vectorización y guardar los vectores generados.

En la Fase 4 el usuario puede iniciar el entrenamiento automático de las cuatro combinaciones de modelos y visualizar el resumen de métricas obtenidas. En la Fase 5 puede abrir el dashboard, visualizar las matrices de confusión comparativas, generar nubes de palabras, exportar los resultados y consultar la conclusión automática del análisis generada por el sistema.

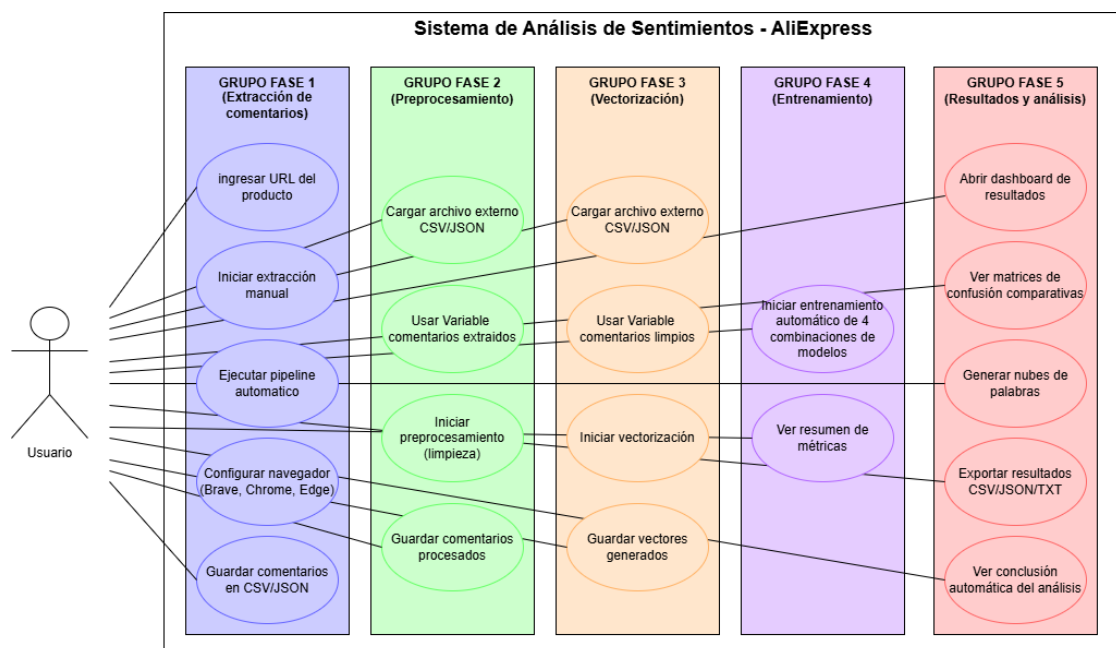


Figura 4. Diagrama de caso de uso del sistema de análisis de sentimientos. Elaboración propia.

3.2.4 DISEÑO DE INTERFACES

El diseño de la interfaz gráfica de usuario se desarrolló bajo el principio de usabilidad progresiva, organizando visualmente las cinco fases del sistema en un flujo vertical descendente dentro de una ventana principal con desplazamiento. Cada fase está representada mediante una tarjeta visual con un color identificador único, un badge numérico y controles específicos que guían al usuario de forma intuitiva a través del proceso de análisis [42].

La ventana principal presenta en su parte superior un encabezado con el nombre del sistema, seguido de las cinco tarjetas de fase en orden secuencial. La Fase 1 contiene el campo de ingreso de URL y los botones de extracción manual y pipeline automático. La Fase 2 presenta los botones de carga de datos, la lista de los diez procesos aplicados y el botón de inicio de limpieza. La Fase 3 muestra las opciones de entrada de datos y el botón de vectorización.

La Fase 4 describe las cuatro combinaciones a entrenar y el botón de inicio del entrenamiento automático. La Fase 5 contiene el botón de acceso al dashboard, que se habilita únicamente tras completar el entrenamiento. Todas las fases incluyen barras de progreso con mensajes de estado en tiempo real que informan al usuario sobre los datos procesados en cada etapa.

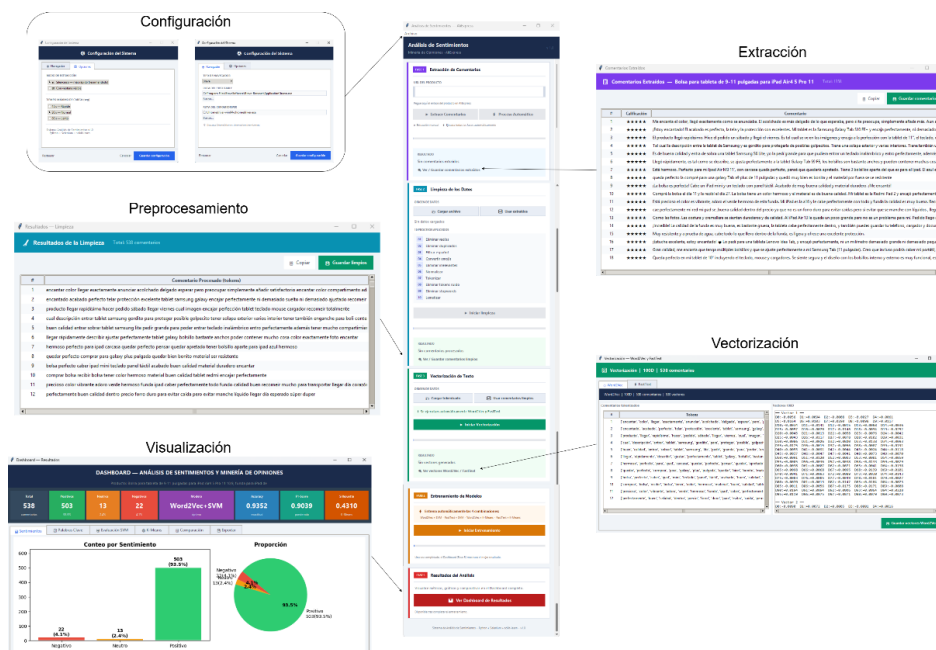


Figura 5. Diseño de la interfaz del sistema de análisis de sentimientos.

3.3 DESARROLLO DEL SISTEMA

El desarrollo del sistema se estructura en cinco fases secuenciales que implementan el pipeline completo de análisis de sentimientos, desde la recolección automatizada de reseñas hasta la presentación de resultados en el dashboard interactivo. Cada fase fue implementada como un módulo independiente dentro de la aplicación de escritorio desarrollada en Python con Tkinter, permitiendo su ejecución individual o de forma completamente automatizada mediante el modo pipeline.

3.3.1 FASE 1 - EXTRACCIÓN DE COMENTARIOS

La primera fase implementa el módulo de extracción automatizada de reseñas mediante Selenium WebDriver, que controla una instancia del navegador Google Chrome para navegar hacia la sección de comentarios del producto indicado por el usuario a través de una URL ingresada en la interfaz gráfica [27].. El sistema ofrece dos modos de operación configurables: el modo headless, que ejecuta la navegación en segundo plano sin mostrar ventana del navegador, y el modo visible, que permite observar el proceso de extracción en tiempo real.

Durante la extracción, el sistema realiza desplazamientos automáticos dentro del contenedor de reseñas para cargar todos los comentarios disponibles mediante renderizado dinámico. Por cada reseña detectada, el módulo extrae el texto completo del comentario y la calificación numérica en estrellas correspondiente, descartando automáticamente cualquier dato identificable del usuario en cumplimiento con la LOPDP. Los datos recopilados se almacenan en formato CSV y JSON, estructurando cada registro con el título del producto, el texto de la reseña y su calificación numérica.

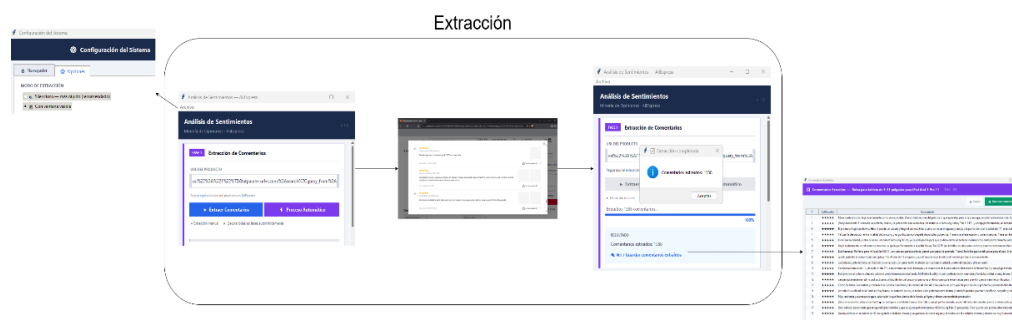


Figura 6. Fase 1 - Extracción de Comentarios.

3.3.2 FASE 2 - PREPROCESAMIENTO Y LIMPIEZA DE DATOS

La segunda fase implementa una cadena de diez procesos secuenciales de limpieza y normalización sobre los comentarios recolectados, con el objetivo de eliminar el ruido lingüístico y estandarizar el texto antes de la vectorización [43]. El módulo acepta como entrada los datos extraídos en la Fase 1 o un archivo CSV y JSON cargado externamente por el usuario, ofreciendo flexibilidad en el flujo de trabajo.

Los diez procesos aplicados de forma secuencial son los siguientes: eliminación de comentarios vacíos y duplicados, filtrado de comentarios en idioma distinto al español, transformación de emojis a su descripción textual equivalente en español, eliminación de comentarios irrelevantes de muy corta extensión, normalización del texto a minúsculas y eliminación de URLs y caracteres especiales, tokenización mediante NLTK, eliminación de tokens de ruido como interjecciones y muletillas, lematización con el modelo `es_core_news_md` de spaCy, y eliminación de stopwords preservando términos con carga semántica relevante como negaciones. Los comentarios procesados se almacenan en formato CSV y JSON manteniendo la correspondencia con sus calificaciones originales.

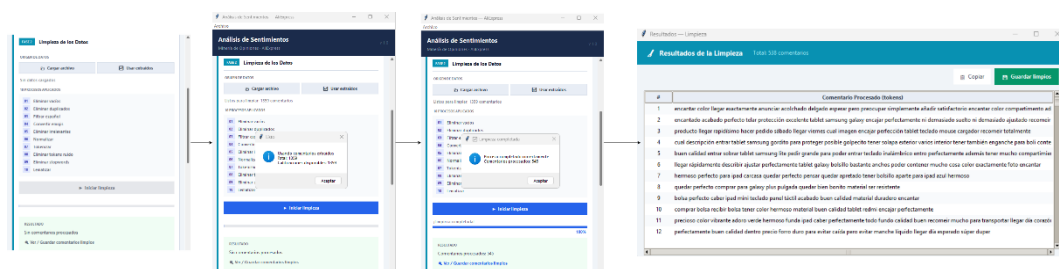


Figura 7. Fase 2 - Preprocesamiento y Limpieza de Datos.

3.3.3 FASE 3 - VECTORIZACIÓN DE TEXTO

La tercera fase transforma los comentarios tokenizados en representaciones numéricas densas mediante el entrenamiento de dos modelos de word embedding sobre el corpus procesado [25]. El módulo acepta como entrada los datos limpios de la Fase 2 o un archivo externo cargado por el usuario, y ajusta automáticamente la dimensionalidad de los vectores entre 100 y 300 dimensiones según el tamaño del corpus disponible.

El primer modelo entrenado es Word2Vec, que aprende representaciones vectoriales capturando las relaciones semánticas entre palabras según sus contextos de aparición. El

segundo modelo entrenado es FastText, que incorpora información de subpalabras mediante n-gramas de caracteres para generar vectores robustos ante términos poco frecuentes o con variaciones ortográficas propias del lenguaje informal de las reseñas. Para ambos modelos, el vector representativo de cada comentario se calcula como el promedio aritmético de los vectores de sus tokens válidos, generando una única representación numérica por reseña apta para los algoritmos de clasificación y agrupamiento.

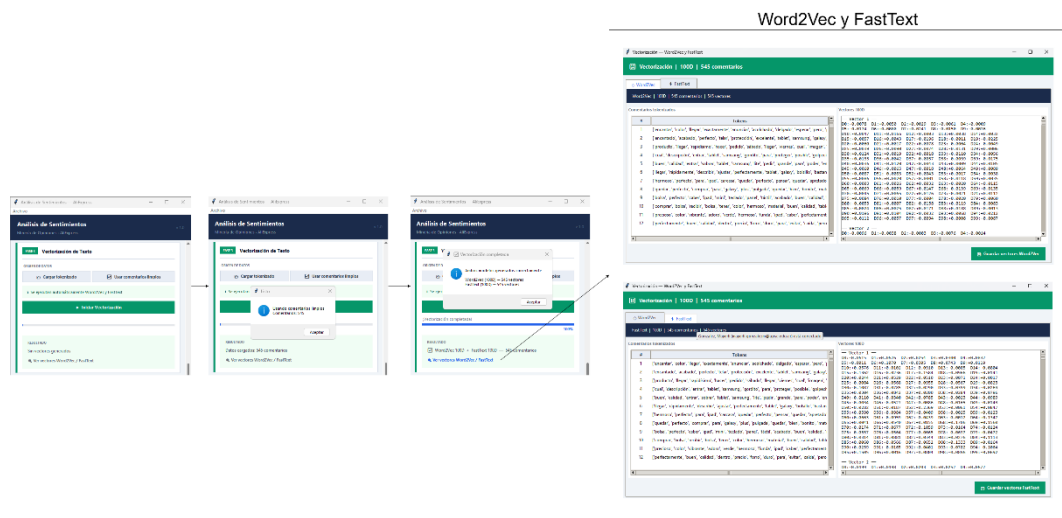


Figura 8. Fase 3 - Vectorización de Texto.

3.3.4 FASE 4 - ENTRENAMIENTO DE MODELOS

La cuarta fase implementa el entrenamiento automático de cuatro combinaciones de modelos de aprendizaje automático sobre los vectores generados en la fase anterior [29]. Las combinaciones ejecutadas son Word2Vec con SVM, FastText con SVM, Word2Vec con K-Means y FastText con K-Means, permitiendo comparar el rendimiento de ambas técnicas de vectorización sobre ambos algoritmos de forma simultánea.

Las etiquetas de clasificación para el modelo supervisado se derivan automáticamente de las calificaciones extraídas en la Fase 1, asignando la categoría Negativo a valoraciones de una y dos estrellas, Neutro a tres estrellas y Positivo a cuatro y cinco estrellas. Antes del entrenamiento, los vectores son normalizados mediante StandardScaler. El modelo SVM utiliza kernel RBF con class_weight igual a balanced, y aplica SMOTE

exclusivamente sobre el conjunto de entrenamiento para mitigar el desbalance de clases, mientras que el conjunto de prueba al 20% permanece sin modificaciones.

El modelo K-Means opera con $K=3$ clústeres para explorar agrupamientos de reseñas por similitud semántica. La elección de $K=3$ se estableció como una decisión experimental que facilita la comparación del comportamiento del agrupamiento con las tres categorías de polaridad utilizadas en el modelo supervisado, sin asumir equivalencia entre los clústeres y dichas categorías.

Es importante aclarar que K-Means no realiza clasificación de sentimientos en este sistema, sino agrupamiento no supervisado de reseñas por similitud semántica. El valor K igual a tres fue seleccionado por correspondencia con las tres categorías de polaridad definidas en el análisis supervisado, lo que facilita la comparación entre ambos enfoques, pero los clústeres generados por K-Means no equivalen directamente a las categorías Positivo, Neutro y Negativo, ya que el algoritmo opera sin etiquetas y agrupa los comentarios únicamente por sus características vectoriales en el espacio semántico.

Para cada combinación se ejecutan dos experimentos: el Experimento 1 sin balanceo de clases y el Experimento 2 con SMOTE y `class_weight` balanceado, permitiendo comparar el impacto de las técnicas de balanceo sobre el rendimiento de los modelos entrenados.

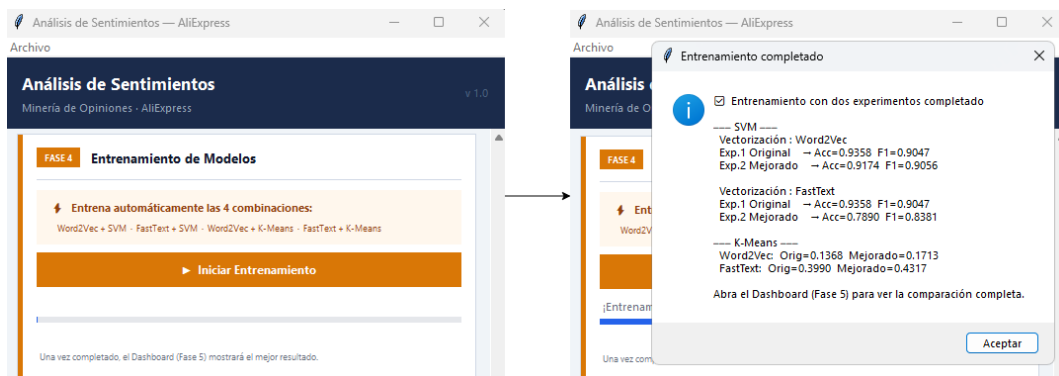


Figura 9. Fase 4 - Entrenamiento de Modelos.

3.3.5 FASE 5 - CONFIGURACIÓN DE MÉTRICAS DE EVALUACIÓN

La quinta fase configura y calcula el conjunto de métricas estadísticas utilizadas para evaluar el rendimiento de los modelos entrenados en la fase anterior [29]. Para los modelos SVM se calculan la exactitud, la precisión ponderada, el recall ponderado y el

F1-score ponderado, junto con el reporte de clasificación detallado por categoría de sentimiento. Para los modelos K-Means se calcula el coeficiente de Silhouette como indicador de la calidad geométrica de los agrupamientos generados.

Adicionalmente, esta fase genera las visualizaciones que conforman el dashboard interactivo presentado al usuario en la interfaz de escritorio. Estas visualizaciones incluyen los gráficos de distribución de sentimientos, las matrices de confusión comparativas entre el Experimento 1 original y el Experimento 2 mejorado con SMOTE, las nubes de palabras diferenciadas por categoría de sentimiento, los gráficos de las diez palabras más frecuentes por clase, la proyección PCA de los clústeres en dos dimensiones y la tabla comparativa acumulativa de todos los experimentos ejecutados.

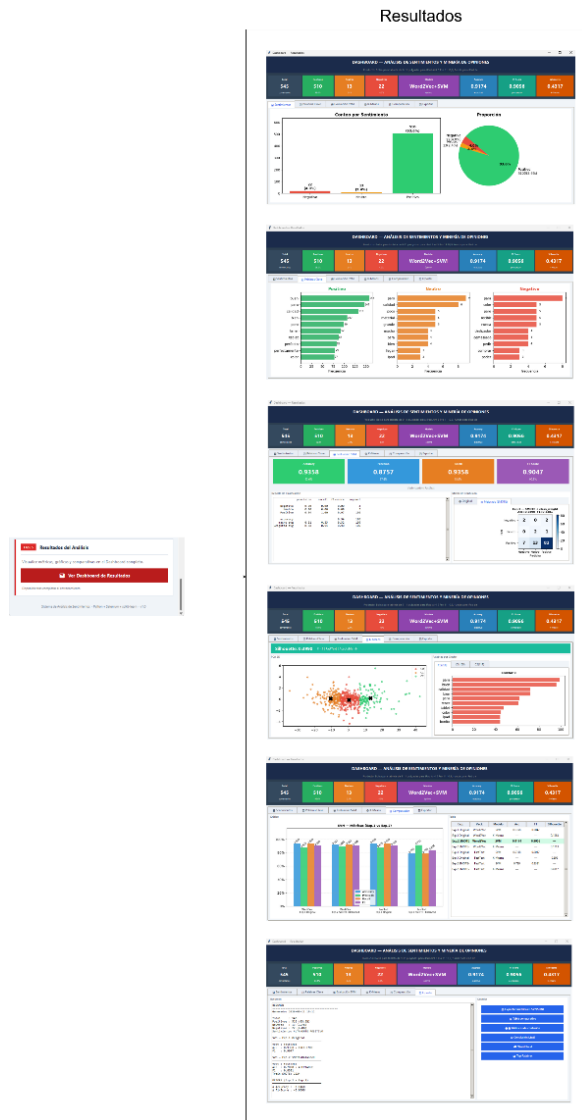


Figura 10. Fase 5 - Visualización de resultados.

3.4 DESARROLLO DEL APLICATIVO Y DASHBOARD

El desarrollo del aplicativo integra los cinco módulos del pipeline de análisis en una única aplicación de escritorio funcional, construida con la biblioteca Tkinter de Python y diseñada para operar en entornos Windows sin requerir conocimientos técnicos avanzados por parte del usuario [42]. La arquitectura del aplicativo separa claramente la lógica de procesamiento de la capa de presentación visual, permitiendo que cada fase del sistema pueda ejecutarse de forma independiente o en modo pipeline automático según las necesidades del usuario.

3.4.1 INTEGRACIÓN DEL MODELO CON LA INTERFAZ PRINCIPAL

La interfaz principal actúa como el punto central de control del sistema, coordinando la ejecución secuencial de las cinco fases del pipeline mediante una arquitectura basada en hilos de ejecución paralelos que evitan el bloqueo de la interfaz gráfica durante el procesamiento [42]. Cada fase del sistema está representada visualmente como una tarjeta independiente con controles propios, barras de progreso en tiempo real y secciones de resultado que informan al usuario sobre el estado del procesamiento en cada etapa.

La integración entre los modelos de machine learning y la interfaz se realiza mediante variables globales compartidas que almacenan los datos procesados entre fases, permitiendo que la salida de cada módulo esté disponible inmediatamente como entrada del siguiente sin necesidad de operaciones adicionales de carga. Los modelos Word2Vec, FastText, SVM y K-Means entrenados en la Fase 4 quedan almacenados en memoria durante toda la sesión de uso, habilitando automáticamente el botón de acceso al dashboard una vez completado el entrenamiento.

El sistema implementa además un modo de ejecución automática denominado pipeline, que permite al usuario ingresar únicamente la URL del producto y ejecutar las cinco fases de forma completamente secuencial sin intervención manual adicional. Este modo gestiona internamente las dependencias entre módulos, el manejo de excepciones en cada etapa y la retroalimentación visual del progreso global mediante mensajes de estado actualizados en tiempo real.



Figura 11. Interfaz Principal

3.4.2 VISUALIZACIÓN DE RESULTADOS EN EL DASHBOARD

El dashboard de resultados constituye el componente visual central del aplicativo, presentando de forma integrada todos los resultados del análisis de sentimientos mediante una ventana de pantalla completa organizada en seis pestañas temáticas navegables [42]. Esta ventana se habilita exclusivamente tras completar el entrenamiento de los modelos en la Fase 4 y muestra en su encabezado un resumen de ocho tarjetas con los indicadores más relevantes del análisis ejecutado.

La primera pestaña denominada Sentimientos presenta la distribución de las categorías Positivo, Neutro y Negativo mediante un gráfico de barras y un gráfico circular con los conteos y porcentajes correspondientes.

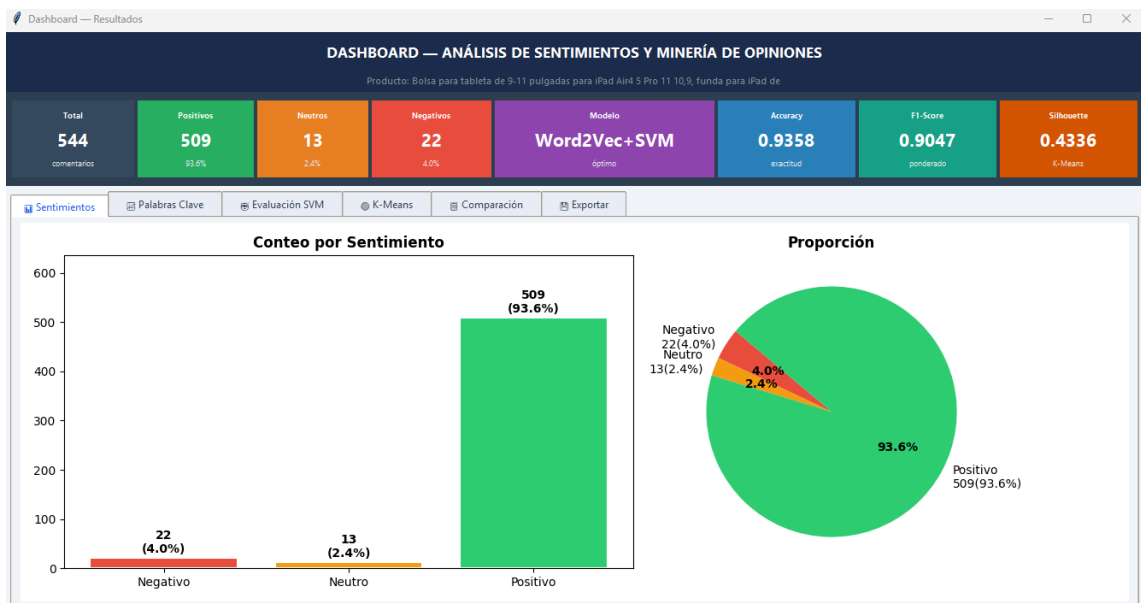


Figura 12. Pestaña de sentimiento

La segunda pestaña denominada Palabras Clave muestra los gráficos de las diez palabras más frecuentes por categoría de sentimiento.

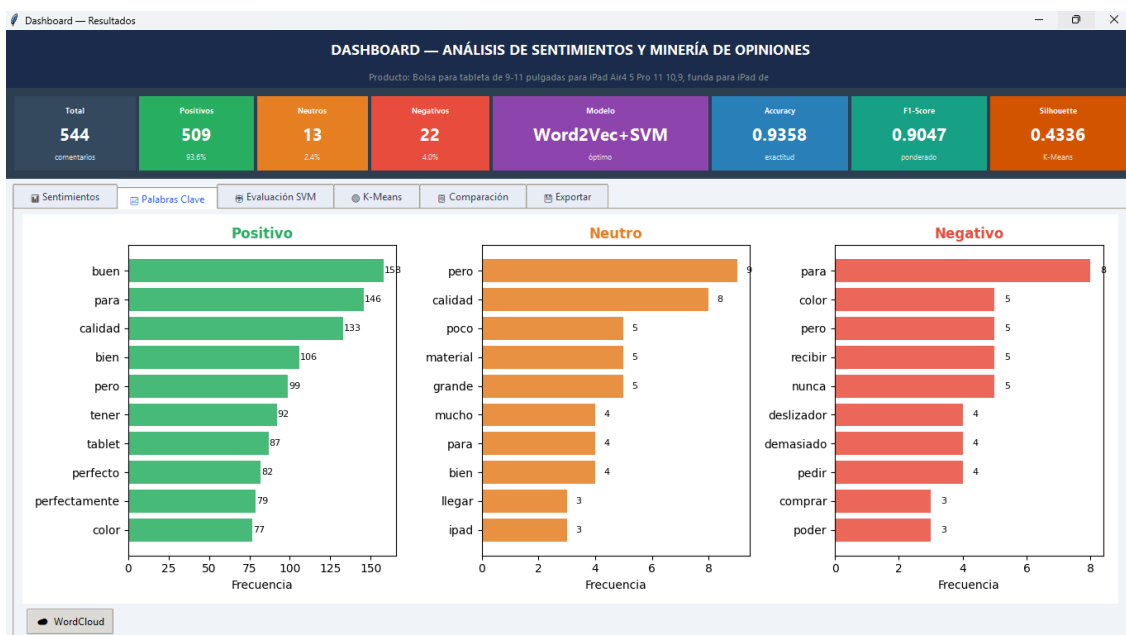


Figura 13. Pestaña de palabras Clave

Dentro de esta pestaña también se encuentra un botón de acceso a la visualización de nubes de palabras generadas con WordCloud.



Figura 14. Nube de palabras

La tercera pestaña denominada Evaluación SVM presenta las tarjetas de métricas con los valores de exactitud, precisión, recall y F1-score, el reporte de clasificación detallado y la sección de Matriz de Confusión organizada en dos sub-pestañas comparativas. La sub-pestaña Experimento 1 muestra la matriz del modelo original sin balanceo de clases, mientras que la sub-pestaña Experimento 2 muestra la matriz del modelo mejorado con SMOTE y class_weight balanceado, permitiendo al usuario identificar visualmente el impacto del balanceo sobre la clasificación.

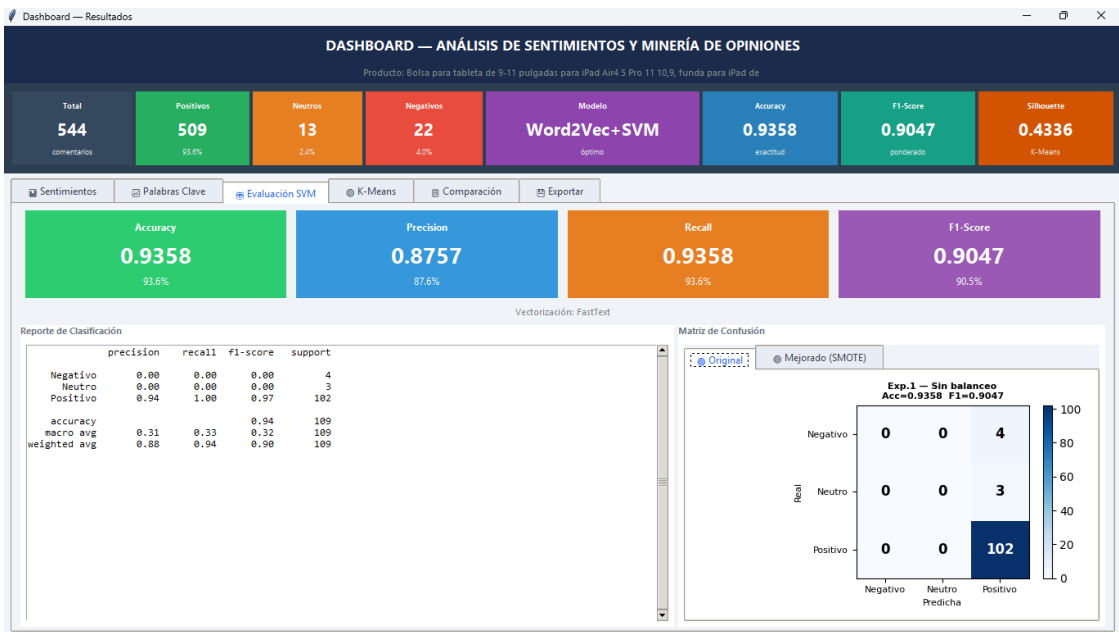


Figura 15. Pestaña de evaluación SVM

La cuarta pestaña denominada K-Means presenta el coeficiente de Silhouette obtenido, la proyección PCA de los clústeres en dos dimensiones con sus centroides marcados.

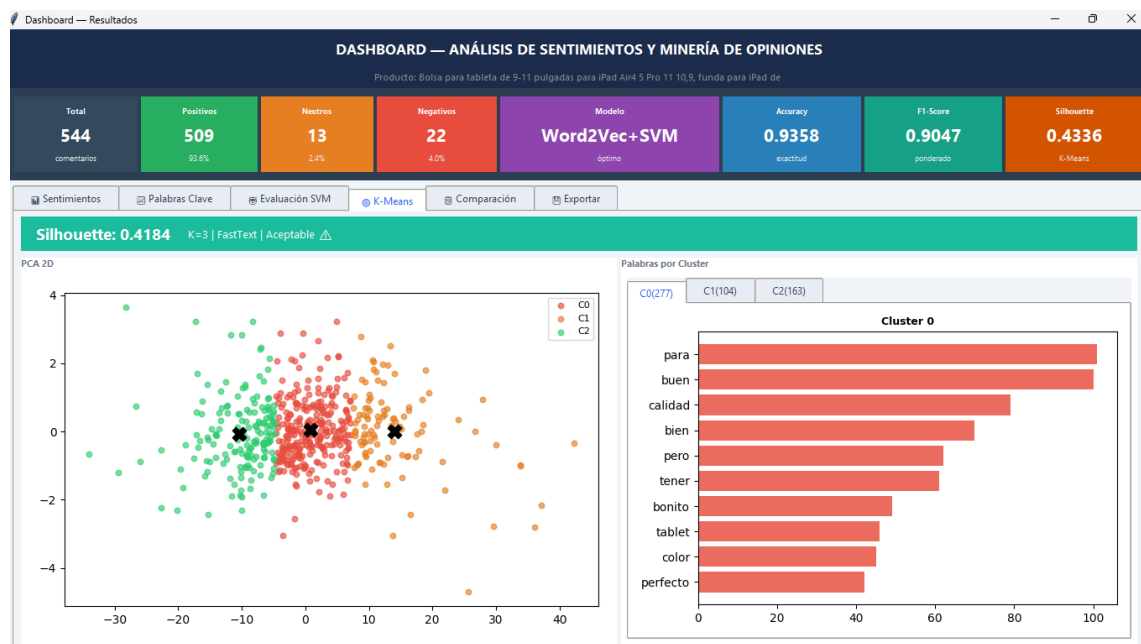


Figura 16. Pestaña de k-means

A lado de la gráfica de PCA se encuentran los gráficos de las palabras más frecuentes por clúster, organizados en subpestañas.



Figura 17. Palabras más frecuentes por clúster

La quinta pestaña denominada Comparación presenta el gráfico de barras agrupadas con las métricas de todos los experimentos ejecutados y la tabla comparativa acumulativa que destaca automáticamente el modelo con mayor F1-score.

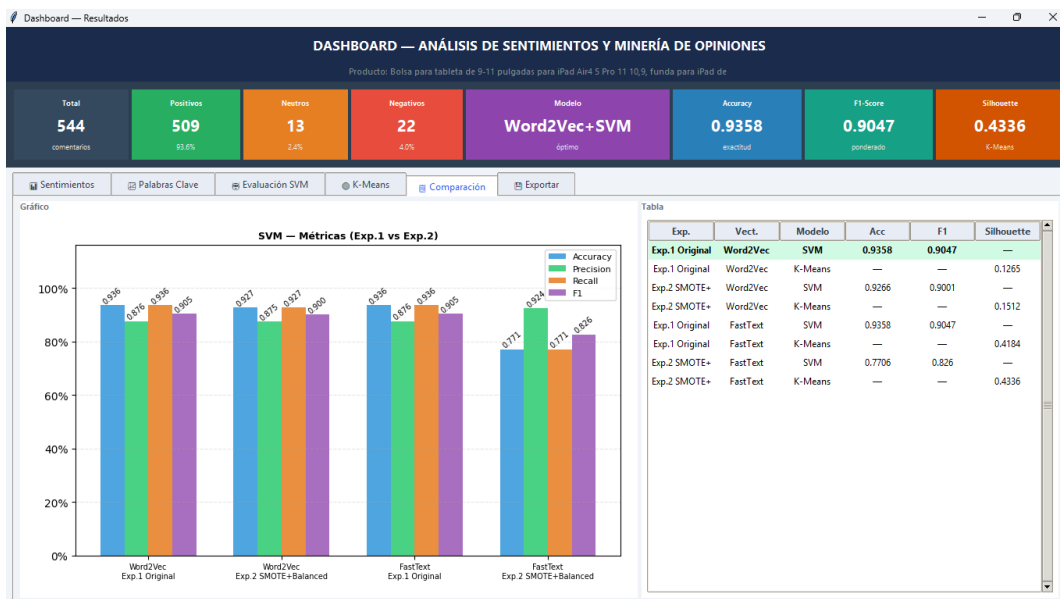


Figura 18. Pestaña de comparación con gráfico de barras y tabla de métricas acumulativas.

La sexta pestaña denominada Exportar presenta el resumen textual completo del análisis incluyendo la comparación entre experimentos y los botones de exportación de resultados en formato CSV, JSON y TXT. Dentro de esta pestaña también existe el botón de presenta la conclusión automática generada por el sistema, describiendo la distribución de sentimientos identificada y el rendimiento comparativo de los modelos entrenados.

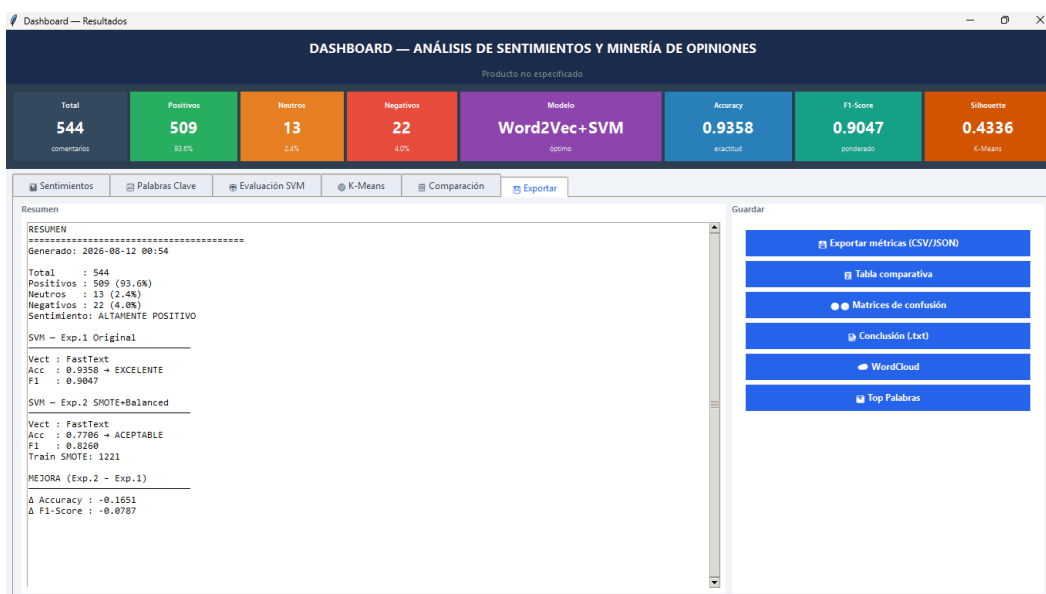


Figura 19. Pestaña exportar

3.5 PRUEBAS DEL SISTEMA

Las pruebas del sistema tienen como objetivo verificar que el aplicativo desarrollado cumple con los requerimientos funcionales y no funcionales establecidos en la sección 3.1, garantizando que cada módulo del pipeline opera correctamente de forma individual y en conjunto. El proceso de pruebas se divide en dos categorías complementarias: las pruebas funcionales, orientadas a verificar el comportamiento correcto de la interfaz y los módulos del sistema, y las pruebas de los modelos, orientadas a evaluar el rendimiento estadístico de los algoritmos de aprendizaje automático implementados.

3.5.1 PRUEBAS FUNCIONALES

Las pruebas funcionales verifican que cada funcionalidad implementada en la interfaz gráfica responde correctamente ante las acciones del usuario y produce los resultados esperados según los requerimientos definidos. Estas pruebas se ejecutaron de forma manual sobre la aplicación de escritorio, simulando los escenarios de uso más representativos del sistema en condiciones reales de operación.

La siguiente tabla presenta los casos de prueba funcionales ejecutados sobre el sistema:

ID	Módulo	Caos de prueba	Resultado Esperado	Resultado Obtenido	Estado
PF-01	Fase 1	Ingresar URL validez de AliExpress e iniciar extracción	El sistema extrae comentarios y calificaciones y los almacena en CSV y JSON	El sistema extrae correctamente los comentarios y genera los archivos de salida	Aprobado
PF-02	Fase 1	Ingresar URL inválida	El sistema muestra mensaje de error	El sistema muestra alerta indicando	Aprobado

				problemas con el ChromeDriver	
PF-03	Fase 1	Ejecutar modo pipeline automático desde URL	El sistema ejecuta las cinco fases secuencialmente sin intervención del usuario	El pipeline ejecutó las cinco fases correctamente	Aprobado
PF-04	Fase 2	Iniciar limpieza con datos extraídos en la Fase 1	El sistema aplica diez procesos y muestra comentarios limpios	Los diez procesos se ejecutaron y los resultados se almacenaron correctamente	Aprobado
PF-05	Fase 2	Cargar archivo CSV externo e iniciar limpieza	El sistema carga el archivo y aplica el preprocesamiento	El sistema procesó el archivo externo correctamente	Aprobado
PF-06	Fase 3	Iniciar vectorización con datos limpios de la Fase 2	El sistema entrena Word2Vec y FastText y genera vectores promedio	Ambos modelos de vectorización se entrenaron y los vectores se almacenaron correctamente	Aprobado
PF-07	Fase 3	Verificar ajuste automático de dimensionalidad	El sistema ajusta la dimensión entre 100 y 300 según el tamaño del corpus.	La dimensionalidad se ajustó automáticamente	Aprobado

				según el corpus disponible	
PF-08	Fase 4	Iniciar entrenamiento automático de los cuatro modelos	El sistema ejecuta los cuatro experimentos y muestra resumen de métricas	Los cuatro modelos se entrenaron y las métricas se presentaron correctamente	Aprobado
PF-09	Fase 4	Verificar aplicación de SMOTE en Experimento 2	SMOTE se aplica solo sobre el conjunto de entrenamiento y no sobre el conjunto de prueba	SMOTE se aplicó exclusivamente sobre x_train e y_train	Aprobado
PF-10	Fase 5	Abrir dashboard tras completar entrenamiento	El dashboard se abre con las seis pestañas y los resultados visibles	El dashboard se desplegó correctamente con todos los componentes activos	Aprobado
PF-11	Fase 5	Visualizar matrices de confusión comparativas	El dashboard muestra dos sub-pestañas con la matriz del Experimento 1 y del Experimento 2	Ambas matrices se mostraron correctamente con sus métricas asociadas	Aprobado

PF-12	Fase 5	Exportar resultados en formato CSV, JSON y TXT	El sistema genera los archivos de exportación en la ruta seleccionada	Los archivos se generaron correctamente en los tres formatos	Aprobado
PF-13	General	Ejecutar el sistema sin conexión a internet en Fase 2 a 5	El sistema opera correctamente sobre datos previamente guardados	Las fases 2 a 5 funcionaron sin requerir conexión a internet	Aprobado
PF-14	General	Verificar que el botón de dashboard permanece deshabilitado antes del entrenamiento	El botón de Fase 5 solo se habilita tras completar la Fase 4	El botón permaneció deshabilitado hasta completar el entrenamiento	Aprobado

Tabla 8. Pruebas funcionales ejecutadas sobre el sistema.

3.5.2 PRUEBAS DE LOS MODELOS

Las pruebas de los modelos evalúan el rendimiento estadístico de los algoritmos SVM y K-Means entrenados durante la Fase 4 del sistema, utilizando el conjunto de datos de prueba reservado al 20% del total de reseñas recopiladas mediante web scraping [29]. Estas pruebas permiten verificar que los modelos implementados cumplen con los niveles de desempeño esperados para la tarea de clasificación de sentimientos y agrupamiento semántico de reseñas de AliExpress.

Para los modelos SVM se evalúan cuatro métricas principales: exactitud, precisión ponderada, recall ponderado y F1-score ponderado, calculadas sobre el conjunto de prueba no modificado por SMOTE. La comparación entre el Experimento 1 sin balanceo

y el Experimento 2 con SMOTE y class_weight balanceado permite cuantificar el impacto de las técnicas de balanceo sobre el rendimiento de la clasificación en las tres categorías de sentimiento.

La siguiente tabla presenta los resultados obtenidos en las pruebas de los modelos SVM:

Experimento	Vectorización	Accuracy	Precision	Recall	F1-Score
Experimento.1 Original	Word2Vec + SVM	0.9358	0.8757	0.9358	0.9047
Experimento.1 Original	FastText + SVM	0.9358	0.8757	0.9358	0.9047
Experimento 2: SMOTE + class_weight='balanced'	Word2Vec + SVM	0.9266	0.8751	0.9266	0.9001
Experimento 2: SMOTE + class_weight='balanced'	FastText + SVM	0.7706	0.9239	0.7706	0.826

Tabla 9. Tabla de resultados en las pruebas de los modelos SVM.

La siguiente tabla presenta los resultados obtenidos en las pruebas de los modelos K-Means:

Experimento	Vectorización	Silhouette Score	Calidad
Exp.1 Original	Word2Vec + K-Means	0.1265	Baja
Exp.1 Original	FastText + K-Means	0.4184	Aceptable

Exp.2 SMOTE	Word2Vec + K-Means	0.1512	Baja
Exp.2 SMOTE	FastText + K-Means	0.4336	Aceptable

Tabla 10. Tabla de resultados en las pruebas de los modelos K-Means

Los resultados de las pruebas SVM evidencian el impacto del desbalance de clases sobre el rendimiento del clasificador. El Experimento 1 sin balanceo alcanzó un Accuracy de 0.9358, pero las matrices de confusión revelan que este valor es engañoso, ya que el modelo clasificó todas las observaciones Negativas y Neutras del conjunto de prueba como Positivas, mostrando incapacidad total para discriminar las clases minoritarias. El Experimento 2 con SMOTE y `class_weight` balanceado en la combinación FastText con SVM redujo el Accuracy a 0.7706, pero adquirió capacidad para identificar las tres categorías de sentimiento, aunque con un desempeño desigual entre ellas, identificando correctamente comentarios Negativos y Neutros que el modelo original ignoraba por completo.

Respecto al agrupamiento no supervisado, los valores de Silhouette Score indican una estructura de agrupamiento parcial con separación moderada entre clústeres. FastText con K-Means presenta la mejor calidad con valores de 0.4184 y 0.4336 en ambos experimentos, clasificados como aceptables, mientras que Word2Vec con K-Means obtiene valores bajos en ambos casos. Estos resultados son coherentes con la naturaleza del corpus analizado, ya que las reseñas de un único producto de AliExpress comparten vocabulario entre categorías de sentimiento, lo que limita la separación geométrica alcanzable mediante agrupamiento no supervisado. Los resultados se analizan con mayor profundidad en la sección 3.6.

3.6 ANÁLISIS DE RESULTADOS

3.6.1 RESULTADOS DE CLASIFICACIÓN SVM

El sistema procesó un total de 1.358 reseñas extraídas del producto Bolsa para tableta de 9 a 11 pulgadas para iPad en AliExpress, de las cuales 544 quedaron disponibles tras

aplicar la cadena de diez procesos de preprocesamiento y limpieza de datos. La reducción de 1.358 a 544 comentarios válidos se explica por la eliminación de reseñas en idiomas distintos al español, comentarios duplicados, textos vacíos y opiniones irrelevantes de muy corta extensión que no aportaban valor semántico al análisis.

La distribución de sentimientos derivada de las calificaciones numéricas muestra una marcada asimetría característica del comercio electrónico: 509 comentarios Positivos correspondientes al 93.6% del corpus, 22 comentarios Negativos equivalentes al 4.0% y 13 comentarios Neutros representando el 2.4% restante. Esta distribución confirma el desbalance severo de clases identificado como problema central en la Fase 4, justificando la necesidad de implementar SMOTE y `class_weight` balanceado en el Experimento 2.

El conjunto de datos fue dividido en 80% para entrenamiento con 435 registros y 20% para prueba con 109 registros, manteniendo la proporción estratificada de las tres categorías. Tras aplicar SMOTE exclusivamente sobre el conjunto de entrenamiento en el Experimento 2, el número de muestras de entrenamiento se expandió a 1.221 registros sintéticos y reales combinados, equilibrando las tres categorías de sentimiento para el proceso de aprendizaje supervisado.

Los resultados del Experimento 1 sin balanceo revelan que, a pesar de alcanzar un Accuracy de 0.9358, el modelo no logró discriminar entre las tres categorías de sentimiento. El reporte de clasificación muestra una Precision y Recall de 0.00 para las clases Negativo y Neutro, y un F1-score macro de 0.32, confirmando que el clasificador asignó la totalidad de los comentarios del conjunto de prueba a la clase Positivo. Este comportamiento es consecuencia directa del desbalance severo del corpus, donde el 93.6% de los registros pertenecen a la categoría positiva, lo que induce al modelo a optimizar su rendimiento global ignorando sistemáticamente las clases minoritarias.

El Experimento 2 con SMOTE produjo resultados diferenciados según la vectorización. Word2Vec con SVM obtuvo un Accuracy de 0.9266 y un F1-score de 0.9001, mostrando una mejora marginal en la distribución de errores, pero sin lograr detectar comentarios Negativos. FastText con SVM alcanzó un Accuracy de 0.7706, una Precision de 0.9239 y un F1-score de 0.8260. En el conjunto de prueba, el modelo identificó correctamente 2 de los 4 comentarios Negativos, equivalente al 50 %; los 3 comentarios Neutros presentes, equivalente al 100 %; y 79 de los 102 comentarios Positivos, equivalente al 77,45 %.

Aunque el Accuracy global disminuyó, FastText con SVM fue el único modelo que demostró capacidad de discriminación entre las tres categorías de sentimiento.

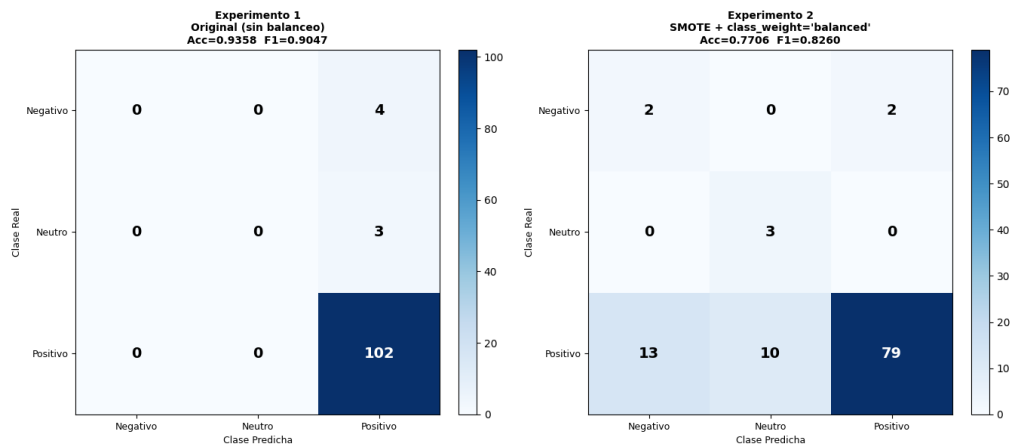


Figura 20. Comparación de Experimentos Original vs Mejorado con SMOTE

La reducción del Accuracy en el Experimento 2 con FastText no representa una degradación del sistema sino una redistribución del error hacia una clasificación más equitativa entre clases. Un clasificador que detecta comentarios Negativos reales tiene mayor valor práctico para compradores y vendedores que uno que ignora sistemáticamente las críticas negativas, aunque su Accuracy global sea menor.

3.6.2 RESULTADOS DE AGRUPAMIENTO K-MEANS

El algoritmo K-Means fue aplicado con K igual a tres clústeres sobre los 544 vectores normalizados de dimensionalidad 100D generados en la Fase 3. Los tres clústeres resultantes agruparon los comentarios de la siguiente manera: el Clúster 0 concentró 277 elementos, el Clúster 1 agrupó 104 elementos y el Clúster 2 reunió 163 elementos, sumando el total de 544 comentarios procesados.

La calidad de los agrupamientos evaluada mediante el coeficiente de Silhouette produjo los siguientes resultados: Word2Vec con K-Means alcanzó valores de 0.1265 en el Experimento 1 y 0.1512 en el Experimento 2, ambos clasificados como calidad baja. FastText con K-Means obtuvo valores de 0.4184 en el Experimento 1 y 0.4336 en el Experimento 2, ambos clasificados como calidad aceptable. Los valores de Silhouette obtenidos indican una estructura de agrupamiento parcial con separación moderada entre clústeres, sin alcanzar niveles de cohesión elevados. El solapamiento léxico observado en

las palabras más frecuentes por grupo evidencia que las reseñas de un único producto comparten vocabulario entre categorías, lo que limita la separación geométrica alcanzable mediante agrupamiento no supervisado en este dominio.

La proyección PCA en dos dimensiones muestra que los tres clústeres presentan cierta separación en el espacio vectorial. El Clúster 2 se concentra en la región izquierda del eje principal, el Clúster 0 ocupa la zona central y el Clúster 1 se distribuye hacia la región derecha. Sin embargo, existe solapamiento entre clases, especialmente entre C0 y C2, lo que explica el valor aceptable pero no excelente del Silhouette Score obtenido.

El análisis de las palabras más frecuentes por clúster revela que los tres grupos comparten términos como calidad, buen, bien y para, lo que indica que los comentarios de distintas categorías de sentimiento emplean un vocabulario muy similar.

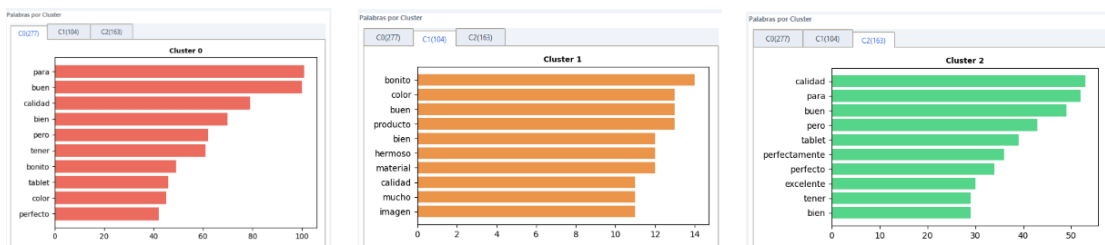


Figura 21. Palabras más frecuentes por clúster.

Esta superposición léxica es la causa principal de los valores moderados de Silhouette, ya que dificulta la separación geométrica nítida de los clústeres en el espacio vectorial de 100 dimensiones generado por FastText.

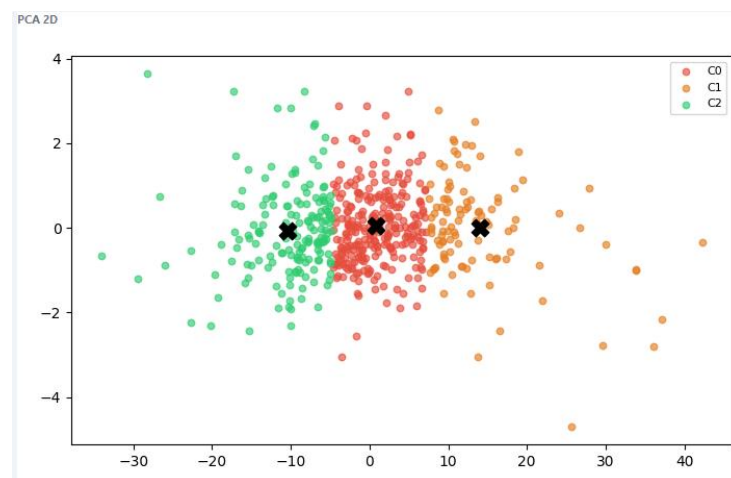


Figura 22. PCA 2D

3.6.3 COMPARACIÓN DE EXPERIMENTOS ORIGINAL VS MEJORADO CON SMOTE

La comparación entre el Experimento 1 original y el Experimento 2 mejorado con SMOTE permite identificar claramente el impacto de las técnicas de balanceo de clases sobre el comportamiento real del sistema de clasificación. La siguiente tabla sintetiza las diferencias entre ambos experimentos:

Criterio	Experimento 1 Original	Experimento 2: SMOTE + class_weight='balanced'
Accuracy (Word2Vec+SVM)	0.9358	0.9266
F1-Score (Word2Vec+SVM)	0.9047	0.9001
Accuracy (FastText+SVM)	0.9358	0.7706
Precision (FastText + SVM)	0.8757	0.9239
F1-Score (FastText+SVM)	0.9047	0.8260
Negativos detectados correctamente	0 de 4	2 de 4
Neutros detectados correctamente	0 de 3	3 de 3
Muestras entrenamiento	435	1.221 con SMOTE

Silhouette FastText+K-Means	0.4184	0.4336
Utilidad práctica real	Muy baja	Mejorada / Moderada

Tabla 11. Tabla de diferencias entre los dos experimentos.

El Experimento 1 presenta métricas globales superiores numéricamente; sin embargo, su utilidad práctica es muy baja, debido a que clasifica incorrectamente todas las observaciones de las clases minoritarias como Positivas. El Experimento 2 presenta una disminución del Accuracy global, pero mejora la capacidad del modelo para identificar las clases minoritarias. En particular, logra identificar correctamente 2 de las 4 reseñas Negativas y 3 de las 3 reseñas Neutras del conjunto de prueba. Esta diferencia evidencia que un modelo con mayor Accuracy no necesariamente presenta un mejor desempeño para la clasificación de clases desbalanceadas [1].

JUSTIFICACIÓN DE LOS PARÁMETROS UTILIZADOS

K=3 en K-Means: Se seleccionó experimentalmente un valor de K=3 con el propósito de explorar la estructura semántica de las reseñas y facilitar la comparación de los agrupamientos con las tres categorías de polaridad utilizadas en el análisis supervisado. Esta selección no implica que los clústeres obtenidos correspondan directamente con las categorías Positivo, Neutro y Negativo. La determinación del número óptimo de clústeres mediante métricas como el coeficiente de Silhouette y métodos complementarios se considera una posible mejora futura.

Kernel RBF en SVM: El kernel de función de base radial fue seleccionado porque los vectores Word2Vec y FastText no son linealmente separables en el espacio original de características, y el kernel RBF proyecta implícitamente los datos a un espacio de mayor dimensionalidad donde la separación lineal sí es posible [38].

Word2Vec y FastText: Se eligieron ambos modelos para comparar el rendimiento de una representación basada en palabras completas contra una basada en subpalabras. FastText ofrece mayor robustez ante términos poco frecuentes y variaciones ortográficas

propias del lenguaje informal de las reseñas, lo que lo hace especialmente adecuado para este dominio [44].

División 80/20: La proporción de entrenamiento y prueba de 80% y 20% respectivamente es la más utilizada en proyectos de machine learning de tamaño pequeño a mediano, garantizando suficientes datos de entrenamiento sin reducir excesivamente el conjunto de evaluación [41].

LIMITACIONES IDENTIFICADAS

El análisis de resultados permite identificar tres limitaciones principales del sistema. La primera es el desbalance severo del corpus, donde el 93.6% de los 544 comentarios son Positivos, lo que restringe la capacidad de aprendizaje del modelo supervisado sobre las clases minoritarias incluso tras aplicar SMOTE, como evidencia que el modelo FastText con SVM logró detectar correctamente 2 de los 4 comentarios Negativos presentes en el conjunto de prueba, equivalente al 50%. La segunda es la dimensionalidad reducida de los vectores, ya que el corpus de 544 comentarios condicionó el ajuste automático a 100 dimensiones, siendo recomendable un corpus de varios miles de registros para entrenar embeddings de mayor calidad semántica. La tercera es la dependencia de un único producto, dado que los resultados obtenidos sobre la bolsa para tableta no son generalizables a otras categorías de productos de AliExpress sin ejecutar nuevamente el sistema sobre un corpus diferente.

3.7 ESTUDIO DE FACTIBILIDAD

3.7.1 FACTIBILIDAD TÉCNICA

La factibilidad técnica evalúa si los recursos de hardware y software disponibles son suficientes para el desarrollo, despliegue y ejecución del sistema en condiciones reales de operación. El sistema fue desarrollado íntegramente en Python 3.12.0, un lenguaje de programación de código abierto que no requiere licencias comerciales, y todas las bibliotecas utilizadas en su implementación son de distribución libre y gratuita [28].

Los requisitos mínimos de hardware para ejecutar el sistema son un procesador de doble núcleo a 2.0 GHz o superior, 4 GB de memoria RAM para corpus de tamaño pequeño y

8 GB recomendados para corpus de mayor volumen, 2 GB de espacio en disco para la instalación de las dependencias y el almacenamiento de los datos generados, y conexión a internet estable únicamente durante la fase de extracción con Selenium. El sistema fue desarrollado y probado en entornos Windows, siendo compatible con cualquier equipo que disponga de Python 3.12.0 instalado y las dependencias declaradas.

La siguiente tabla presenta las herramientas tecnológicas utilizadas en el desarrollo del sistema:

Herramienta	Versión	Función	Licencia
Python	3.12.0	Lenguaje de programación principal	PSF License
Tkinter	8.6.13	Interfaz gráfica de escritorio	PSF License
Selenium WebDriver	4.19.0	Automatización del navegador	Apache 2.0
ChromeDriver	151.0.7922.77	Control del navegador	BSD
NLTK	3.9.1	Tokenización y stopwords	Apache 2.0
spaCy	3.8.5	Lematización en español	MIT
Gensim	4.3.3	Word2Vec y FastText	LGPL
Scikit-learn	1.5.1	SVM y K-Means	BSD

Imbalanced-learn	0.14.2	SMOTE	MIT
NumPy	1.26.4	Operaciones matriciales	BSD
Pandas	2.2.2	Manipulación de datos	BSD
Matplotlib	3.9.0	Visualizaciones y gráficos	PSF License
WordCloud	1.9.6	Nubes de palabras	MIT

Tabla 12. Herramientas tecnológicas utilizadas en el desarrollo del sistema

Todas las herramientas listadas son de código abierto y distribución gratuita, lo que garantiza que el sistema puede ser replicado y ejecutado sin incurrir en costos de licenciamiento de software.

3.7.2 FACTIBILIDAD OPERATIVA

La factibilidad operativa evalúa si el sistema puede ser utilizado de manera efectiva por los usuarios finales en el entorno previsto, considerando su nivel de usabilidad, el esfuerzo de aprendizaje requerido y la sostenibilidad de su operación a lo largo del tiempo. El sistema fue diseñado bajo el principio de usabilidad progresiva, organizando las cinco fases del pipeline en una interfaz gráfica intuitiva que guía al usuario paso a paso sin requerir conocimientos previos en programación o machine learning [42].

La operación del sistema requiere únicamente que el usuario disponga de un equipo con Windows, Python 3.12.0 instalado y las dependencias del proyecto configuradas. Una vez satisfechos estos requisitos, el usuario puede ejecutar el análisis completo de cualquier producto de AliExpress ingresando únicamente su URL en la interfaz principal y presionando el botón de pipeline automático. Los resultados son presentados de forma visual e interpretable en el dashboard interactivo, eliminando la necesidad de interpretar código o salidas de consola.

Desde la perspectiva de mantenibilidad, el código está organizado en módulos funcionales claramente separados por fase, lo que facilita futuras modificaciones en los parámetros de los modelos, la adición de nuevas técnicas de vectorización o la adaptación del sistema a otras plataformas de comercio electrónico. Los parámetros clave del sistema, como la ruta del navegador, el modo headless y los hiperparámetros de los modelos, son configurables desde la interfaz sin necesidad de modificar el código fuente.

CONCLUSIONES

- Mediante la recolección de 1.358 reseñas del producto seleccionado en AliExpress mediante web scraping con Selenium WebDriver, extrayendo el texto del comentario y la calificación en estrellas de cada reseña pública disponible, se dio cumplimiento al primer objetivo específico del proyecto. De las reseñas extraídas, 544 resultaron válidas tras el preprocesamiento, con una distribución de 509 positivas, 22 negativas y 13 neutras, evidenciando el desbalance severo característico de las plataformas de comercio electrónico.
- La aplicación de los diez procesos de preprocesamiento con NLTK y spaCy, seguida del entrenamiento de modelos Word2Vec y FastText sobre el corpus de 544 comentarios, generó vectores promedio de 100 dimensiones que representaron semánticamente cada reseña, dando cumplimiento al segundo objetivo específico. La comparación entre ambas técnicas demostró que FastText produce representaciones más robustas ante el vocabulario informal y las variaciones ortográficas presentes en las reseñas de AliExpress en español.
- La implementación del algoritmo SVM para la clasificación supervisada de polaridad y del algoritmo K-Means para el agrupamiento no supervisado de reseñas por tópicos dio cumplimiento al tercer objetivo específico. El análisis de palabras por clúster identificó que los grupos generados concentran términos relacionados con aspectos como la calidad del material, el color y la funcionalidad del producto, revelando los tópicos de mayor impacto en las opiniones de los usuarios.
- La evaluación mediante métricas de validación estadística demostró que el Accuracy no es un indicador suficiente para corpus con desbalance severo, dando cumplimiento al cuarto objetivo específico. El Experimento 1 sin balanceo obtuvo un Accuracy de 0.9358 con un F1-score macro de 0.32, mientras que el Experimento 2 con SMOTE y class_weight balanceado, utilizando FastText con SVM, alcanzó una Precision de 0.9239 y un F1-score de 0.8260. En el conjunto de prueba, el modelo identificó correctamente 2 de los 4 comentarios Negativos presentes, equivalente al 50%, y los 3 de los 3 comentarios Neutros presentes, equivalente al 100%. Estos resultados evidencian que las técnicas de balanceo permitieron mejorar la capacidad del clasificador para reconocer las clases minoritarias.

- La visualización de los resultados mediante la interfaz gráfica desarrollada en Tkinter, que integra un dashboard interactivo con seis pestañas incluyendo métricas comparativas, matrices de confusión, nubes de palabras y proyección PCA de clústeres, dio cumplimiento al quinto objetivo específico, proporcionando una experiencia intuitiva para la interpretación de los resultados del análisis de sentimientos y agrupamiento temático de las reseñas analizadas.

RECOMENDACIONES

- Se recomienda extender el sistema para soportar la extracción y análisis de reseñas en otras plataformas de comercio electrónico como Temu, Amazon o Mercado Libre, adaptando el módulo de Selenium a las estructuras HTML específicas de cada plataforma. Esta expansión permitiría comparar patrones de sentimiento entre distintos mercados y ampliar la aplicabilidad del sistema más allá de AliExpress.
- Se recomienda explorar técnicas de vectorización más avanzadas como los modelos de lenguaje preentrenados BERT y RoBERTa en su versión para español denominada BETO, que incorporan representaciones contextuales capaces de capturar el significado de las palabras según su posición en la oración. Su implementación sobre el mismo pipeline podría mejorar significativamente las métricas de clasificación en las categorías minoritarias, especialmente en corpus con desbalance severo como el analizado en este trabajo.
- Se recomienda ampliar el sistema con un módulo de análisis de sentimientos a nivel de aspecto, que permita identificar no solo la polaridad global de una reseña sino también qué características específicas del producto generan cada reacción, como la calidad del material, el tiempo de envío o la correspondencia con la descripción. Esta mejora requeriría la anotación de un corpus etiquetado con atributos específicos y la implementación de modelos de extracción de aspectos sobre el pipeline existente.
- Se recomienda aplicar técnicas de optimización de hiperparámetros como Grid Search o búsqueda aleatoria para encontrar la configuración óptima del kernel SVM y los parámetros de entrenamiento de Word2Vec y FastText, con el objetivo de mejorar el rendimiento de los modelos sobre corpus de distintos tamaños y distribuciones de clases sin requerir ajustes manuales.
- Se recomienda implementar un módulo de actualización periódica del corpus que permita ejecutar la extracción de nuevas reseñas de forma programada, acumulando datos en el tiempo para detectar cambios en la tendencia de sentimientos de un producto. Esta funcionalidad convertiría el sistema en una herramienta de monitoreo continuo con mayor valor práctico para compradores y vendedores en plataformas de comercio electrónico internacional.

REFERENCIAS

- [1] B. Liu, «Sentiment Analysis and Opinion Mining,» Springer Nature, [En línea]. Available: <https://doi.org/10.2200/S00416ED1V01Y201204HLT016>. [Último acceso: 12 07 2026].
- [2] B. Pang y L. Lee, «Opinion Mining and Sentiment Analysis,» Emerald Publishing Limited, [En línea]. Available: <https://www.emerald.com/ftinr/article-abstract/2/1-2/1/1321550/Opinion-Mining-and-Sentiment-Analysis?redirectedFrom=fulltext>. [Último acceso: 12 07 2026].
- [3] S. M. Mudambi y D. Schuff, «What Makes a Helpful Online Review? A Study of Customer Reviews on Amazon.com,» JSTOR, [En línea]. Available: <https://www.jstor.org/stable/20721420>. [Último acceso: 12 '7 2026].
- [4] T. Joachims, «Text categorization with Support Vector Machines: Learning with many relevant features,» Springer Nature, [En línea]. Available: <https://link.springer.com/chapter/10.1007/BFb0026683>. [Último acceso: 12 07 2026].
- [5] C. C. Aggarwal y C. Zhai, «A Survey of Text Clustering Algorithms,» Springer Nature, [En línea]. Available: https://link.springer.com/chapter/10.1007/978-1-4614-3223-4_4. [Último acceso: 13 07 2026].
- [6] P. Bojanowski, E. Grave, A. Joulin y T. Mikolov, «Enriching Word Vectors with Subword Information,» arxiv, [En línea]. Available: <https://arxiv.org/abs/1607.04606>. [Último acceso: 13 07 2026].
- [7] G. A. Martínez Misal, «UTADEO - Universidad Jorge Tadeo lozano,» 2021. [En línea]. Available: <https://expeditiorepositorio.utadeo.edu.co/handle/20.500.12010/19138>. [Último acceso: 03 05 2024].

- [8] M. Becerra, «SEDICI - Repositorio Institucional de la Universidad Nacional de La Plata,» 2017. [En línea]. Available: <https://sedici.unlp.edu.ar/handle/10915/65440>. [Último acceso: 03 05 2024].
- [9] H. D. Yamunaque Villegas, «Uso del análisis de sentimientos de reseñas en restaurantes de comidas peruanas,» Repositorio Digital de la Universidad César Vallejo, 2022. [En línea]. Available: <https://repositorio.ucv.edu.pe/handle/20.500.12692/115238>. [Último acceso: 03 05 2024].
- [10] S. G. UPSE, «Universidad Estatal Peninsula de Santa Elena,» 22 07 1998. [En línea]. [Último acceso: 03 05 2024].
- [11] Organisation for Economic Co-operation and Development, «OECD Recommendation of the Council on Consumer Protection in E-Commerce,» OECD Publishing, 2016. [En línea]. Available: <https://doi.org/10.1787/9789264255258-en>. [Último acceso: 10 09 2026].
- [12] Conferencia de las Naciones Unidas sobre Comercio y Desarrollo, «Fomentar la confianza en los mercados digitales mediante una mayor protección del consumidor en las plataformas en línea,» Naciones Unidas, 21 04 2023. [En línea]. Available: https://unctad.org/system/files/official-document/cicplpd34_es_0.pdf. [Último acceso: 10 09 2026].
- [13] S. Russell y P. Norvig, «Artificial Intelligence: A Modern Approach,» University of California, Berkeley, [En línea]. Available: <https://aima.cs.berkeley.edu/>. [Último acceso: 13 07 2026].
- [14] E. Alpaydm, «Introduction to Machine Learning,» MIT Press, 2026. [En línea]. Available: <https://mitpress.mit.edu/9780262043793/introduction-to-machine-learning/>. [Último acceso: 13 07 2026].

- [15] T. Hastie, R. Tibshirani y J. Friedman, «The Elements of Statistical Learning: Data Mining, Inference, and Prediction.,» Springer Nature, [En línea]. Available: <https://link.springer.com/book/10.1007/978-0-387-84858-7>. [Último acceso: 14 07 2026].
- [16] C. Cortes y V. Vapnik, «"Support-vector networks", Machine Learning,» Springer Nature, [En línea]. Available: <https://link.springer.com/article/10.1007/BF00994018>. [Último acceso: 14 07 2026].
- [17] H. He y E. A. Garcia, «Learning from Imbalanced Data,» IEEE Xplore, [En línea]. Available: <https://ieeexplore.ieee.org/document/5128907>. [Último acceso: 14 07 2026].
- [18] N. V. Chawla, K. W. Bowyer, L. O. Hall y K. P. Kegelmeyer, «SMOTE: Synthetic Minority Over-sampling Technique,» JAIR, [En línea]. Available: <https://www.jair.org/index.php/jair/article/view/10302>. [Último acceso: 14 07 2026].
- [19] J. MacQueen, «Some methods for classification and analysis of multivariate observations,» Project Euclid, a Duke University Press initiative, [En línea]. Available: <https://projecteuclid.org/euclid.bsmmsp/1200512992>. [Último acceso: 14 07 2026].
- [20] P. J. Rousseeuw, «Silhouettes: A graphical aid to the interpretation and validation of cluster analysis,» Elsevier, [En línea]. Available: [https://doi.org/10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7). [Último acceso: 14 07 2026].
- [21] D. Jurafsky y J. H. Martin, «Speech and Language Processing,» Stanford University, [En línea]. Available: <https://web.stanford.edu/~jurafsky/slp3/>. [Último acceso: 14 07 2026].

- [22] J. Lyons, «Linguistic Semantics: An Introduction,» Cambridge University Press, [En línea]. Available: <https://www.cambridge.org/core/books/linguistic-semantics/760D6C09FC63FDF55FF50C1AAE5F663F>. [Último acceso: 14 07 2026].
- [23] S. Bird, E. Klein y E. Loper, «Natural Language Processing with Python,» O'Reilly Media, [En línea]. Available: <https://www.nltk.org/book/>. [Último acceso: 14 07 2026].
- [24] P. Kralj Novak, J. Smailović, B. Sluban y I. Mozetič, «Transformación de Emojis,» Public Library of Science (PLOS), 2015. [En línea]. Available: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0144296>. [Último acceso: 14 07 2026].
- [25] T. Mikolov, K. Chen, G. Corrado y J. Dean, «Efficient Estimation of Word Representations in Vector Space,» arXiv (Cornell University), [En línea]. Available: <https://arxiv.org/abs/1301.3781>. [Último acceso: 15 07 2026].
- [26] R. Mitchell, «Web Scraping with Python: Collecting More Data from the Modern Web,» CA, EE. UU.: O'Reilly Media, 2018. [En línea]. Available: <https://www.oreilly.com/library/view/web-scraping-with/9781491985564/>. [Último acceso: 15 07 2026].
- [27] Software Freedom Conservancy, «The Selenium Browser Automation Project,» Software Freedom Conservancy, 2024. [En línea]. Available: <https://www.selenium.dev/documentation/>. [Último acceso: 15 07 2026].
- [28] G. van Rossum y F. L. Drake, «The Python Language Reference,» Python Software Foundation, 2024. [En línea]. Available: <https://docs.python.org/3/>. [Último acceso: 15 07 2026].

- [29] F. Pedregosa, G. Varoquaux, A. Gramfort y V. Michel, «Scikit-learn: Machine Learning in Python,» Microtome Publishing, 2011. [En línea]. Available: <https://jmlr.org/papers/v12/pedregosa11a.html>. [Último acceso: 15 07 2026].
- [30] J. E. Grayson, «Python and Tkinter Programming,» Manning Publications, 2000. [En línea]. Available: <https://www.manning.com/books/python-and-tkinter-programming>. [Último acceso: 15 07 2026].
- [31] Asamblea Nacional del Ecuador, «Ley Orgánica de Protección de Datos Personales,» Registro Oficial del Ecuador, 26 05 2021. [En línea]. Available: <https://www.gob.ec/sites/default/files/regulations/2025-01/01%20Ley%20Org%C3%A1nica%20de%20Protecci%C3%B3n%20de%20Datos%20Personales.pdf>. [Último acceso: 19 07 2026].
- [32] Congreso Nacional del Ecuador, «Ley de Comercio Electrónico, Firmas y Mensajes de Datos,» Registro Oficial del Ecuador, 17 04 2022. [En línea]. Available: <https://www.telecomunicaciones.gob.ec/wp-content/uploads/downloads/2012/11/Ley-de-Comercio-Electronico-Firmas-y-Mensajes-de-Datos.pdf>. [Último acceso: 19 07 2026].
- [33] Asamblea Nacional del Ecuador, «Ley Orgánica de Telecomunicaciones,» Registro Oficial del Ecuador, 18 02 2015. [En línea]. Available: <https://www.telecomunicaciones.gob.ec/wp-content/uploads/downloads/2016/05/Ley-Org%C3%A1nica-de-Telecomunicaciones.pdf>. [Último acceso: 17 07 2026].
- [34] AliExpress, «Términos de uso de AliExpress.com,» AliExpress, 01 06 2026. [En línea]. Available: https://terms.alicdn.com/legal-agreement/terms/platform_service/20220531142242826/20220531142242826.html. [Último acceso: 19 07 2026].

- [35] World Intellectual Property Organization (WIPO), «Standing Committee on Copyright and Related Rights (SCCR),» WIPO, [En línea]. Available: <https://www.wipo.int/en/web/sccr>. [Último acceso: 19 07 2026].
- [36] M. J.R y W. P.R, «The Language of Evaluation,» Springer Nature, [En línea]. Available: <https://link.springer.com/book/10.1057/9780230511910>. [Último acceso: 23 07 2026].
- [37] H. Zellig. S, «Distributional Structure,» [En línea]. Available: <https://doi.org/10.1080/00437956.1954.11659520>. [Último acceso: 23 07 2026].
- [38] V. Vladimir. N, «The Nature of Statistical Learning Theory,» Springer Nature, [En línea]. Available: <https://doi.org/10.1007/978-1-4757-2440-0>. [Último acceso: 23 07 2026].
- [39] J. A. K, «Pattern Recognition Letters,» Elsevier B.V, [En línea]. Available: <https://doi.org/10.1016/j.patrec.2009.09.011>. [Último acceso: 23 07 2026].
- [40] SeleniumHQ, «Selenium WebDriver Documentation,» Selenium Project, [En línea]. Available: <https://www.selenium.dev/documentation/webdriver/>. [Último acceso: 26 07 2026].
- [41] J. Han, M. Kamber y J. Pei, «Data Mining: Concepts and Techniques,» 2012. [En línea]. Available: <https://books.google.com.ec/books?id=pQws07tdpjoC>. [Último acceso: 26 07 2026].
- [42] Python Software Foundation, «tkinter — Python interface to Tcl/Tk,» Python Documentation, [En línea]. Available: <https://docs.python.org/3/library/tkinter.html>. [Último acceso: 10 08 2026].
- [43] M. Honnibal y I. Montani, «spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing,» 2017. [En línea]. Available: <https://spacy.io/>. [Último acceso: 12 08 2026].

- [44] A. Joulin, E. Grave, P. Bojanowski y T. Mikolov, «Bag of Tricks for Efficient Text Classification,» [En línea]. Available: <https://arxiv.org/abs/1607.01759>. [Último acceso: 13 08 2026].
- [45] V. U. I. d. Valencia, «VIU Universidad Internacional de Valencia,» 29 08 2022. [En línea]. Available: <https://www.universidadviu.com/es/actualidad/nuestros-expertos/claustering-que-es-y-que-aplicaciones-tiene>. [Último acceso: 03 05 2024].
- [46] J. Suárez Mosquera, «Repository Universidad Distrital Francisco José de Caldas,» 19 Diciembre 2020. [En línea]. Available: <https://repository.udistrital.edu.co/handle/11349/29594>. [Último acceso: 03 junio 2024].
- [47] D. C. G. BRYAN y T. M. B. Gilson, «Repository UG,» 2021. [En línea]. Available: <https://repository.ug.edu.ec/items/3920dc79-f6ba-4d89-adb8-285f68f0992b>. [Último acceso: 03 junio 2024].
- [48] I. F. Mondino, «Repository UCA,» 2022. [En línea]. Available: <https://repository.uca.edu.ar/handle/123456789/15200>. [Último acceso: 03 junio 2024].
- [49] M. d. T. y. d. l. S. d. l. Información, «telecomunicaciones.gob.ec,» [En línea]. Available: <https://www.telecomunicaciones.gob.ec/wp-content/uploads/2021/06/Ley-Organica-de-Datos-Personales.pdf>.
- [50] J. C. S. Sande, «universitat oberta de catalunya,» 06 2018. [En línea]. Available: <https://openaccess.uoc.edu/server/api/core/bitstreams/da2062be-a1fb-42f0-b694-0fd4d26c7770/content>. [Último acceso: 16 04 2026].
- [51] GeeksforGeeks, «Elbow Method for optimal value of k in KMeans,» GeeksforGeeks, 11 11 2025. [En línea]. Available:

- <https://www.geeksforgeeks.org/machine-learning/elbow-method-for-optimal-value-of-k-in-kmeans/>. [Último acceso: 16 04 2026].
- [52] EBSCO Information Services, «Semántica,» EBSCO Information Services, [En línea]. Available: <https://www.ebsco.com/research-starters/language-and-linguistics/semantics>. [Último acceso: 18 04 2026].
- [53] B. Williams, «Polaridad y subjetividad en el análisis de sentimientos: explicación,» Insight7, [En línea]. Available: <https://insight7.io/polarity-and-subjectivity-in-sentiment-analysis-explained/>. [Último acceso: 18 04 2026].
- [54] Elasticsearch, «¿Qué es el análisis de sentimiento?,» Elasticsearch, [En línea]. Available: <https://www.elastic.co/es/what-is/sentiment-analysis>. [Último acceso: 16 04 2026].
- [55] L. Ramírez, «Algoritmo k-means: ¿Qué es y cómo funciona?,» IEBS Biztech School, 05 enero 2023. [En línea]. Available: <https://www.iebschool.com/blog/algoritmo-k-means-que-es-y-como-funciona-big-data/>. [Último acceso: 03 junio 2024].
- [56] The MathWorks, «Introducción a Support Vector Machines (SVM),» The MathWorks, [En línea]. Available: <https://la.mathworks.com/discovery/support-vector-machine.html>. [Último acceso: 03 junio 2024].
- [57] D. A. Morey, «Web scraping en Python con Selenium y BeautifulSoup,» Medium, 27 07 2023. [En línea]. Available: <https://medium.com/@repopkukato/web-scraping-en-python-con-selenium-y-beautiful-soup-19fccb62908e>. [Último acceso: 16 04 2026].
- [58] Plotly, «Getting Started with Plotly in Python,» Plotly, [En línea]. Available: <https://plotly.com/python/getting-started/>. [Último acceso: 26 04 2026].
- [59] G. S. Desirée, «idUS. Depósito de Investigación de la Universidad de Sevilla,» Ereding, 09 07 2021. [En línea]. Available:

<https://biblus.us.es/bibing/proyectos/abreproy/71999/fichero/TFM-1999+GARC%C3%8DA+SORIANO%2C+DESIREE.pdf>. [Último acceso: 23 04 2025].

- [60] SOTC, «Linguistics Fundamentals: Essential Concepts And Approaches,» SOTC, [En línea]. Available: <https://spires.co/online-linguistics-tutors/undergraduate/linguistics-fundamentals-essential-concepts-and-approaches>. [Último acceso: 17 04 2026].
- [61] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer y R. Wirth, «CRISP-DM 1.0: Step-by-step Data Mining Guide.,» SPSS Inc, 2000. [En línea]. Available: <https://www.dataprix.com/files/CRISP-DM.pdf>. [Último acceso: 26 07 2026].

ANEXOS

Anexo 1. Observación de comentarios Negativos relacionados con la baja calidad del producto

Este anexo presenta una muestra de comentarios recopilados de productos de diversas categorías en AliExpress, donde los usuarios han expresado insatisfacción relacionada con la calidad del producto y la discrepancia entre lo descrito y lo recibido. Los siguientes ejemplos muestran cómo los consumidores mencionan problemas como descripciones engañosas, productos defectuosos y la percepción de baja calidad.

The image shows a screenshot of an AliExpress product page for a digital camera. The product is a "Cámara Digital 4K" with a price of US \$9.31, marked down from US \$13.90 (-33% dto.). The product description mentions "50 megapíxeles de fotografía HD de viaje" and "Selfie cámara de 2,4 pulgadas de nivel de entrada para niños Campus Mini cámara Selfie". The product has 4.1 stars from 95 reviews and 700+ sold.

Below the product listing, there are three negative reviews:

- Review 1:** User "T***a" (11 MAY 2024) gave 1 star. Comment: "Color:Silver ¡Esta es la peor cámara del mundo! ¡No tiene Resolución 4K!". Includes two photos of the camera.
- Review 2:** User "A***r" (25 SEP 2024) gave 2 stars. Comment: "Color:Black Fotos de muy mala calidad... No es de 50 megapíxeles, eso es una completa mentira, sólo un sensor de 8 megapíxeles... Honestamente se siente como un juguete barato... No esperaba mucho por el precio, me decepcionó.".
- Review 3:** User "Anónimo" (12 SEP 2024) gave 2 stars. Comment: "Color:Silver No de alta resolución, la descripción es una mentira. Estafa.".

Figura 23. Anexo 1. Producto 1: Cámara digital.

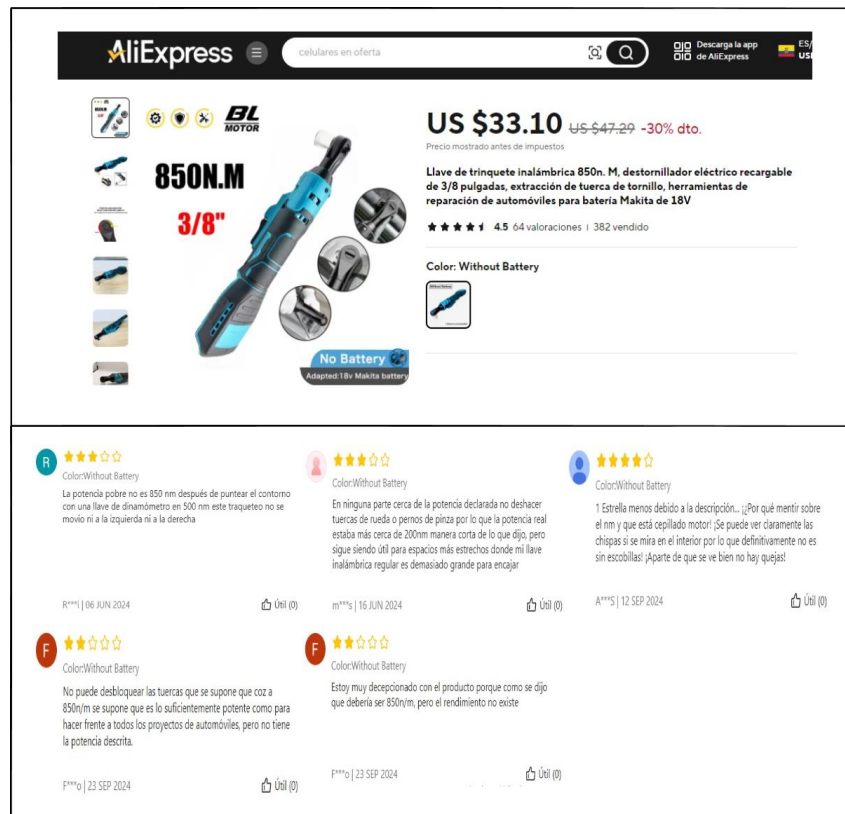


Figura 26. Anexo 1. Producto 4: Llave de trinquete inalámbrica

A través de la observación de estos comentarios Negativos, se destaca la importancia de realizar un análisis más profundo de las reseñas con el fin de identificar patrones de insatisfacción que puedan guiar mejoras en la confianza del consumidor y en la calidad de los productos ofrecidos.

Anexo 2. Observación de reseñas contradictorias en productos con alto volumen de comentarios

Este anexo presenta una muestra de comentarios recopilados de productos con un alto número de reseñas en AliExpress, donde se puede observar la variabilidad en las opiniones de los usuarios. Estos comentarios muestran tanto críticas positivas como negativas, lo que dificulta a los compradores potenciales obtener una evaluación objetiva del producto.

[audífonos bluetooth inalámbricos](#)

US \$6.01 ~~US \$12.27~~ -51% dto.

Precio mostrado antes de impuestos

Xiaomi-auriculares inalámbricos por Bluetooth, intrauditivos con micrófono incorporado

★★★★☆ 4.2 1374 valoraciones | 5.000+ vendido

Color: White 1pcs

Comentario positivo

★★★★★

Color:White 1pcs

A**** | 08 SEP 2024

Auriculares Xiaomi Bluetooth color blanco con buen sonido en ambos dispositivos derecho e izquierdo

★★★★★

Color:White 1pcs

Llegó mucho antes de lo esperado, el paquete estaba intacto, buen sonido, buena calidad, durabilidad está por probar, fácil de conectar, buen diseño, todo perfecto, recomiendo el artículo, buen producto, estoy satisfecho.

M****a | 20 JUN 2024

Útil (0)

Comentario negativo:

★★★★☆

Color:White 1pcs

(La calidad de los auriculares es repugnante, el plástico es triste, las lagunas de rejilla son visibles! El sensor vive por sí mismo. El sonido es terrible incluso para este precio, acaba de tirar el dinero por el desajuste...

K****m | 21 SEP 2024

Útil (0)

★★★★☆

Color:White 1pcs

Esto es tan falso. He comprado algunos auriculares inalámbricos por Xiaomi. De lejos, este es el peor. La batería es tan débil, me duró sólo 2 horas. La calidad del sonido es muy mala. No carga bien.

P****a | 02 AGO 2024

Útil (0)

Figura 27. Anexo 2. Producto 1: Auriculares inalámbricos.

[aspiradora portátil](#)

\$6.45 ~~\$11.95~~ -46% dto.

Precio mostrado antes de impuestos

Xiaomi-Robot aspirador automático 3 en 1 para el hogar, aspiradora pequeña lhome recargable por Usb, barredora portátil en seco y húmedo

★★★★☆ 4.5 322 valoraciones | 800+ vendido

Color: Black

Comentario positivo

★★★★★

Color:White

Me encanta esta aspiradora. Acabo de configurarlo para hacer su propia cosa en la noche y estoy muy impresionado con la cantidad de suciedad, arena, pelo de mascota que esta cosa recoge. ¡Tratar de mantener una casa limpia con 3 niños y un perro es difícil, pero este pequeño trabajador milagroso ha sido muy útil!

J****d | 09 AGO 2024

Helpful (0)

★★★★★

Color:Black

¡He intentado varios otros aspiradores, pero este es de lejos el mejor! Tiene una potente succión, una gran función de fregona, y es realmente "inteligente". Ninguna otra aspiradora en esta categoría también funciona. ¡Altamente recomendado!

Andnimo | 12 AGO 2024

Helpful (0)

Comentario negativo:

★★★★☆

Color:White

Lo siento, pero estoy muy decepcionado con la compra. Coger casi nada, se queda atascado en ciertos ángulos y luego me è que lo ha sacado de allí. Por supuesto, por el precio que aplican no puede esperar mucho, pero siendo que yo mismo tengo que aspirar todo de nuevo, è una compra innecesaria.

1****y | 05 SEP 2024

Helpful (0)

Figura 28. Anexo 2. Producto 2. Robot aspirador.

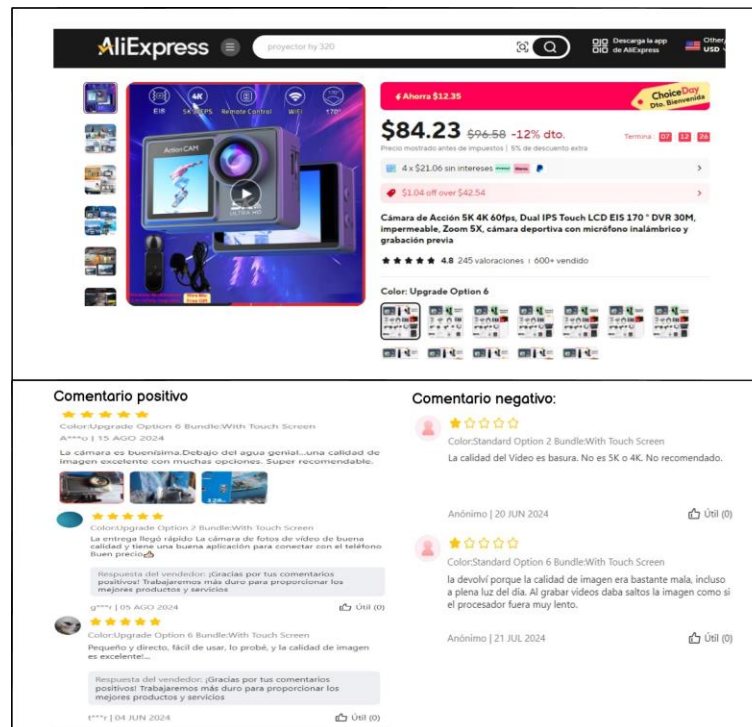


Figura 29. Anexo 2. Producto 3. Cámara deportiva.

A través de la observación de estos comentarios contradictorios, se destaca la necesidad de un análisis más detallado para identificar patrones claros en las reseñas, con el objetivo de proporcionar a los compradores potenciales una mejor orientación y mejorar la precisión de la evaluación de la calidad de los productos.

Anexo 3. Observación de patrones en comentarios de productos con grandes volúmenes de reseñas

Este anexo presenta una muestra de comentarios recopilados de productos con alto volumen de reseñas en AliExpress, donde la variedad temática de las opiniones evidencia la necesidad de aplicar técnicas de agrupamiento para organizar y clasificar las percepciones de los usuarios de manera estructurada. La diversidad de aspectos mencionados por los compradores, tales como la calidad del material, el tiempo de entrega, la correspondencia con la descripción y la funcionalidad del producto dificulta la identificación manual de los temas predominantes cuando el volumen de reseñas es elevado.

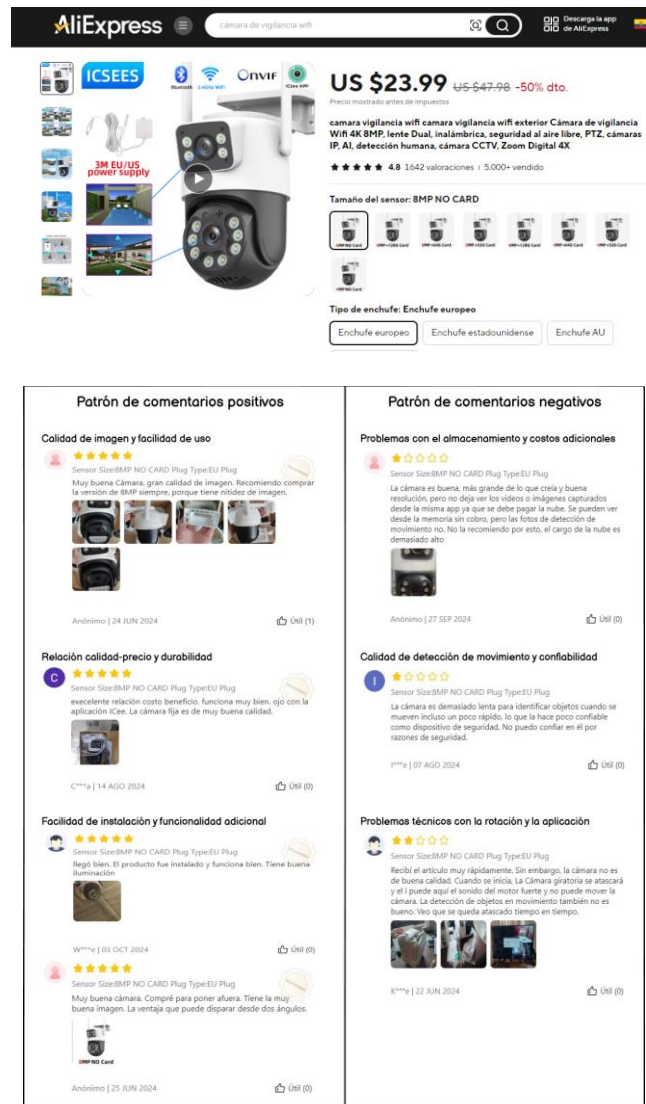


Figura 30. Anexo 3. Producto 1. Cámara vigilancia wifi



Figura 31. Anexo 3. Producto 2. Humidificador de aire.

Patrón de comentarios positivos	Patrón de comentarios negativos
Duración del uso y eficiencia ★★★★★ Color:White aunque es pequeño, consume muy poca agua. Con el depósito lleno, me ha durado toda una jornada laboral entera y me ha sobrado la mitad. Muy contento. Anónimo 21 JUL 2024 Útil (0) ★★★★★ Color:White El humidificador funciona desde la red en dos modos: antinatural e intermitentemente. Después de 3,5 horas el dispositivo se apaga. Para continuar el trabajo requiere re-inclusión. El resto del dispositivo es consistente con la descripción. Adecuado para habitaciones pequeñas de hasta 10 m. sq. n***t 22 JUL 2024 Útil (0) Diseño compacto y estética ★★★★★ Color:Green Está bien, es ligero, el color es idéntico, funciona con cable USB, tiene que estar conectado para funcionar, es para espacios pequeños y cerrados Y***S 27 AGO 2024 Útil (0) ★★★★★ Color:White Pequeño pero muy lindo. Múltiples modos y retroiluminación Anónimo 20 JUL 2024 Útil (0) Rapidez en la entrega y facilidad de uso ★★★★★ Color:White Llegó muy rápido en Brasil, producto simple pero cumple con la promesa, puede apagar el led para mantener la habitación oscura y tiene un buen flujo de agua que sale de ella Anónimo 27 JUL 2024 Útil (0) ★★★★★ Color:White Buen artículo, cumple su función. Cambia de color y tiene 2 modos de funcionamiento. Difusor continuo o difusor por ráfagas. Material algo endeble, pero sencillo y elegante J***z 04 SEP 2024 Útil (0)	Defectos y mal funcionamiento ☆☆☆☆☆ Color:White Recibí el paquete rápidamente, pero no lo recomiendo, no funciona, solo cambian los colores, la humidificación solo tiene una corriente de humo muy delgada, dinero desperdiciado a***r 18 JUL 2024 Útil (0) ☆☆☆☆☆ Color:Green Las luces no se pueden apagar, por lo que es inútil si usted está tratando de dormir. Después de un tiempo comienza a hacer un sonido de "gritos", muy probablemente defectuoso. Anónimo 09 AGO 2024 Útil (0) Publicidad engañosa y calidad de construcción ☆☆☆☆☆ Color:White El dispositivo enviado no es el anuncio, ya que no es Xiaomi, ni es recargable. El anuncio electrónico no es el que aparece en él: puro engaño con los clientes A***s 23 JUL 2024 Útil (0) ☆☆☆☆☆ Color:White No es Xiaomi marca (Mi) es un modelo sin marca muy genérico. La caja llegó bien abollada y la construcción de la placa anunciada es diferente de la Enviada. La conexión USB vino hundido, tuvo que hacer cierta fuerza para conectar. Totalmente insatisfecho. L***j 06 AGO 2024 Útil (0) Problemas con el embalaje y tamaño ☆☆☆☆☆ Color:Green El artículo vino con la caja abollada y el humidificador no es de la marca Xiaomi a***n 03 AGO 2024 Útil (0) ☆☆☆☆☆ Color:White El producto no es marca Xiaomi y comparar con la imagen es muy pequeño en realidad y***i 09 OCT 2024 Útil (0)

A través de la observación de estos patrones temáticos, se justifica el uso del algoritmo K-Means en la Fase 4 del sistema, el cual permite agrupar automáticamente los comentarios según sus similitudes semánticas para facilitar una interpretación más clara y estructurada de las opiniones de los consumidores.