



UNIVERSIDAD NACIONAL  
AUTÓNOMA DE MÉXICO

Facultad de Ingeniería

## AMAZON SENTIMENT ANALYSIS

Análisis de sentimientos en reseñas de Amazon

### **Alumnos:**

Basile Álvarez Andrés José  
Ceres Martínez Hanna Sofía  
Keller Ascencio Rodolfo Andrés

**Profesor:** M.P. Octavio Augusto Sánchez Velázquez

**Grupo:** 01

19 de junio de 2023



# Índice

<b>1. Resumen</b>	<b>3</b>
<b>2. Abstract</b>	<b>3</b>
<b>3. Introducción</b>	<b>3</b>
3.1. Objetivo . . . . .	4
3.2. Motivación . . . . .	4
3.3. Posible Impacto . . . . .	4
<b>4. Estado de la cuestión</b>	<b>4</b>
4.1. Documento “Sentiment and Topic Modeling on Social S. Books” . . . . .	4
4.2. Documento “Amazon Product Sentiment Analysis using BERT” . . . . .	5
4.3. Documento “A machine learning-based sentiment analysis of online product reviews with a novel term weighting and feature selection approach.” . . . . .	5
4.4. Documento “How to Fine-Tune BERT for Text Classification?” . . . . .	5
<b>5. Marco Teórico</b>	<b>6</b>
5.1. <i>Bidirectional Encoder Representations from Transformers (BERT)</i> . . . . .	6
5.2. DistilBERT . . . . .	6
5.3. Latent Dirichlet Allocation (LDA) . . . . .	6
5.4. Optimizador ADAM . . . . .	7
5.5. Learning rate . . . . .	7
5.6. Épocas . . . . .	7
5.7. Herramientas de conocimiento necesarias para desarrollar el proyecto . . . . .	7
<b>6. Metodología, Experimentación y Resultados</b>	<b>8</b>
6.1. Entendimiento del negocio . . . . .	8
6.2. Entendimiento del Conjunto de Datos . . . . .	8
6.2.1. Análisis y preprocesamiento del conjunto de datos inicial . . . . .	9
6.2.2. Selección y análisis de un conjunto reducido de datos . . . . .	12
6.3. Análisis de Frecuencia de Tokens . . . . .	14
6.3.1. Frecuencia de Palabras por Aparición . . . . .	14
6.3.2. Frecuencia TF-IDF . . . . .	15
6.3.3. Frecuencia TF-IDF texto Lematizado . . . . .	15
6.4. WordClouds . . . . .	16
6.4.1. WordClouds por nombre del producto en categorías . . . . .	16
6.4.2. WordClouds por Tipo y Reseña . . . . .	17
6.4.3. WordClouds de reseñas por cantidad de estrellas . . . . .	20
6.5. Análisis de tópicos . . . . .	22
6.5.1. Análisis de tópicos en conjunto reducido de datos . . . . .	22



---

6.5.2. Análisis de tópicos en conjunto de datos de interés . . . . .	23
6.6. Análisis de sentimientos con DistilBERT . . . . .	25
<b>7. Conclusión</b>	<b>27</b>
<b>8. Bibliografía</b>	<b>28</b>
<b>9. Anexo</b>	<b>29</b>
9.1. WordClouds por nombre del producto en categorías . . . . .	29
9.2. Tópicos del conjunto reducido de datos. . . . .	30
9.3. Tópicos del conjunto de datos de interés. . . . .	33



## 1. Resumen

En los últimos años, el análisis de textos se ha convertido en un área de investigación de interés, con lo cual, aunado al aumento de popularidad del comercio electrónico, podemos encontrar una necesidad en el análisis de sentimientos a partir de las reseñas de productos de venta en línea. En este sentido, este proyecto propone diversos métodos para el análisis de reseñas de productos para el beneficio de empresas, enfocándonos en el análisis de reseñas para beneficio de nuestra propia empresa para el desarrollo de calentadores de agua eléctricos. Finalmente, proponemos el uso de DistilBERT como un modelo para el análisis de reseñas para la clasificación en número de estrellas para que, de esta forma, una empresa pueda tener un mejor control de la calidad de sus productos, así como en el proceso de venta, garantía y servicio al cliente.

**Palabras clave**— Análisis de sentimientos, DistilBERT, Amazon, Reseñas, Estrellas, Calentadores de agua eléctricos, Acero inoxidable

## 2. Abstract

In recent years, text analysis has become an interesting area of research. Coupled with the increasing popularity of e-commerce, we can identify a need or opportunity in sentiment analysis based on online product reviews. This project proposes various methods for analyzing reviews of products for the benefit of companies, focusing on the analysis of reviews for the benefit of our own company, specifically in the development of electric water heaters. Therefore, we propose the use of DistilBERT as a model for review analysis to classify them based on star ratings. This way, a company can have better control over the quality of its products, as well as improve the sales process, warranty, and customer service.

**Keywords**— Sentiment Analysis, DistilBERT, Amazon, Reviews, Stars, Electric Water Heater, Stainless Steel

## 3. Introducción

Amazon es una empresa estadounidense fundada en 1994 por Jeff Bezos, empresa la cual se ha convertido en una de los gigantes del comercio electrónico a nivel mundial. Además de su amplia selección de productos, una de las características que destaca en Amazon es su sistema de reseñas por parte de los usuarios.

La empresa utiliza herramientas de análisis de opiniones para analizar las reseñas que los usuarios escriben sobre sus productos. Según Amazon, estas herramientas permiten determinar el tono emocional de la reseña y clasificarla como positiva, negativa o neutra, siendo este un análisis que permite que los proveedores y vendedores de productos en Amazon puedan reconocer las cualidades de sus productos, así como las áreas de oportunidad de los mismos.Services (sf)

En la actualidad, las empresas tienen grandes cantidades de datos de texto que incluyen correos electrónicos, chats de atención al cliente, comentarios en redes sociales y reseñas. Las herramientas de análisis de opiniones pueden escanear este texto automáticamente para determinar la actitud del autor hacia un tema. Las empresas pueden utilizar esta información para mejorar el servicio al cliente y mejorar la imagen de la marca.Services (sf)

El análisis de opiniones es de suma importancia para empresas como Amazon y el campo del marketing. En el caso de Amazon, el análisis de opiniones ayuda a mejorar la experiencia del cliente al identificar sus preferencias y necesidades. Asimismo, permite gestionar la reputación y el branding al monitorear y abordar rápidamente problemas o quejas, así como fortalecer la imagen de la empresa con opiniones positivas. En el ámbito del marketing, el análisis de opiniones ayuda a identificar oportunidades de mejora en las estrategias de marketing y comunicación, además de permitir la identificación de tendencias y preferencias del consumidor.



Esta capacidad permite a las empresas adaptarse a las cambiantes necesidades de los clientes y tomar decisiones informadas para mantenerse competitivas en el mercado. En particular, si consideramos la opinión de una empresa que vende sus productos en Amazon y está buscando realizar un análisis de mercado para los productos que está a punto de lanzar, es aún más crucial profundizar en este aspecto. Esto ayudaría a la empresa a identificar áreas de mejora y mantenerse atenta a los detalles que podrían marcar la diferencia en su estrategia de mercado, siendo ésta la motivación principal de nuestro proyecto.

El análisis de opiniones funciona con tecnologías de Procesamiento de Lenguaje Natural (PLN) que entrena al software de computación para que pueda ser capaz de entender textos de manera similar a los seres humanos. En el preprocesamiento, el análisis de opiniones identifica las palabras clave que destacan el mensaje principal del texto. Esto se logra a través de la *tokenización*, que divide la oración en elementos individuales llamados *tokens*. Además, se utiliza la lematización para convertir las palabras a su forma raíz y la eliminación de palabras vacías para filtrar las palabras que no aportan un valor significativo a la oración. Services (sf)

### 3.1. Objetivo

El principal objetivo general de nuestro proyecto se centra en realizar un análisis de sentimientos de las reseñas de productos de Amazon con el fin de comprender qué buscan los clientes. Se busca identificar las características de los productos que las personas resaltan en las reseñas positivas y negativas, y utilizar esta información para desarrollar nuestros propios productos.

Por otro lado, se busca obtener una herramienta que sea capaz de categorizar reseñas de clientes dándoles como valor final un número de estrellas relacionado con el comentario o reseña que hagan de un producto. Se busca encontrar una herramienta que sea capaz de realizar esto para poder obtener métricas de clientes que se encuentren fuera de plataformas como Amazon, implementando nuestro propio sistema de recolección de opiniones y reseñas.

Como objetivo particular, mediante el análisis de las opiniones de los clientes, se pretende identificar patrones y tendencias en las preferencias y necesidades de los usuarios relacionados a productos específicos.

### 3.2. Motivación

Contamos con una empresa de calentadores eléctricos de agua de alta calidad, donde nuestra empresa podrá utilizar este análisis de reseñas y estos datos para realizar ajustes y mejoras en los productos que se encuentra diseñando y que está por lanzar al mercado, adaptándolos a las demandas del mercado y satisfaciendo las expectativas de los clientes de manera más efectiva aún antes de realizar campañas para la venta del producto. El proyecto tiene como objetivo final proporcionar a nuestra propia empresa información valiosa y accionable basada en las reseñas de productos, con el fin de impulsar la innovación y la mejora continua de nuestros productos y servicios.

### 3.3. Posible Impacto

Consideramos que este proyecto puede impactar de forma positiva el futuro de una empresa, debido a que el prestigio de una empresa resulta sumamente importante, por lo tanto, antes de lanzar un nuevo producto en desarrollo por parte de una empresa el análisis de productos que ya se encuentran en el mercado y de las opiniones de los clientes puede resultar crucial para el éxito de un producto y el crecimiento de la empresa.

## 4. Estado de la cuestión

### 4.1. Documento “Sentiment and Topic Modeling on Social S. Books”

El proyecto “Sentiment and Topic Modeling on Social S. Books” realizado por el usuario HalilErgul en la plataforma Kaggle utiliza modelados de tópicos, *Word-Clouds*, redes neuronales de tipo DistilBERT y modelado de temas con *Latent Dirichlet Allocation* (LDA) para realizar un análisis de sentimientos en las reseñas de los



consumidores escritas para algunos libros de ciencias sociales y abordar la cuestión de hasta qué punto existe una diferencia sentimental significativa y una variación basada en el tema entre estas reseñas en varias disciplinas de las ciencias sociales, como la psicología, la política y sociología, entre otras. Como resultado de su preprocesamiento de datos y uso del modelo DistilBERT, el modelo de sentimiento basado en BERT entrenado obtuvo buenos resultados al etiquetar el sentimiento asociado de un libro determinado. Además, el modelado de temas con LDA a través de la biblioteca Gensim funciona de buena forma, aunque requiere algunas mejoras, como sugieren su puntuación de coherencia y perplejidad. HalilErgul (2022)

Pese a estas observaciones, la mayoría de nuestro proyecto se encuentra basado en el proceso que realiza este autor en el desarrollo de su proyecto, debido a que los resultados que obtuvo este usuario se relacionan directamente con los resultados que buscamos obtener al finalizar nuestro trabajo, donde la diferencia radica en el hecho de que nosotros nos encontraremos trabajando bajo ciertos productos de interés, en lugar de libros por categoría.

## 4.2. Documento “Amazon Product Sentiment Analysis using BERT”

En el artículo publicado por el autor Yash Inaniya para *Analytics Vidhya*, utilizan un modelo de clasificación preentrenado “DistilBert” para realizar un análisis de sentimientos en las reseñas de productos electrónicos y móviles de Amazon, haciendo uso de un conjunto similar al que planeamos utilizar. El proceso que se encuentra desarrollado en este documento resulta parecido al que se planteó en el proyecto que realizamos, con la diferencia de que en este caso se hace uso del modelo “TFDistilBertForSequenceClassification”, pero con un entrenamiento de dos épocas, un *batch size* de dieciséis y una tasa de aprendizaje de 0.00005 haciendo uso del optimizador Adam. Su modelo dió como resultado una precisión mayor del 90 %, sin embargo, se tiene que tomar en cuenta que en el caso de este documento la clasifica-

ción de reseñas se realiza a tres clases, dando una clasificación positiva, neutra y negativa, en lugar de un valor numérico de número de estrellas como lo que buscamos con nuestro proyecto. Finalmente, su modelo funcionaba de forma excelente para predecir el sentimiento en las oraciones de las reseñas, siendo un modelo más ligero y rápido de implementar que un BERT normal. Inaniya (2021)

## 4.3. Documento “A machine learning-based sentiment analysis of online product reviews with a novel term weighting and feature selection approach.”

En este trabajo publicado en la revista *ScienceDirect* se propone el uso de LSIBA-ENN, un modelo de aprendizaje automático, para analizar las polaridades de las reseñas de productos en línea. Se realizaron experimentos utilizando dos conjuntos de datos de aplicaciones y películas/programas de televisión recopilados de Amazon. Los resultados muestran que LSIBA-ENN obtuvo el mejor rendimiento en comparación con otros clasificadores existentes. Además, al utilizar el método de ponderación de términos propuesto (LTF-MICF), se lograron obtener los resultados más eficientes en las métricas de clasificación, trabajando con valores de precisión mayores al 85 %. Los valores de precisión de LSIBA-ENN fueron superiores a los de los clasificadores existentes en ambos conjuntos de datos. Esto demuestra que LSIBA-ENN es un clasificador eficaz para el análisis de sentimientos de reseñas de productos. Pese a esto, nosotros preferimos hacer uso de DistilBERT como modelo para nuestro proyecto. Zhao (2021)

## 4.4. Documento “How to Fine-Tune BERT for Text Classification?”

En este trabajo se investiga el pre-entrenamiento de modelos de lenguaje, enfocándose en BERT, que ha demostrado excelentes resultados en tareas de comprensión del lenguaje. Se realizan experimentos exhaustivos para



explorar diferentes métodos de ajuste fino de BERT en la clasificación de texto, obteniendo una solución general para su ajuste. Los hallazgos experimentales incluyen el uso de la capa superior de BERT para la clasificación de texto, la disminución de la tasa de aprendizaje por capa para mejorar la precisión, y el beneficio del pre-entrenamiento adicional dentro de la misma tarea y dominio. Se demuestra que BERT puede mejorar el rendimiento incluso con conjuntos de datos pequeños. Los resultados obtenidos superan el estado del arte en ocho conjuntos de datos ampliamente estudiados dentro de este documento. Se plantea la necesidad de una mayor comprensión del funcionamiento de BERT en futuras investigaciones. Sun (2020)

## 5. Marco Teórico

En esta sección se proporciona una base sólida para entender los conceptos esenciales y las herramientas requeridas para desarrollar el proyecto.

### 5.1. Bidirectional Encoder Representations from Transformers (BERT)

BERT es un modelo de lenguaje basado en la arquitectura *Transformer*. Se entrenó en un *corpus* masivo de texto sin supervisión y logró avances significativos en varias tareas de procesamiento del lenguaje natural (NLP), como la clasificación de texto, la extracción de información y la traducción automática. BERT utiliza un enfoque bidireccional para capturar el contexto de una palabra en función de las palabras que la rodean, lo que le permite comprender mejor las sutilezas del lenguaje natural. Delvin (2018)

Para lograr esto BERT realiza lo siguiente:

- *Masked Language Model (MLM)*: BERT oculta aleatoriamente algunas palabras en las oraciones de entrada y luego intenta predecir esas palabras ocultas. Esto obliga al modelo a comprender el contexto y las relaciones entre las palabras circundantes para realizar una predicción precisa.

- *Next Sentence Prediction (NSP)*: BERT recibe pares de oraciones y debe predecir si la segunda oración sigue a la primera en el texto original. Esto ayuda a BERT a entender la relación entre las oraciones y capturar la coherencia del texto en un nivel más alto. Horev (2021)

### 5.2. DistilBERT

DistilBERT es una versión más pequeña y compacta de BERT. Fue desarrollado para ser más eficiente en términos de recursos computacionales y memoria, al tiempo que mantiene un rendimiento comparable a BERT en varias tareas de NLP. DistilBERT logra esto utilizando técnicas de destilación de conocimiento para transferir el conocimiento de BERT a un modelo más compacto.

El proceso de destilación del conocimiento implica entrenar inicialmente un modelo BERT completo en un gran *corpus* de texto, utilizando las tareas de pre-entrenamiento de BERT, como *Masked Language Model (MLM)* y *Next Sentence Prediction (NSP)*. Una vez que BERT está pre-entrenado, se toman los pesos y las capas del modelo BERT y se transfieren a DistilBERT. Sanh (2019)

### 5.3. Latent Dirichlet Allocation (LDA)

LDA es un modelo probabilístico utilizado para el análisis de temas en textos. Se basa en la suposición de que cada documento se compone de múltiples temas y que cada tema se compone de palabras con una cierta probabilidad. LDA es capaz de identificar y agrupar automáticamente las palabras en temas significativos sin la necesidad de etiquetas o conocimiento previo.

El proceso de inferencia en LDA implica la estimación de dos distribuciones:

- Distribución de temas en los documentos: Se calcula la probabilidad de que un documento pertenezca a cada tema.
- Distribución de palabras en los temas: Se calcula la probabilidad de que una palabra pertenezca a



cada tema.

Blei (2003)

## 5.4. Optimizador ADAM

ADAM (*Adaptive Moment Estimation*) es un algoritmo de optimización utilizado para ajustar los pesos y las tasas de aprendizaje en los modelos de aprendizaje automático. Combina los beneficios del optimizador de descenso de gradiente estocástico (SGD) y el método de estimación de momento para converger de manera más eficiente a una solución óptima. ADAM adapta las tasas de aprendizaje para cada parámetro individual y realiza actualizaciones más rápidas en las direcciones donde los gradientes son consistentes. Kingma (2014)

## 5.5. Learning rate

La tasa de aprendizaje es un hiperparámetro crítico en el entrenamiento de modelos de aprendizaje automático. Representa la magnitud de los ajustes realizados en los pesos del modelo durante el proceso de entrenamiento. Una tasa de aprendizaje adecuada es crucial para un entrenamiento estable y eficiente. Si la tasa de aprendizaje es demasiado baja, el modelo puede converger lentamente o quedarse atascado en mínimos locales. Por otro lado, si la tasa de aprendizaje es demasiado alta, el modelo puede oscilar o saltar sobre el mínimo global. Brownlee (2018)

## 5.6. Épocas

En el contexto del aprendizaje automático, una época se refiere a una pasada completa a través del conjunto de datos de entrenamiento durante el proceso de entrenamiento de un modelo. Cada época implica alimentar todos los ejemplos del conjunto de datos al modelo y ajustar los pesos del modelo en función de los errores cometidos. El número de épocas es un hiperparámetro que determina cuántas veces se repetirá este proceso. Una época insuficiente puede resultar en un modelo subentrenado, mientras que un número excesivo de épocas puede

llevar a un sobreajuste del modelo a los datos de entrenamiento. Es importante encontrar un equilibrio adecuado al determinar el número de épocas para obtener un modelo con un buen rendimiento generalizado.

## 5.7. Herramientas de conocimiento necesarias para desarrollar el proyecto

- Fundamentos de procesamiento del lenguaje natural (NLP): Es importante tener una comprensión sólida de los conceptos y técnicas utilizados en el procesamiento del lenguaje natural. Esto incluye el conocimiento de modelos de lenguaje, técnicas de representación de texto, algoritmos de clasificación y agrupamiento, así como la comprensión de métricas de evaluación y evaluación de modelos de NLP.
- Conocimiento de modelos de aprendizaje automático: Es esencial estar familiarizado con los fundamentos del aprendizaje automático y los diferentes tipos de modelos utilizados en NLP, como redes neuronales, modelos de secuencia, modelos de atención y modelos generativos. Esto incluye el conocimiento de cómo entrenar, ajustar y evaluar estos modelos.
- Experiencia con bibliotecas y *frameworks* de NLP: Es necesario tener experiencia con bibliotecas y *frameworks* populares utilizados en NLP, como *TensorFlow*, *PyTorch* o *NLTK (Natural Language Toolkit)*. Estas herramientas proporcionan implementaciones eficientes y optimizadas de algoritmos de NLP y facilitan el desarrollo y la experimentación con modelos de lenguaje.
- Habilidades de programación en lenguajes como Python: Python es uno de los lenguajes más utilizados en el campo de la inteligencia artificial y el procesamiento del lenguaje natural. Es importante tener habilidades sólidas de programación en Python para implementar algoritmos, manipular



datos, crear modelos de lenguaje y realizar análisis de textos.

- Conocimientos en preprocesamiento de texto: Antes de aplicar modelos de NLP, es necesario realizar un preprocesamiento adecuado de los textos. Esto incluye técnicas como tokenización, eliminación de *stopwords*, normalización de texto y codificación de palabras. Familiarizarse con estas técnicas y saber cómo aplicarlas de manera eficiente es fundamental para obtener buenos resultados.
- Capacidad para interpretar y visualizar resultados: Una parte crucial del desarrollo de proyectos de NLP es la capacidad de interpretar y visualizar los resultados. Esto implica comprender las salidas de los modelos, analizar las predicciones y evaluar el rendimiento del modelo utilizando métricas adecuadas. También se requiere la capacidad de presentar los resultados de manera clara y comprensible para diferentes audiencias.

## 6. Metodología, Experimentación y Resultados

La metodología que usaremos para desarrollar esta investigación es CRISP-DM (*Cross-Industry Standard Process for Data Mining*). Al seguir la metodología CRISP-DM, la empresa podrá obtener una comprensión más profunda de las opiniones de los clientes sobre los productos existentes en el mercado.

A continuación se describe cómo aplicamos esta metodología para desarrollar la investigación.

### 6.1. Entendimiento del negocio

En este proyecto, contamos con una empresa mexicana de calentadores eléctricos de agua de alta calidad, la cual busca analizar las reseñas de los calentadores de agua eléctricos y productos similares en Amazon, como lo serían productos de acero inoxidable, productos de la

competencia, así como elementos que forman parte de los calentadores. El objetivo que la empresa busca satisfacer sería comprender las necesidades y preferencias de los clientes en relación a este tipo de productos, con lo cual la empresa podrá familiarizarse con los aspectos más relevantes que deberá cuidar al momento de desarrollar y lanzar su nuevo producto al mercado. Buscamos identificar las características clave mencionadas en las reseñas positivas y negativas para crear y mejorar nuestros propios calentadores de agua eléctricos. Este análisis nos permitirá adaptar nuestros productos a las demandas del mercado, satisfacer las expectativas de los clientes y fomentar la mejora continua en nuestra línea de calentadores de agua eléctricos.

Por otro lado, en un futuro se busca que la misma empresa cuente con un sistema de gestión de reseñas y opiniones, por lo cual será necesario desarrollar un método que permita a la empresa reconocer el número de estrellas que los usuarios le pondrían a sus productos, sin la necesidad de que ellos directamente indiquen este valor.

### 6.2. Entendimiento del Conjunto de Datos

Los datos que utilizaremos para el desarrollo de nuestro proyecto provienen directamente de la página de Amazon, siendo éstas las reseñas de productos de cuatro distintas categorías elegidas por nosotros, en los que podemos encontrar productos electrónicos, muebles, productos de oficina y herramientas. Este conjunto de datos se obtuvo de la página web <https://s3.amazonaws.com/amazon-reviews-pds/readme.html> y contiene las opiniones y calificaciones de los clientes sobre estos productos que nos ayudará en el análisis.

Diccionario de datos:

- **marketplace**: Código de país de dos letras del mercado donde se escribió la reseña.
- **customer\_id**: Identificador aleatorio que se puede utilizar para agrupar las reseñas escritas por un



mismo autor.

- review\_id: ID único de la reseña.
- product\_id: ID único del producto al que se refiere la reseña.
- product\_parent: Identificador aleatorio que se puede utilizar para agrupar las reseñas del mismo producto.
- product\_title: Título del producto.
- product\_category: Categoría amplia del producto que se puede utilizar para agrupar las reseñas (también se utiliza para dividir el conjunto de datos en partes coherentes).
- star\_rating: Calificación de 1 a 5 estrellas de la reseña.
- helpful\_votes: Número de votos útiles.
- total\_votes: Número total de votos que recibió la reseña.
- vine: La reseña se escribió como parte del programa Vine.
- verified\_purchase: La reseña es de una compra verificada.
- review\_headline: El título de la reseña.
- review\_body: El texto de la reseña.
- review\_date: La fecha en que se escribió la reseña.

De este conjunto completo de datos se decidió trabajar únicamente con los siguientes valores:

Diccionario de datos final:

- product\_title: Título del producto.
- product\_category: Categoría amplia del producto que se puede utilizar para agrupar las reseñas (también se utiliza para dividir el conjunto de datos en partes coherentes).

- star\_rating: Calificación de 1 a 5 estrellas de la reseña.
- review\_headline: El título de la reseña.
- review\_body: El texto de la reseña.

Nuestro conjunto de datos contaba con 7,427,551 reseñas de cuatro categorías diferentes, 2,891,062 reseñas de la categoría *Electronics*, 2,406,499 reseñas de la categoría *Office Products*, 1,574,301 reseñas de la categoría *Tools* y 762,582 reseñas de la categoría *Furniture*.

### **6.2.1. Análisis y preprocesamiento del conjunto de datos inicial**

La preparación de los datos es una etapa esencial en el análisis de sentimientos de las reseñas de los productos. A continuación, se detallan las acciones realizadas durante esta fase:

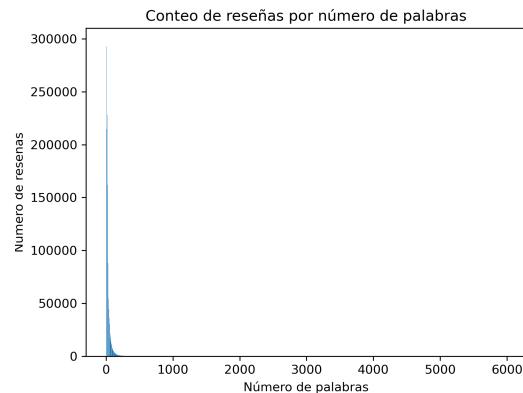
- Descarga de datos: Se seleccionaron y descargaron los cuatro conjuntos de datos antes mencionados a partir de las categorías que pudieran ser de interés para la empresa.
- Unión de conjuntos de datos: Se combinan todos los conjuntos de datos disponibles en uno único. Esto asegura que tengamos una visión completa y consistente de todas las reseñas de las cuatro categorías.
- Filtrado de columnas: Se realiza un filtrado de las columnas para retener solo las que vamos a utilizar en el análisis, como el título del producto, la categoría del producto, el número de estrellas de la reseña, el título de la reseña y el cuerpo de la reseña. Eliminamos cualquier otra columna que no sea relevante para nuestro objetivo.
- Eliminación de elementos vacíos: A través de toda la información buscamos elementos vacíos en las reseñas para que éstos sean eliminados debido a que no nos sirven para el análisis.

- Eliminación de elementos duplicados: A través de toda la información buscamos elementos duplicados para que éstos sean eliminados.
- Concatenación del título y cuerpo de las reseñas: Se fusionan el título y el cuerpo de las reseñas en un único texto. Esto nos permite tener una representación completa de la reseña, que puede contener información relevante para el análisis de sentimientos, donde tanto el título como el cuerpo de la reseña contienen información fundamental de la opinión del cliente hacia un producto.
- Preprocesamiento de los títulos de los productos y de las reseñas: Ambos conjuntos de interés se someten a técnicas de preprocesamiento, como la eliminación de *stopwords* (palabras sin significado para el sentimiento de las reseñas) y la eliminación de contracciones. Esto ayuda a reducir el ruido y a obtener una representación más limpia de los textos al eliminar palabras que no aportan información relevante y a normalizar el texto para un análisis más preciso.

Una vez completada la preparación de los datos, estaremos listos para realizar un análisis más completo del conjunto de reseñas y de sus características, así como proceder posteriormente al análisis de sentimientos en las reseñas de los productos. Los datos habrán sido filtrados, unificados y preprocesados de manera adecuada, lo que nos permitirá obtener resultados más precisos y significativos.

Como siguiente paso, se buscó realizar un análisis cuantitativo del número de palabras contenido en las reseñas de los productos, para que, de esta manera, pudiéramos apreciar si existía algún tipo de relación entre el número de palabras dentro de una reseña con el número de estrellas que el usuario le daba a un producto.

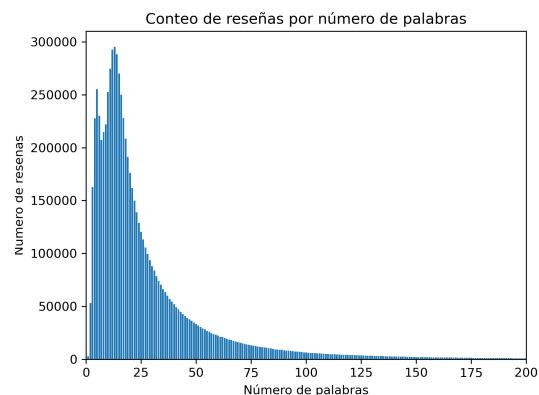
Para esto, decidimos graficar el número de reseñas por número de palabras contenidas en ella, con lo cual de manera global podemos apreciar el siguiente gráfico.



**Figura 1:** Conteo de reseñas por número de palabras.

Para esto, nos dimos cuenta de que contábamos con trece reseñas que contenían reseñas con entre cuatro mil y seis mil palabras, siendo estas las reseñas con mayor número de palabras, donde estas reseñas en su mayoría pertenecían a la categoría de Electrónicos, y habían punтуado con cinco estrellas a sus productos.

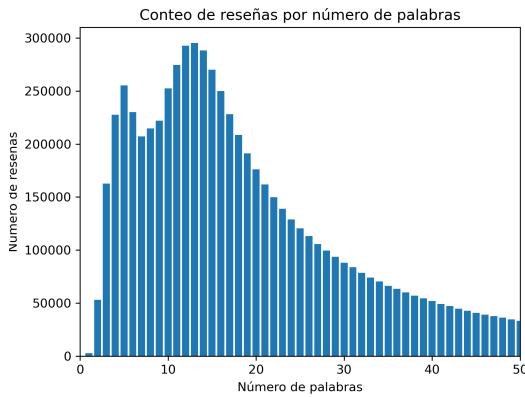
Para poder hacer un análisis más detallado decidimos acotar el número de palabras de análisis a un rango de entre cero y doscientas palabras, donde esta distribución de reseñas se muestra en el siguiente gráfico.



**Figura 2:** Conteo de reseñas acotado a 200 palabras.

En este gráfico se observa de mejor forma la distribución de las reseñas por números de palabras, sin embar-

go, se aprecia un fuerte sesgo positivo hacia la derecha, con lo cual en un último gráfico decidimos realizar una visualización de un rango que va de cero a cincuenta palabras.



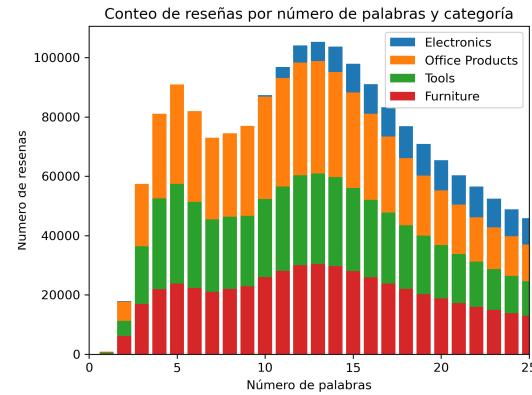
**Figura 3:** Conteo de reseñas acotado a 50 palabras.

A partir de este último gráfico, podemos apreciar que para nuestro conjunto de datos los usuarios prefieren realizar reseñas de trece palabras, mientras que el mayor número de reseñas cuentan con entre dos y cuarenta palabras.

Para ahondar aún más en nuestro análisis, decidimos también analizar cómo se comportaba esta distribución del número de palabras para cada categoría, con lo cual decidimos agrupar estos valores por número de palabras y categorías, siendo que el resultado obtenido es el siguiente.

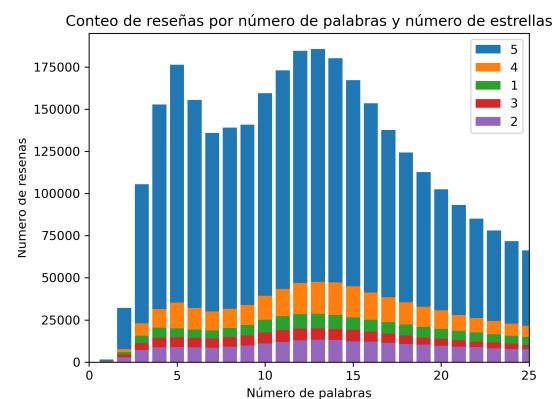
Este gráfico nos indica que contamos con un mayor número de reseñas de la categoría de Electrónicos y Productos de Oficina comparado con el número de reseñas de productos pertenecientes a las categorías de Herramientas y Muebles. Por otra parte, se aprecia una distribución similar del número de palabras por reseña en cada una de las cuatro categorías, alcanzando un mayor número de reseñas que contienen alrededor de trece palabras, con un pico significativo en las reseñas de cinco palabras.

Otro de los gráficos que consideramos sería de interés



**Figura 4:** Conteo de reseñas por número de palabras y categoría.

se presentaba con la relación que existía entre el número de palabras contenidas en las reseñas y el número de estrellas que puntuaba una persona.



**Figura 5:** Conteo de reseñas por número de palabras y número de estrellas.

A partir de este análisis nos dimos cuenta que la distribución de las reseñas a partir del número de palabras no tiene una relación directa con el número de estrellas con las que un usuario puntuó un producto, es decir, a mayor número de palabras utilizadas en una reseña no quiere decir que el usuario va a dar un mejor o peor puntaje a un producto. Por otra parte, nos dimos cuenta que el número de reseñas de productos puntuados con cinco



estrellas era bastante mayor con respecto a las reseñas con menor puntuación.

### 6.2.2. Selección y análisis de un conjunto reducido de datos

Tras haber realizado el análisis global de todo el conjunto inicial de datos que contenía más de siete millones de reseñas, decidimos acotar nuestro conjunto de datos para trabajar con un conjunto de mayor interés, para esto, decidimos reducir el conjunto de datos en dos subconjuntos.

El primer subconjunto lo denominados como *filtered\_data*, el cual contenía reseñas de productos con palabras clave de nuestro interés en su nombre. Decidimos filtrar el conjunto global de datos buscando productos de marcas competitivas para nuestro producto de interés, de componentes que se relacionaran con el producto final, o de sinónimos y productos hermanos al producto a desarrollar por nuestra empresa, así como el material principal del producto que nos encontramos desarrollando, de esta forma seleccionamos productos que se relacionaran con alguna de las siguientes diez categorías:

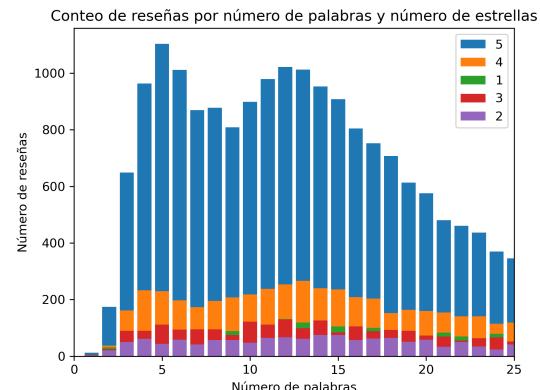
- Stiebel
- Rheem
- Siemens
- Thermocouple
- Temperature Controller
- Tankless Water Heater
- Electric Water Heater
- Boiler
- Water Heater
- Stainless Steel

```
Number of reviews with ' stainless steel ' are: 37356
Number of reviews with ' stainless steel box ' are: 100
Number of reviews with ' water heater ' are: 653
Number of reviews with ' boiler ' are: 60
Number of reviews with ' electric water heater ' are: 8
Number of reviews with ' tankless water heater ' are: 157
Number of reviews with ' thermocouple ' are: 381
Number of reviews with ' temperature controller ' are: 289
Number of reviews with ' stiebel ' are: 9
Number of reviews with ' rheem ' are: 27
Number of reviews with ' siemens ' are: 1981
```

**Figura 6:** Número de reseñas por tipo de producto de interés.

De esta forma, el número de reseñas por cada una de las categorías antes mencionadas quedó de la siguiente manera:

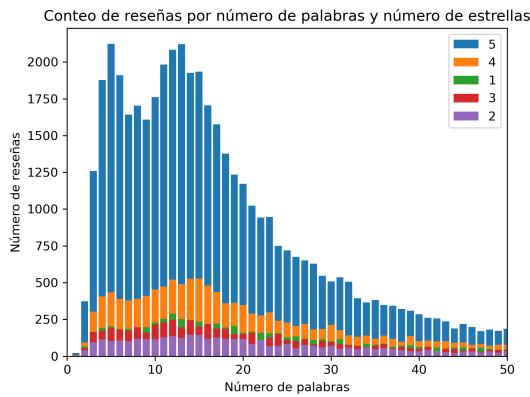
Donde contamos con 43,519 reseñas de interés para nuestro análisis. La relación de las reseñas con el número de estrellas dentro de nuestro conjunto de interés quedaría entonces de la siguiente manera:



**Figura 7:** Conteo de reseñas de conjunto de interés por número de palabras y número de estrellas.

Tras esto, decidimos aumentar nuestro conjunto de interés a partir de la selección aleatoria de diez mil reseñas por cada una de las cuatro categorías de nuestro conjunto de datos, para que, de esta manera, contáramos con más de ochenta mil reseñas, donde poco más de la mitad de estas reseñas del conjunto reducido pertenecían a nuestro conjunto de interés. La relación de las reseñas

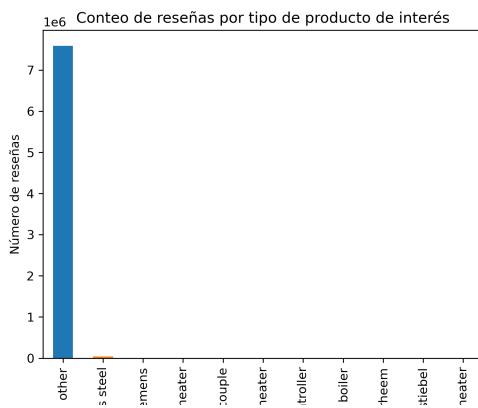
con el número de estrellas dentro de nuestro conjunto reducido quedaría entonces de la siguiente manera:



**Figura 8:** Conteo de reseñas de conjunto reducido por número de palabras y número de estrellas.

Esto decidió realizarse para añadir un poco de diversidad a nuestro conjunto de interés y para poder hacer un análisis que nos brindara más información con respecto a las palabras contenidas en las reseñas.

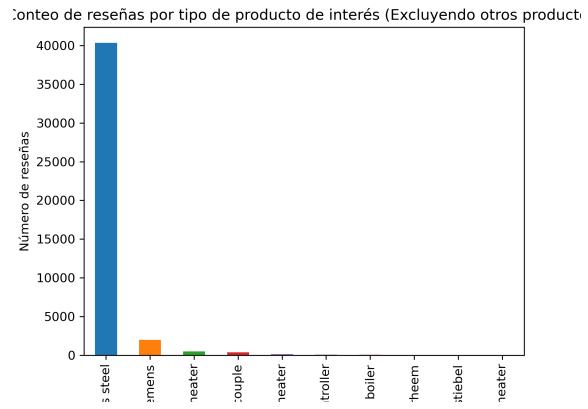
La distribución de las categorías de los productos por el tipo de producto se graficó, quedando de la siguiente manera:



**Figura 9:** Conteo de reseñas en conjunto reducido.

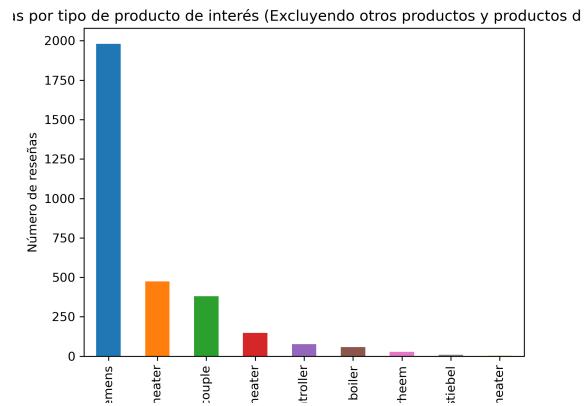
Donde la mitad de nuestro conjunto reducido pertenece a productos de tipo *other*, por lo que al graficar úni-

camente los productos de nuestro interés obtenemos la siguiente gráfica comparativa con respecto al número de reseñas:



**Figura 10:** Conteo de reseñas en conjunto de interés.

Donde se aprecia que los productos relacionados con el acero inoxidable, material de la carcasa de nuestro producto a desarrollar, cuentan con un importante número de reseñas a diferencia del resto de las categorías, siendo muy reducida su participación dentro del conjunto de interés, donde al eliminar a los productos de acero inoxidable del gráfico obtenemos:



**Figura 11:** Conteo de reseñas en conjunto de interés sin tomar en cuenta acero inoxidable.

Siendo este gráfico el que nos da una idea de la participación de las reseñas de cada producto por cada una de

las categorías que nosotros decidimos implementar para el análisis de reseñas.

De esta manera, en resumen, la agrupación del número de reseñas por distintas categorías y áreas de interés como lo son el número de estrellas quedaría de la siguiente manera:

```
Agrupación de productos por número de estrellas:
5      4419731
4      1326702
1      846097
3      611995
2      429919
Name: star_rating, dtype: int64

Agrupación de productos por categoría:
Electronics        2891062
Office Products    2406499
Tools              1574301
Furniture          762582
Name: product_category, dtype: int64

Agrupación de productos por tipo de producto:
other             7590925
stainless steel    40364
siemens           1981
water heater       473
thermocouple       381
tankless water heater 147
temperature controller 77
boiler              57
rheem               27
stiebel              9
electric water heater 3
Name: product_type, dtype: int64

Número de productos únicos: 647010
```

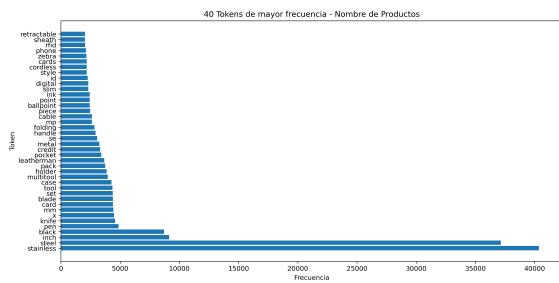
**Figura 12:** Información resumida del conjunto de datos.

## 6.3. Análisis de Frecuencia de Tokens

Como siguiente área de análisis, decidimos reconocer la diferencia que existe entre un análisis de frecuencia de aparición de palabras, un análisis de frecuencia TF-IDF y un análisis de frecuencia TF-IDF a partir de un texto que haya pasado por el proceso de lematizado.

### 6.3.1. Frecuencia de Palabras por Aparición

Para esto, decidimos hacer un conteo del número de apariciones que tiene cada palabra dentro del nombre del producto, con lo cual a continuación se muestran los cuarenta *Tokens* de mayor aparición dentro de este rubro.

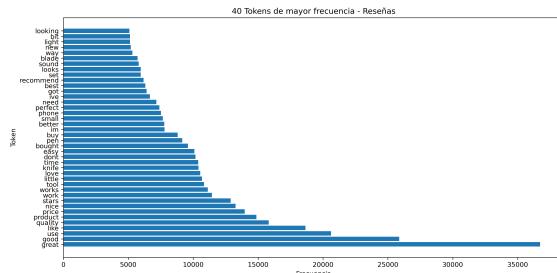


**Figura 13:** Frecuencia por aparición en nombre del producto.

Se puede apreciar que las palabras de mayor frecuencia de aparición se relacionan con las características de los productos de nuestro interés, como lo son las palabras *stainless*, *steel*, *inch*, *pen*, *knife*, *multitool*, donde no se nos hace raro encontrar este tipo de descripciones de productos debido a que, en general, reconocemos que los productos como navajas y cuchillos son los que cuentan con estas características, como el hecho de que están construidas con acero inoxidable. De misma forma, podemos encontrar palabras que describen la manera en cómo se presentan los productos, como lo son las palabras *pack*, *piece*, *set*, *case*, *pocket*.

Por parte de las reseñas, las palabras con mayor frecuencia de aparición se relacionaban con adjetivos que calificaban a un producto, como lo son las palabras *great*, *good*, *like*, *nice*, *love*, *perfect*, así como las características

con las cuales debe cumplir un producto de éxito para la venta como lo son *quality, price, work, works*

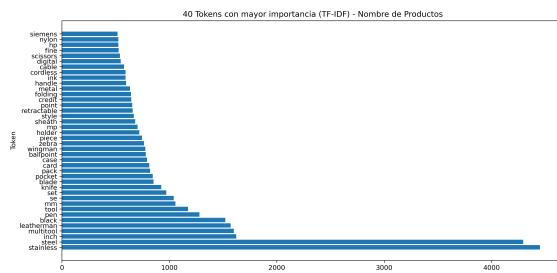


**Figura 14:** Frecuencia por aparición en reseña.

Sin embargo, también contamos con palabras que no tienen tanto significado para nosotros, o palabras que al aparecer podrían resultar más significativas para nuestro análisis, sin embargo, éstas se pueden ver opacadas por la aparición y repetición de otras palabras. Es por esto por lo que se decidió probar la diferencia entre la frecuencia común de aparición de palabras y el uso de la frecuencia TF-IDF.

### 6.3.2. Frecuencia TF-IDF

Por parte del análisis TF-IDF, hicimos uso de la función *TFIdfVectorizer()*, con lo cual para las palabras que se encontraban en el nombre de los productos apreciamos lo siguiente:

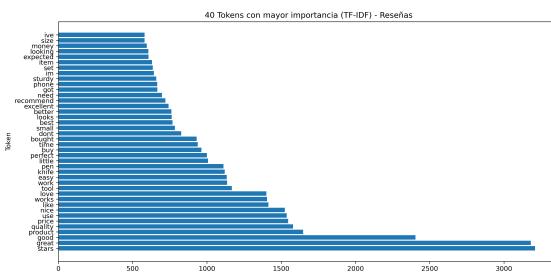


**Figura 15:** Frecuencia TF-IDF en nombre del producto.

Siendo que seguimos teniendo las palabras *stainless*, *steel*, *inch*, *pen*, *knife*, *multitool* dentro de nuestro conjunto, sin embargo, la importancia y orden han cambiado.

do, por ejemplo, la palabra *multitool* escaló varias posiciones de importancia situándose muy cerca de las palabras *stainless* y *steel*. Por otro lado, hay palabras que ya no aparecen en este nuevo gráfico como lo son *x* y *id*, las cuales no nos proporcionaban información importante al análisis, y se ven añadidas las palabras *siemens*, *fine*, que en su lugar nos proporcionan mayor información.

Para el caso de la frecuencia TF-IDF en las reseñas, tenemos el siguiente gráfico.



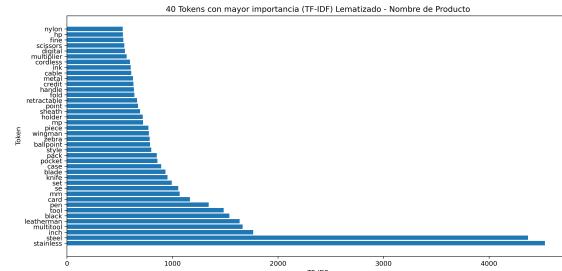
**Figura 16:** Frecuencia TF-IDF en reseña.

Donde podemos apreciar que seguimos contando con los adjetivos que calificaban a un producto *great, good, like, nice, love, perfect*, así como las características con las cuales debe cumplir un producto de éxito para la venta como lo son *quality, price, work, works*, sin embargo, el orden se ve modificado y este conjunto de palabras que nos interesa se encuentra más cercano a las palabras de mayor importancia para el caso de TF-IDF. Por otro lado, se aprecia cómo la palabra *stars*, escaló un gran número de lugares siendo esto de interés debido a que nuestro objetivo final busca que los consumidores puntúen a nuestros productos con un alto número de estrellas. Se aprecia un conjunto similar de palabras, pero cambia el orden.

### 6.3.3. Frecuencia TF-IDF texto Lematizado

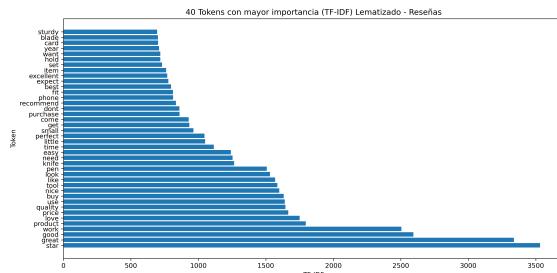
Finalmente, decidimos realizar un tercer análisis a partir de la lematización de nuestro texto y el uso de TF-IDF para el análisis de frecuencias. En este caso, entre las gráficas anteriores y ésta apreciamos el cambio en la palabra *folding* por *fold*, así como un cambio mínimo en

el orden de las palabras, siendo que este cambio se debe al proceso de lematización.



**Figura 17:** Frecuencia TF-IDF con texto lematizado en nombre del producto.

Para el caso de las reseñas, el cambio observado es mayor, pues además de cambiar palabras como lo es de *stars* a *star*, los conjuntos de palabras *work/works*, *buy/bought*, *looks/looking* que antes aparecían han sido fusionadas a *work*, *buy* y *look*, donde su nivel de importancia aumentó, dando paso a que estas palabras similares ocupen un menor espacio y mayor importancia, con lo cual nuevas palabras pueden aparecer en el análisis como lo son las palabras *purchase o year*, así como se eliminan palabras como *money o size*.



**Figura 18:** Frecuencia TF-IDF con texto lematizado en reseña.

Por lo tanto, consideramos que este último tipo de análisis y herramientas son las que más nos pueden ayudar a reconocer lo que los clientes buscan en un producto.

## 6.4. WordClouds

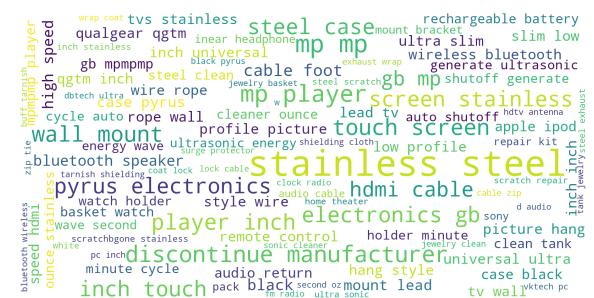
Por otra parte, de manera visual se puede realizar un análisis de palabras a partir de la creación de *WordClouds*. A través del conjunto reducido de datos hemos decidido realizar tres tipos de agrupaciones de *WordClouds* para el análisis, un primer análisis a las cuatro categorías generales y los nombres de los productos, un segundo análisis a los tipos de productos de interés y las reseñas, y un tercer análisis de reseñas por agrupación de estrellas en tres categorías, reseñas altas, medias y bajas.

### 6.4.1. WordClouds por nombre del producto en categorías

Dentro del primer grupo de análisis contamos con las agrupaciones por categorías generales de las reseñas y los nombres de los productos.

En esta primera imagen, se visualiza un *WordCloud* relacionado con los productos de la categoría de Electrónicos y el nombre de los productos.

Se aprecia que las palabras más representativas en los nombres de los productos de la categoría de Electrónicos sería *Stainless Steel*, *discontinue manufacturer*, *electronics gb*, *steel case*, *pyrus electronics*, *player inch*, *touch screen*, *screen stainless*, *mp player*, *wall mount*, entre otros.



**Figura 19:** WordCloud de nombres de productos en electrónicos.

A partir de estas imágenes nos dimos cuenta de que realmente no nos interesa mucho el conocer los tipos de

productos que se encuentran en cada categoría. Es interesante el darse cuenta de la relación que existe entre la categoría y el nombre del producto, sin embargo, para efectos de nuestro análisis esto no nos interesó, por lo tanto, se podrá revisar los otros tres *WordClouds* en la parte del anexo de este documento.

#### **6.4.2. WordClouds por Tipo y Reseña**

Como otra área de análisis tratamos de revisar la relación que existe entre las reseñas y los tipos de productos para que, de esta forma, reconozcamos las palabras más importantes de las reseñas de los usuarios relacionados con los distintos productos de interés.

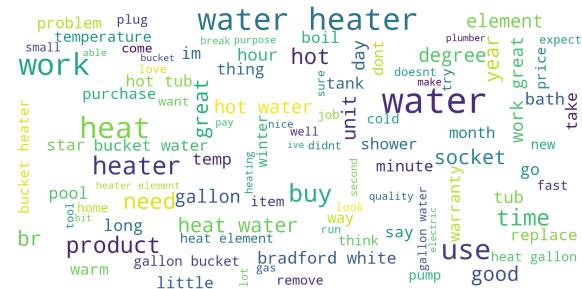
Como primer tipo de producto de interés tenemos a los productos en cuyo nombre se encuentra la palabra *Boiler*.



**Figura 20:** WordCloud de reseñas de boilers.

Entre estas reseñas, las palabras que nos interesan destacar serían *leak*, *replace*, *year*, *hot water*, *fail*, *disappointed*, *issue*, *delicate*, *plumber*, *broken*. Hemos escogido estas palabras debido a que a pesar de que no todas son las más significativas, son las que más nos interesan, pues a partir de ellas podemos reconocer que los usuarios no han tenido buenas experiencias con este tipo de productos, siendo que hay que buscar qué puntos habría que mejorar, como lo es la parte de evitar filtraciones, hacer productos robustos, fáciles de instalar, que no requieran de plomeros.

Otro sinónimo para los *Boilers* serían los calentadores de agua, siendo de alto grado de interés para nosotros.



**Figura 21:** WordCloud de reseñas de calentadores de agua.

Las palabras a destacar para los calentadores de agua serían *water heater, heat, heater, water, hot water, heat water, good, use, time, work great, price, heater element, temperature, greate, love*. Donde, podemos apreciar que los calentadores de agua cuentan con reseñas o palabras más generalizadas y menos sesgadas hacia una opinión positiva o negativa que los *boilers*, por lo cual sería interesante reconocer el motivo, y de esta forma darse cuenta que el nombre del producto y la descripción también resulta importante ante la opinión de los consumidores. Podemos apreciar que, en general, a los clientes les gusta este producto, aunque las reseñas se enfocan más en su uso.

Como siguiente grupo de productos de interés tenemos a los calentadores de agua eléctricos, siendo un producto más específico que los *Boilers* y calentadores de agua, y de mayor interés, pues es el grupo de competencia directa del producto que nos encontramos desarrollando.



**Figura 22:** WordCloud de reseñas de calentadores de agua eléctricos.

Entre estas reseñas, las palabras que nos interesan destacar serían *check valve, air, installation, lifetime, warranty, service, replace, damage, problem*. Hemos escogido estas palabras debido a que nos indican que los usuarios se encuentran consternados por la garantía de los productos, debido a que se puede requerir de servicios o reemplazo de piezas. Es por esto por lo que nuestra empresa deberá dar un gran soporte técnico y hacer uso de piezas de calidad para evitar la necesidad de servicios y reemplazo de piezas. De misma forma, habrá que proponer una instalación sencilla para el usuario, así como considerar el uso de válvulas de alivio en los productos, las cuales para el caso específico de nuestro producto no son necesarias.

Por parte de otro producto a analizar, siendo parte de la familia de los calentadores de agua, pero siendo un producto que nosotros no vamos a desarrollar, tenemos a los calentadores de agua que carecen de tanques de almacenamiento, como lo serían los calentadores de paso.

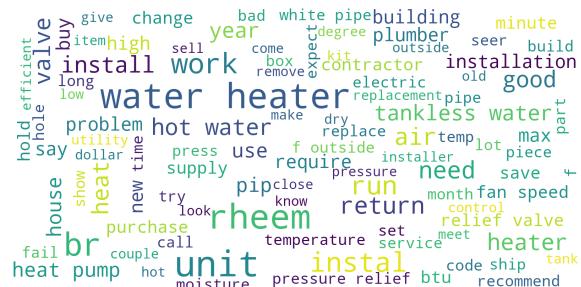


**Figura 23:** WordCloud de reseñas de calentadores de agua sin tanques de almacenamiento.

Las palabras a destacar en los calentadores de agua de paso serían *water heater, hot water, install, water temperature, tankless, problem, replace, warranty, replacement, experience, fail, failure, remove, hard, leak, customer service*. Estas palabras resultan interesantes debido a que representan un área de calentadores de agua que compiten directamente con nuestro producto, siendo calentadores de agua de paso. Estos productos nosotros no los desarrollamos debido a la potencia eléctrica requerida, por lo que la instalación debe ser más robusta, además de que la temperatura a la que se calienta el agua en

un calentador de paso es de máximo 15 grados centígrados, sin embargo, resulta interesante conocer la opinión de los clientes con respecto a este producto, pues al parecer cuentan con muchos fallos y problemas de garantía en este tipo de productos.

Entrando ahora al análisis de reseñas de productos de la competencia a partir de su marca, tenemos la imagen de Rheem.

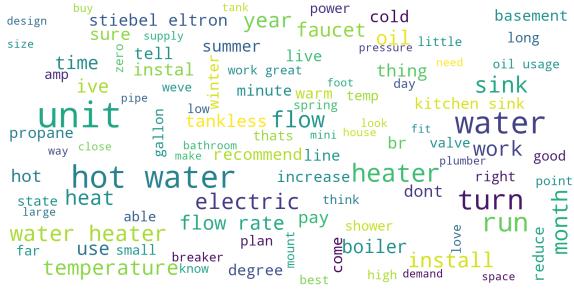


**Figura 24:** WordCloud de reseñas de productos de la marca Rheem.

Para los productos de Rheem, las palabras que nos interesan destacar son *water heater, hot water, rheem, tankless water*, con lo cual observamos que la marca se centra bastante en los productos que vamos a desarrollar, en palabras más pequeñas encontramos *fail, bad, ugly, slow, old, high, price, efficient, fast, replacement*. A partir de estas palabras buscamos darnos cuenta de los conceptos que debemos cuidar en nuestro producto, como lo es el diseño, la calidad, el precio, la eficiencia para calentar el agua, el servicio de instalación y el reemplazo de piezas. Esto nos sirve como base para reconocer en qué áreas hay que prepararse una vez que se haya lanzado el producto, así como las áreas de oportunidad que tenemos como empresa con respecto al diseño y desarrollo del mismo.

Para la marca de calentadores eléctricos Stiebel tenemos la siguiente imagen.

Las palabras que nos interesan destacar en los productos de la marca Stiebel serían *electric, hot water, install, work great, stiebel eltron, water, basement, kitchen*



**Figura 25:** WordCloud de reseñas de productos de la marca Stiebel.

*sink, heater, minute, recommend, plumber.* De estas palabras podemos darnos cuenta que estos calentadores de agua en general son buenos, no cuentan con tanques de almacenamiento y resultan ser muy efectivos debido a que calientan el agua a medida que pasa por lo que son muy recomendados además de que al ser eléctricos y compactos pueden colocarse en cualquier lugar. Este tipo de reseñas deberían relacionarse con los de los calentadores de agua de paso.

Por parte de la marca Siemens, cuyos productos formaran parte de los elementos de los calentadores eléctricos a desarrollar, encontramos lo siguiente.

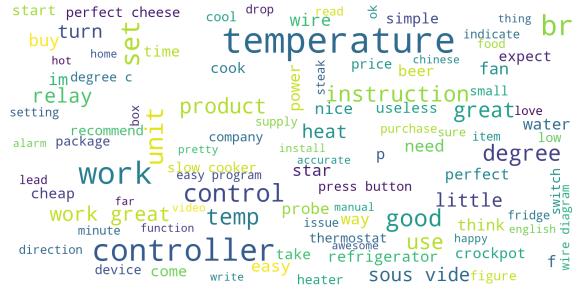


**Figura 26:** WordCloud de reseñas de productos de la marca Siemens.

Las palabras que nos interesan destacar con respecto a Siemens son *sound quality, phone, caller, interference, handset, headset, speakerphone, cell phone, battery life*, con lo cual observamos que los productos que se encuentran en Amazon de esta marca se relacionan más al área de telefonía, por lo tanto el análisis que realicemos de las

reseñas no nos será muy útil.

Así como ya hablamos de Siemens, como marca de algunos de los elementos que utilizamos en los calentadores, ahora hablaremos de los elementos más importantes de los calentadores, comenzando con el controlador de temperatura.



**Figura 27:** WordCloud Temperature Controller Reseña.

Las palabras a destacar en los controladores de temperatura serían *temperature, work great, easy, controller, little, perfect, manual, accurate, fahrenheit, quality, celsius, awesome, easy program*. Estos elementos electrónicos son un tanto complejos, sin embargo, le dan un alto grado de calidad a un producto. Al parecer los clientes que hacen uso de estos elementos agradecen la precisión y calidad del producto, siendo la razón por la cual en nuestra empresa hemos decidido implementarlo para el desarrollo de nuestros calentadores de agua de alta calidad, pues los controladores de temperatura son bastante precisos y permiten a los usuarios conocer con precisión la temperatura del agua que usan. Cuenta con un gran número de reseñas positivas y de ideas positivas.

Por parte de los termopares, estos dispositivos reemplazan a los termostatos dentro de un calentador de agua, siendo dispositivos más precisos.

Las palabras a destacar en los termostatos serían *controller, thermocouple, temperature controller, thermoset, thermometer, temperature, wire, switch*, siendo estos sinónimos de los elementos más representativos en un calentador de agua. Con respecto a la calidad y trabajo encontramos *good, work, work great, easy, perfect, right*,



**Figura 28:** WordCloud Thermocouple Reseña.

*buy, use, great product, price*, con lo cual los usuarios al apreciar se encuentran atraídos hacia los termopares debido al trabajo que realizan y al combinarlos con los controladores de temperatura puede ser que obtengamos un excelente resultado y producto, con elementos de alta calidad a los cuales los consumidores les encanta.

Por parte de la carcasa y exterior de nuestro calentador de agua hacemos uso del acero inoxidable, con lo cual queremos realizar un análisis de los productos que hacen uso de este material para darnos cuenta del punto de vista de los usuarios.



**Figura 29:** WordCloud de reseñas de productos de acero inoxidable.

Para los productos de acero inoxidable, las palabras que nos interesan destacar son *good price, high quality, stainless steel, highly recommend, work great, expensive, beautiful, look great, easy use, good product, durable*, con lo cual nos damos cuenta de que en general los productos de acero inoxidable se relacionan con una calidad excelente, siendo que a pesar de su precio, los clientes se encuentran satisfechos con el producto, siendo es-

to un punto importante debido a que nuestros productos buscan ser de calidad pese a que su precio será elevado, con lo cual, observamos que existe mercado que reconoce estas características en un producto.

Por parte del grupo del conjunto extendido de datos, distinto a nuestro conjunto de interés, obtenemos la siguiente imagen.



**Figura 30:** WordCloud de reseñas de otros productos.

Las palabras que nos interesan destacar son *great product paint, high quality, good quality, look great, sound quality*. Por lo tanto, los usuarios lo que más buscan dentro de un producto sería la calidad del producto tanto físicamente como en su uso, así como buenos acabados. Habrá que enfocarse mucho en un buen diseño de producto con acabados que llamen la atención de los clientes.

#### **6.4.3. WordClouds de reseñas por cantidad de estrellas**

Por otro lado, como última área de interés para el análisis de *WordClouds*, tenemos el análisis de reseñas por cantidad de estrellas, donde hemos dividido este análisis en tres categorías, cuando un usuario puntúa reseñas con un número alto de estrellas, cuando lo hace con un número medio de estrellas y cuando lo hace con un número bajo de estrellas.

Para las reseñas con pocas estrellas las palabras a destacar son *product, problem, time, issue, replace, break, quality, expect, small, sound, bad, price, cheap*. A partir de estas palabras podemos darnos cuenta que los clientes al momento de puntuar de manera baja un produc-



**Figura 31:** WordCloud reseñas con pocas estrellas.

to lo hacen por ser productos de baja calidad, donde el producto no tiene un precio relacionado con lo esperado, baratos, pequeños, con problemas, donde el tiempo lo consideran importante como podría ser el tiempo de entrega o el tiempo de vida del producto.

Para el caso de los productos con puntuaciones de tres estrellas encontramos la siguiente imagen.



**Figura 32:** WordCloud reseñas con tres estrellas.

Donde las palabras a destacar serían *good, nice, look, bad, love, sound, light, break, easily, problem, easy, hard, time, little, small, perfect, work great, cheap*. Las palabras se relacionan con un punto medio entre baja calidad y alta calidad, podemos encontrar tanto palabras positivas como negativas, por lo que el análisis se debería dar con un mayor contexto.

Finalmente, para los productos con muchas estrellas las reseñas cuentan con las siguientes palabras.

Donde las palabras que destacamos son *excellent, great*



**Figura 33:** WordCloud reseñas con muchas estrellas.

*product, good quality, high quality, stainless steel, perfect, great, nice, good, happy, time, purchase, buy, great price, clean, right, work, best, build, design.* Para este caso podemos apreciar que las ideas se relacionan con palabras positivas que tienen que ver con el acabado del producto, la calidad, el precio, su diseño, la sensación de la persona al verlo y usarlo, entre otros, siendo estas cualidades que buscamos para nuestros productos.



## 6.5. Análisis de tópicos

Por otro lado, también decidimos trabajar con el análisis de tópicos de las reseñas, donde para esto hemos decidido realizar dos tipos de análisis, un primer análisis para las reseñas de nuestro conjunto reducido de datos, trabajando con dieciocho tópicos, y un segundo análisis para nuestro conjunto de datos de interés, donde trabajamos diez tópicos. Al ser bastante grande el número de tópicos a analizar, en este reporte únicamente presentaremos los tópicos de mayor interés ya sea tanto para la materia como para el objetivo del proyecto que sería el diseño y la venta de nuestros productos, por lo que el resto de tópicos se podrán apreciar en el anexo de este documento.

### 6.5.1. Análisis de tópicos en conjunto reducido de datos

Lo primero que se logra apreciar en nuestro gráfico para el análisis de tópicos de nuestro conjunto reducido de datos son los *Tokens* ó términos de mayor relevancia en nuestras reseñas.

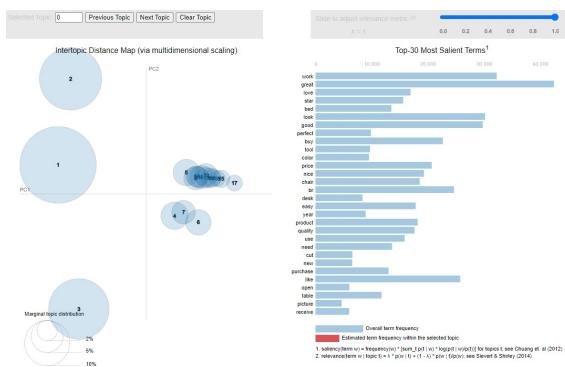


Figura 34: Términos más relevantes en conjunto reducido de datos.

En esta imagen podemos apreciar cómo los términos más relevantes se relacionan con adjetivos calificativos hacia los productos, como lo serían las palabras *great, love, good, perfect, easy*, así como otras palabras relacionadas a características o artículos de los que se hicieron la reseña como lo son las palabras *bed, look, tool, color, price, chair, quality, picture*.

Por parte del análisis de tópicos, el primer tópico que nos pareció más interesante fue el segundo tópico.

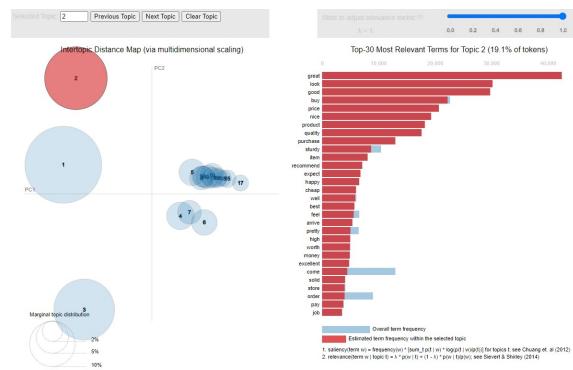


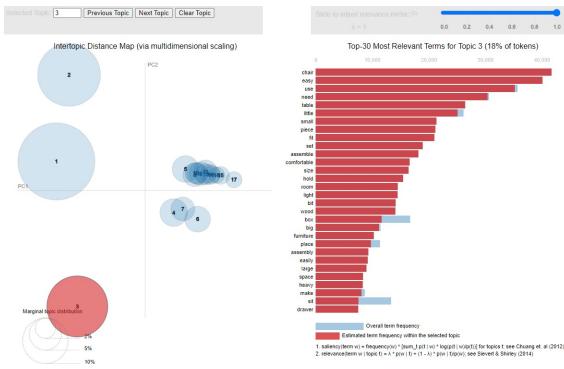
Figura 35: Tokens para segundo tópico en conjunto reducido de datos.

Este tópico contaba con una distribución del 19.1% de los *tokens* totales, donde las palabras que se presentan son adjetivos o características de interés para las personas que se encuentran comprando un producto, como lo son *great, look, good, price, nice, product, quality, purchase, happy, cheap, best, feel, high, worth, excellent, solid*, entre otros. Estas palabras nos resultan interesantes debido a que describen lo que se espera de nuestro producto, es decir, que se vea bien, que su relación calidad-precio sea buena, que sea sólido, que valga la pena.

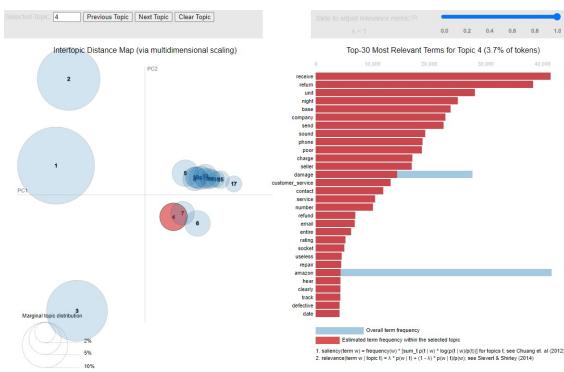
Otro tópico de interés para el caso de la asignatura, en el cual se puede demostrar una distribución de términos por relación en su significado sería el tercer tópico.

Esto se debe a que en este tercer tópico encontramos términos relacionados con muebles y recámaras o cuartos, como lo son los términos *chair, table, assemble, comfortable, room, light, wood, box, big, space, heavy*. Decidimos destacar este tópico debido a que demuestra de forma práctica la teoría vista a lo largo de la clase y cómo se forman los tópicos a partir de la relación de las palabras y su aparición.

El cuarto tópico de este conjunto reducido de datos se relacionaba con el envío y la relación que existe entre el vendedor, el consumidor y Amazon como intermediario.



**Figura 36:** Tokens para tercer tópico en conjunto reducido de datos.

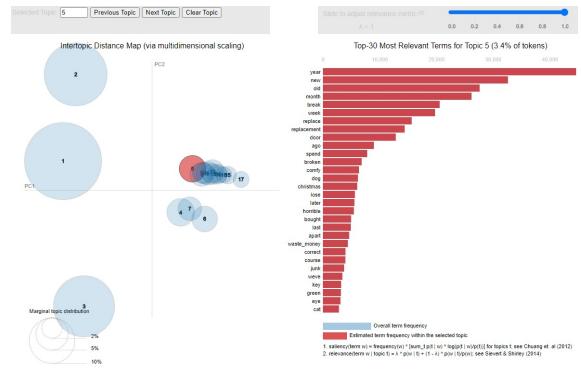


**Figura 37:** Tokens para cuarto tópico en conjunto reducido de datos.

En este tópico se aprecian palabras como *receive*, *return*, *unit*, *company*, *send*, *phone*, *poor*, *charge*, *seller*, *damage*, *customer\_service*, *contact*, *service*, *refund*, *number*, *email*, *rating*, *repair*, *amazon*, *track*. Estas palabras las relacionamos con el proceso de compra, venta y envío de un producto, siendo interesante cómo se agrupan en un tópico.

Como último tópico a analizar en esta sección tenemos al quinto tópico que representa el 3.4% de los *tokens* de nuestro conjunto reducido de reseñas.

En esta agrupación encontramos palabras relacionadas con el tiempo, como lo son *year*, *new*, *old*, *month*, *week*, *ago*, *later*, *last*. Estas palabras también nos ayudan a darnos cuenta del funcionamiento de la teoría vista en



**Figura 38:** Tokens para quinto tópico en conjunto reducido de datos.

clase.

### 6.5.2. Análisis de tópicos en conjunto de datos de interés

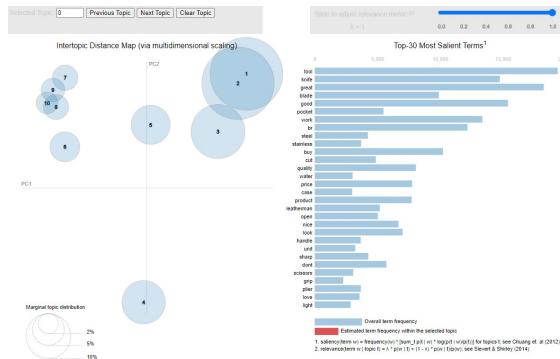
A partir del análisis anterior, consideramos que íbamos a tener mejores resultados al trabajar un análisis de tópicos únicamente a partir de nuestro conjunto de datos de interés. Hay que tomar en consideración que la distribución de los datos era bastante poco uniforme, sin embargo, decidimos trabajar así de todas maneras.

Hay que recordar que para este caso se trabajaron diez tópicos, donde únicamente vamos a reportar los tópicos más relevantes, mientras que el resto podrá apreciarse en el anexo de este documento.

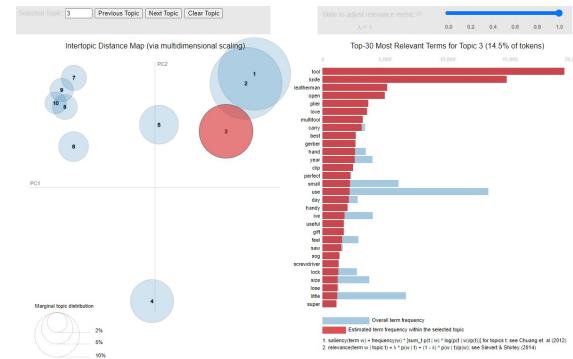
Dentro de los términos más relevantes de nuestro conjunto de datos de interés se aprecian palabras relacionadas con el conjunto de datos más significativo, siendo este el de *stainless steel*, con palabras como *tool*, *knife*, *great*, *blade*, *steel*, *stainless*, *cut*, *quality*, entre otras.

Como primer tópico de interés tenemos al primer tópico del conjunto de datos de interés que se parece bastante al segundo tópico del conjunto reducido de datos.

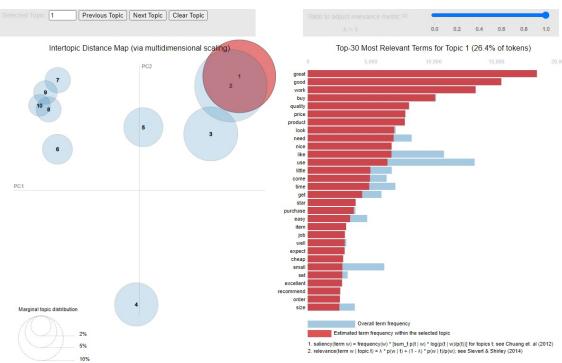
Aquí apreciamos las palabras *great*, *good*, *work*, *buy*, *quality*, *price*, *product*, *look*, *need*, *nice*, *like*, *use*, *little*,



**Figura 39:** Términos más relevantes en conjunto de datos de interés.



**Figura 41:** Tokens para tercer tópico en conjunto de datos de interés.



**Figura 40:** Tokens para primer tópico en conjunto de datos de interés.

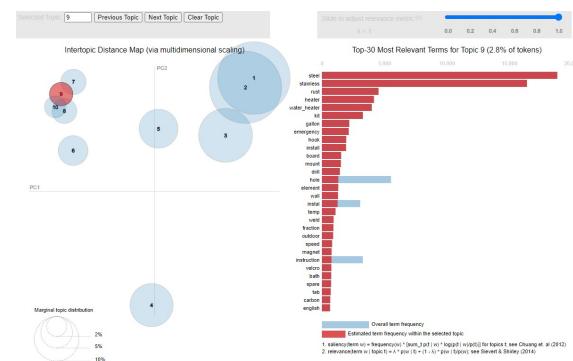
*easy, cheap, small, excellent*, entre otras, las cuales representan adjetivos y características que describen a los productos. Este tópico contiene el 26.4% de los *tokens* de nuestro conjunto de interés.

De misma manera, el tercer tópico representa información relacionada con productos como cuchillos y navajas, siendo que este conjunto de productos era representativo en nuestro conjunto de datos de interés debido a su fuerte relación con el acero inoxidable dentro de la categoría de herramientas. Este tópico contiene el 14.5% de los *tokens* de nuestro conjunto.

Se pueden apreciar palabras como *tool, knife, leatherman, multitool, carry, gerber, hand, clip, use, handy, gift, screwdriver, lock*, siendo éstas palabras relacionadas con

navajas, cuchillos y multiherramientas, con lo cual nuevamente demostramos el funcionamiento del análisis por tópicos a partir de la relación existente en la aparición de palabras.

Finalmente, como último tópico de interés tenemos al noveno tópico de nuestro conjunto de datos de interés donde participan el 2.8% de los *tokens*.



**Figura 42:** Tokens para noveno tópico en conjunto de datos de interés.

Dentro de este tópico podemos apreciar palabras como *stainless, steel, rust, heater, water\_heater, gallon, install, mount, drill, hole, element, wall, temp, weld, bath*, con las cuales se tiene completa relación con los productos que desarrolla nuestra empresa, siendo los calentadores de agua eléctricos de alta calidad.



Este análisis de tópicos resulta sumamente útil debido a que nos permite apreciar de forma acotada ideas relevantes en nuestro conjunto de datos, como lo son las características que los clientes evalúan al comprar un producto, el proceso de compra y envío de productos, o inclusive elementos o ideas que conforman a productos como navajas o calentadores de agua.

## 6.6. Análisis de sentimientos con DistilBERT

En el paso de modelado, utilizamos un modelo basado en BERT, una arquitectura de redes neuronales pre-entrenada, para realizar el análisis de opiniones sobre las reseñas de productos y clasificarlas según el número de estrellas. BERT, que significa *Bidirectional Encoder Representations from Transformers*, es un modelo de lenguaje que ha demostrado excelentes resultados en tareas de comprensión del lenguaje natural.

Para nuestro análisis de sentimientos, adaptamos hicimos uso de un modelo de BERT más pequeño, ligero y barato en cuanto a recursos de cómputo, el cual cuenta con 40 % menos parámetros de BERT, corre un 60 % más rápido y preserva un 95 % de su precisión. Face (2020a)

Decidimos hacer uso de *DistilBERT for Uncased Sentiment*, el cual es un modelo afinado para el análisis de sentimientos en reseñas de productos en seis idiomas, que predice el sentimiento de la reseña como un número de estrellas que va entre una y cinco estrellas. Este modelo está diseñado para ser utilizado directamente como un modelo de análisis de sentimientos para reseñas de productos, o para un posterior afinamiento en tareas relacionadas de análisis de sentimientos. Face (2020b)

En nuestro caso decidimos no realizar ningún tipo de entrenamiento o afinación al modelo, así como buscamos predecir los sentimientos del conjunto reducido de reseñas para obtener la clasificación de las reseñas a partir de un número de estrellas que fuera de una a cinco. Esto tenía como objetivo encontrar un modelo que nos permitiera realizar análisis de sentimientos a reseñas a pesar de que éstas se obtengan de otras plataformas que no cuenten con un clasificador por número de estrellas.

Para poder realizar el análisis de sentimientos tuvimos que trabajar con la plataforma de Google Colab y activar el entorno con GPU, donde al pasar las 83,519 reseñas de nuestro conjunto reducido de datos tras tres horas con cincuenta y tres minutos obtuvimos la clasificación de las reseñas por número de estrellas.



```

Archivos
Código → Texto
Carga de modelo DistilBERT
+ Tokenizer + Autoregressor para pre-entrenado ("clipbert-base-multilingual-miscellaneous") 
model = AutoregressorSequenceClassification.from_pretrained("clipbert-base-multilingual-miscellaneous")
model.train()
model.save_pretrained("./distilbert")

Downloading: tokenizer_config.json (100%)
29/370 [00:00<00:00, 1.64MB/s]
Downloading: [model_name].config (100%)
803/930 [00:00<00:00, 33.55MB/s]
Downloading: [model_name].vocab (100%)
873/873 [00:00<00:00, 0.18MB/s]
Downloading: [model_name].merges (100%)
113/113 [00:00<00:00, 1.64MB/s]
Downloading: [model_name].state_dict (100%)
446/446 [00:00<00:00, 20.64MB/s]

[Evaluación de reseñas con distilBERT]
[distilbert] sentiment_score(review):
    tokens = tokenize_and_truncate_review(return_sentence=True, max_length = 512, truncation=True)
    return torch.argmax(result.logits)

[distilbert] sentiment = sentiment([review_text], progress_apply=lambda x: sentiment_score(x))
100% [██████████] 00:00:00.000/00:00.5595s

```

**Figura 43:** Tiempo de ejecución del análisis de sentimientos con DistilBERT.

El reporte de clasificación obtenido se aprecia en la siguiente imagen.

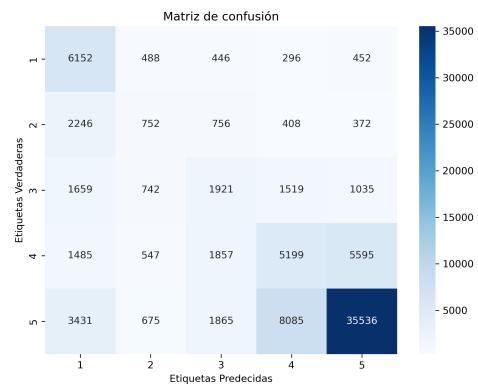
	precision	recall	f1-score	support
1	0.00	0.00	0.00	0
2	0.41	0.79	0.54	7834
3	0.23	0.17	0.19	4534
4	0.28	0.28	0.28	6876
5	0.34	0.35	0.34	14683
micro avg	0.59	0.59	0.59	83519
macro avg	0.35	0.38	0.35	83519
weighted avg	0.62	0.59	0.60	83519

**Figura 44:** Reporte de Clasificación DistilBERT.

Donde se observa que se obtuvo un promedio de precisión del 62 %, mientras que en la página web del modelo nos indica que se llegó a obtener una precisión de hasta el 67 % con reseñas en este mismo idioma, que es el inglés.

Por otro lado, a partir de una matriz de confusión o clasificación resulta más sencillo el análisis de la precisión del modelo, donde habrá que tomar en cuenta que se tenía un mucho mayor número de reseñas de cinco estrellas que del resto de estrellas.

Dentro de la matriz de confusión se aprecia que se obtuvieron 49,560 reseñas clasificadas de manera correcta y 33,959 reseñas clasificadas de manera incorrecta, con lo cual se tuvo una precisión total del 68.52 %, lo cual consideramos bastante alto.



**Figura 45:** Matriz de confusión DistilBERT.

Si ampliamos el margen de error buscando que el modelo pueda equivocarse por una estrella hacia arriba o hacia abajo, aumentamos en 21,287 reseñas el número de reseñas que entran dentro de una categoría aceptable, con lo cual contaríamos con 70,847 reseñas clasificadas de forma aceptable, con un margen de error de una estrella, y 12,672 reseñas clasificadas de forma inaceptable, teniendo 82.11 % de precisión a partir de nuestro margen de error de una estrella, pues consideramos que este margen de error sigue siendo funcional para los fines de nuestro uso.

Por otra parte, se puede apreciar cómo el modelo dió a 22,592 reseñas un valor de estrellas menor que el real, mientras que a 11,367 reseñas les dió un mayor puntaje de estrellas, con lo cual el modelo utilizado es dos veces más capaz de reducir el número de estrellas que de aumentarlo. Esto nos aparece excelente debido a que en la empresa preferimos considerar que el cliente piensa que nuestro producto tiene más áreas de oportunidad para que, de esta manera, podamos enfocarnos en mejorar lo más que podamos.

Por otro lado, también obtuvimos una tabla a partir de los resultados de la clasificación por cada tipo de producto de nuestro interés.

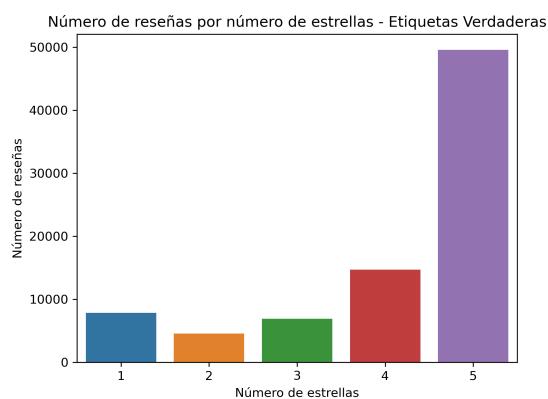
El modelo presentó una raíz de error cuadrático medio de 1.2605 y se tuvo una correlación del 63.78 % entre

	product_type	r2	MSE	rmse
0	boiler	0.512286	1.368421	1.169795
1	electric water heater	-0.038462	3.000000	1.732051
2	other	0.108828	1.606675	1.267547
3	rheem	-0.425000	3.518519	1.875771
4	siemens	0.354099	1.753660	1.324258
5	stainless steel	-0.026936	1.552720	1.246082
6	stiebel	-5.000000	2.666667	1.632993
7	tankless water heater	0.471239	1.653061	1.285714
8	temperature controller	0.175431	1.597403	1.263884
9	thermocouple	0.044930	1.784777	1.335955
10	water heater	-0.038957	2.200846	1.483525

**Figura 46:** Resultados de clasificación por tipo de producto de interés.

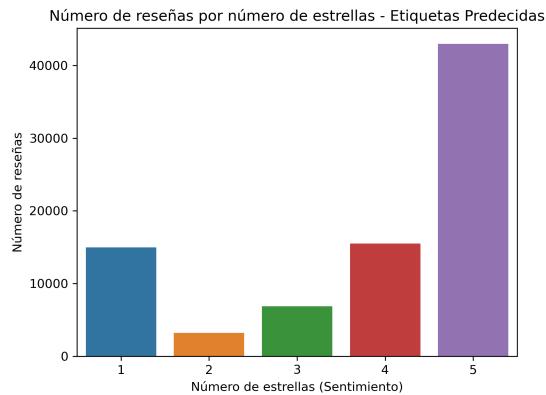
las estrellas dadas por el usuario y el sentimiento obtenido con nuestro modelo.

De forma gráfica la clasificación real de las reseñas en nuestro conjunto reducido era el siguiente.



**Figura 47:** Clasificación real de las reseñas en conjunto reducido.

Mientras que la clasificación obtenida en la predicción mediante el modelo DistilBERT fue el siguiente.



**Figura 48:** Predicción de clasificación de las reseñas en conjunto reducido.

## 7. Conclusión

A lo largo de este proyecto final estuvimos trabajando con el análisis de reseñas de productos en Amazon, así como el análisis de sentimientos de estos mismos productos.

Consideramos que los objetivos de este proyecto se cumplieron en su totalidad debido a que logramos realizar correctamente un análisis de sentimientos de las reseñas de productos de Amazon logrando así comprender lo que buscan los clientes en productos en específico, siendo que fuimos capaces de identificar las características de los productos que las personas resaltan en las reseñas positivas y negativas, así como utilizar esta información para beneficio del desarrollo de nuestro propio producto.

De mismo modo, fuimos capaces de obtener una herramienta que fuera capaz de categorizar reseñas de clientes dándoles como valor final un número de estrellas relacionado con el comentario o reseña que hagan de un producto, donde esta herramienta fue el modelo “DistilBERT for Uncased Sentiment”, el cual nos permite obtener resultados satisfactorios para nuestras necesidades.

Como objetivo particular, deseábamos identificar patrones y tendencias en las preferencias y necesidades de los usuarios relacionados a productos específicos, siendo



estos productos los que nuestra empresa de calentadores eléctricos busca desarrollar. Gracias a este proyecto fuimos capaces de darnos cuenta del cómo los clientes tienen una buena valoración de elementos de los cuales haremos uso en nuestros calentadores de agua eléctricos como son los controladores digitales de temperatura, los termopares y una carcasa de acero inoxidable de alta calidad.

Como trabajo futuro a partir de este proyecto consideramos una buena idea el hecho de trabajar con reseñas más actuales, así como obtener conjuntos de reseñas más grandes relacionados a nuestros productos de interés. Por otro lado, podría considerarse el adaptar o afinar el modelo *DistilBERT for Uncased Sentiment* para que funcione de mejor manera obteniendo precisiones mayores al momento de realizar la clasificación por número de estrellas, además de intentar trabajar con un conjunto de datos de productos en español, pues este será el idioma predominante para el nicho que buscamos abarcar como clientes.

Finalmente, consideramos una buena idea el desarrollar una aplicación o plataforma de análisis de reseñas diseñada directamente para nuestros productos, puesto que a medida que se vayan recibiendo reseñas o comentarios de los clientes se podrá mejorar los productos o la respuesta hacia éstos. Por otro lado, esta misma idea podría desarrollarse para otras empresas y productos de sus áreas, siendo una área en la cual podamos obtener ganancias a partir del proyecto desarrollado.

Consideramos que este fue un proyecto muy completo que abarcó una enorme cantidad de elementos vistos en la materia, por lo tanto, nos encontramos muy satisfechos por los resultados obtenidos.

## 8. Bibliografía

### Referencias

- Blei, D., N. A. . J. M. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022.
- Brownlee, J. (2018). A gentle introduction to learning rate. *Machine Learning Mastery*.
- Delvin, J., C. M. L. K. . T. K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Face, H. (2020a). Distilbert - [https://huggingface.co/docs/transformers/model\\_doc/distilbert](https://huggingface.co/docs/transformers/model_doc/distilbert).
- Face, H. (2020b). Distilbert uncased sentiment - <https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment>.
- HalilErgul (2022). Sentiment and topic modeling on social s. books. - kaggle - <https://www.kaggle.com/code/halilergul/sentiment-and-topic-modeling-on-social-s-books/notebook>.
- Horev, R. (2021). Todo sobre bert, el modelo lingüístico más avanzado para nlp a día de hoy - <https://www.ibidemgroup.com/edu/bert-nlp-machine-translation/>.
- Inaniya, Y. (2021). Sentiment analysis using bert - amazon review sentiment analysis. - <https://www.analyticsvidhya.com/blog/2021/06/amazon-product-review-sentiment-analysis-using-bert/>.
- Kingma, D., . B. J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Sanh, V., D. L. C. J. . W. T. (2019). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.



Services, A. W. (s/f). ¿qué es el análisis de opiniones?

- explicación del análisis de opiniones - aws. (n.d.).
- amazon aws - <https://aws.amazon.com/es/what-is/sentiment-analysis/>.

Sun, C., Q. X. X. Y. . H. X. (2020). How to fine-tune bert for text classification? retrieved mayo 23, 2023, from <https://arxiv.org/pdf/1905.05583.pdf> %22https://arxiv.org/pdf/1905.05583.pdf%ef%bc%89%22.

Zhao, H., L. Z. Y. X. . Y. Q. (2021). A machine learning-based sentiment analysis of online product reviews with a novel term weighting and feature selection approach. sciencedirect. - [https://www.sciencedirect.com/science/article/pii/s0306457321001448?casa\\_token=cdp0dsrnf1qaaaaa:xmzcuhxy6bb0erozludcle8kp17lzjrtbjgndu2mhokeylksza8ck3xvice5ozwjq\\_jpc95m](https://www.sciencedirect.com/science/article/pii/s0306457321001448?casa_token=cdp0dsrnf1qaaaaa:xmzcuhxy6bb0erozludcle8kp17lzjrtbjgndu2mhokeylksza8ck3xvice5ozwjq_jpc95m).

## 9. Anexo

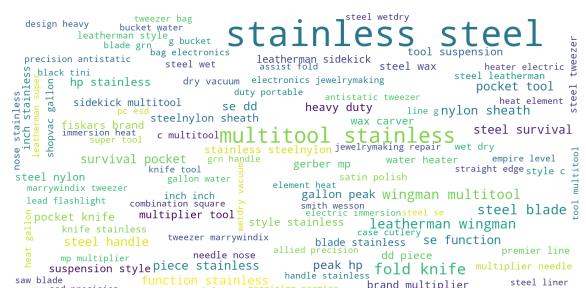
### 9.1. WordClouds por nombre del producto en categorías



**Figura 49:** WordCloud de nombres de productos en muebles.

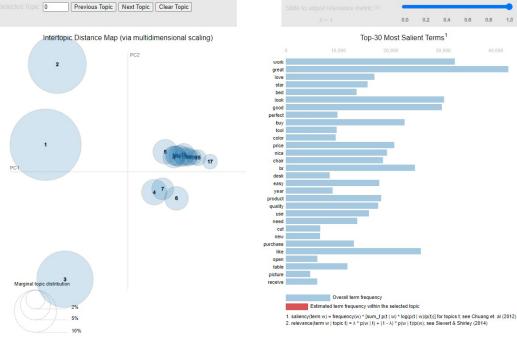


**Figura 50:** WordCloud de nombres de productos en productos de oficina.

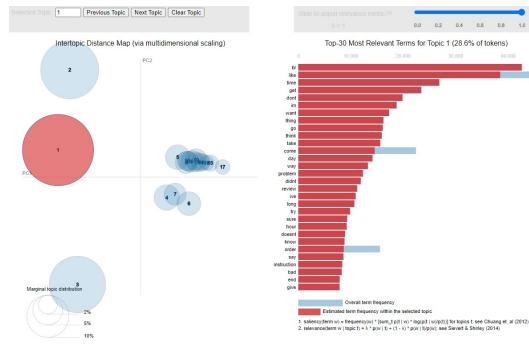


**Figura 51:** WordCloud de nombres de productos en herramientas.

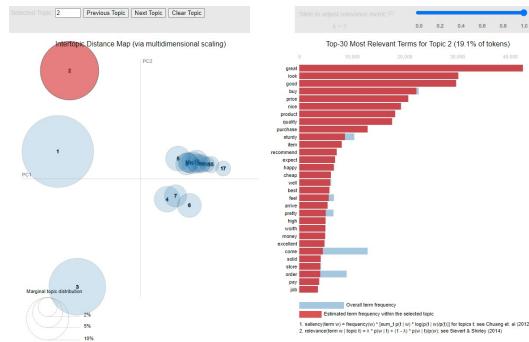
## 9.2. Tópicos del conjunto reducido de datos.



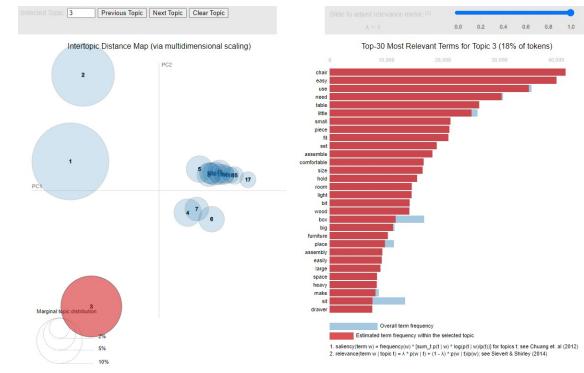
**Figura 52:** Tokens más relevantes para conjunto reducido de datos.



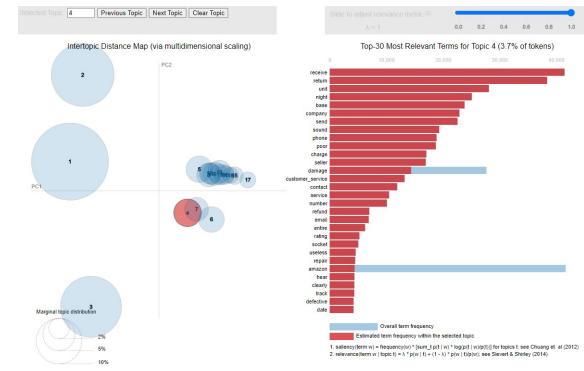
**Figura 53:** Tokens para primer tópico en conjunto reducido de datos.



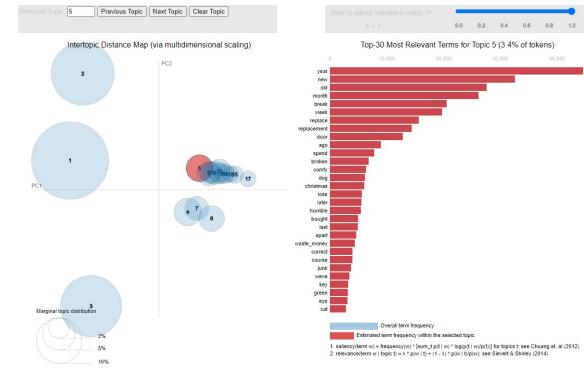
**Figura 54:** Tokens para segundo tópico en conjunto reducido de datos.



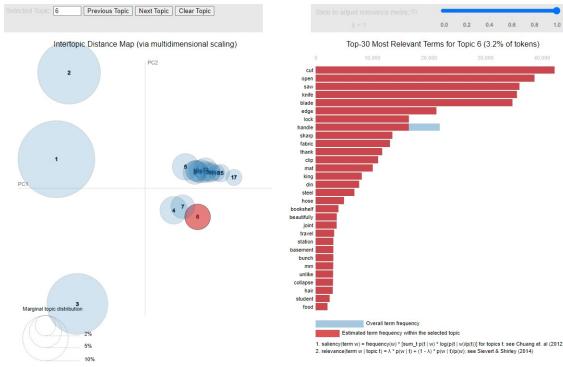
**Figura 55:** Tokens para tercer tópico en conjunto reducido de datos.



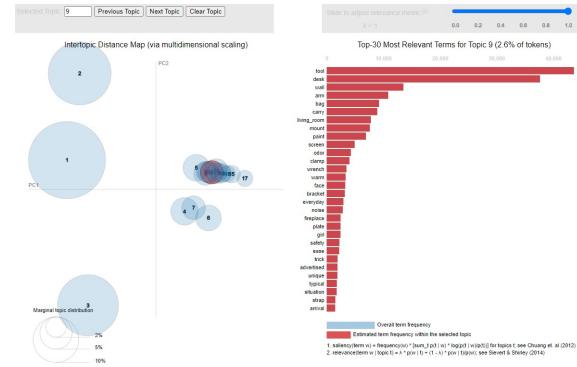
**Figura 56:** Tokens para cuarto tópico en conjunto reducido de datos.



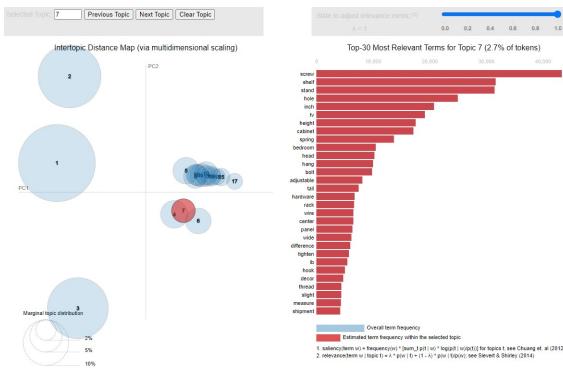
**Figura 57:** Tokens para quinto tópico en conjunto reducido de datos.



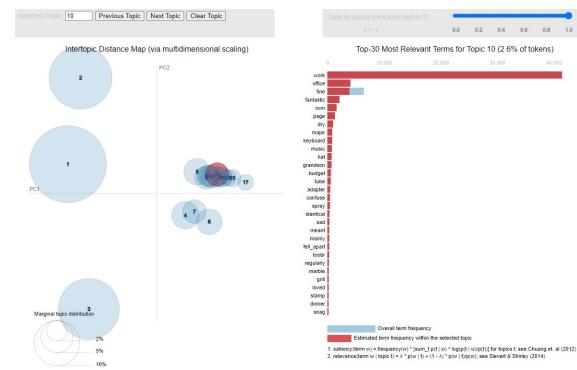
**Figura 58:** Tokens para sexto tópico en conjunto reducido de datos.



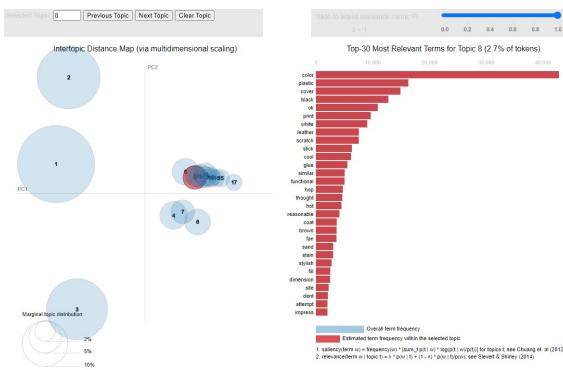
**Figura 61:** Tokens para noveno tópico en conjunto reducido de datos.



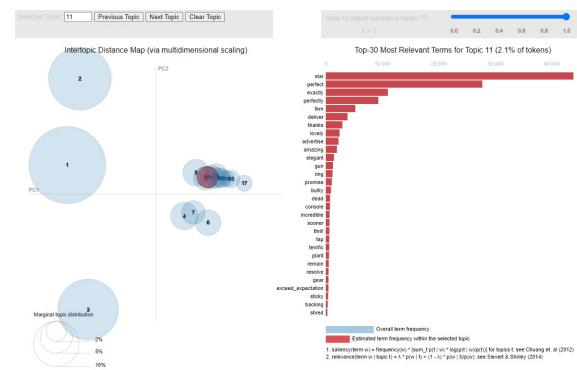
**Figura 59:** Tokens para séptimo tópico en conjunto reducido de datos.



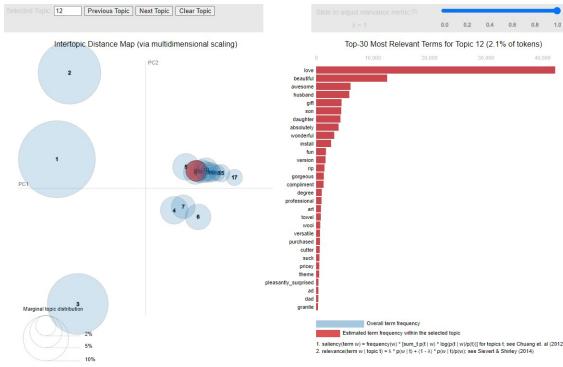
**Figura 62:** Tokens para décimo tópico en conjunto reducido de datos.



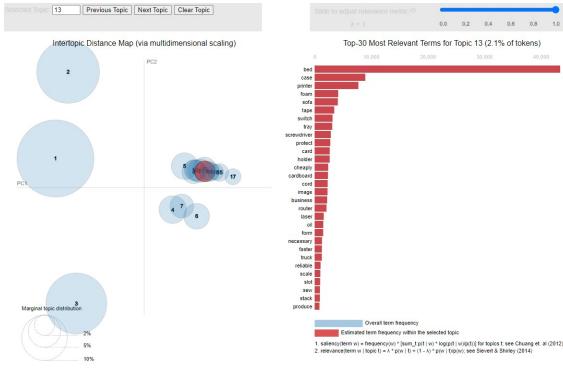
**Figura 60:** Tokens para octavo tópico en conjunto reducido de datos.



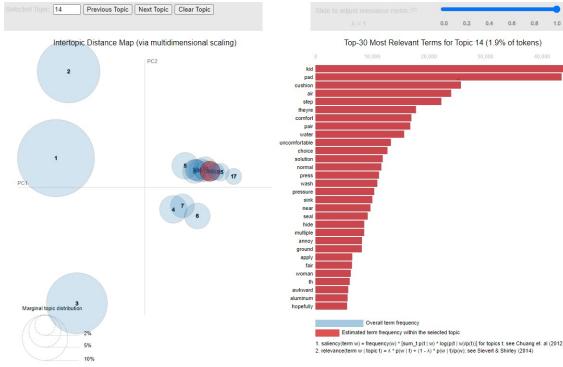
**Figura 63:** Tokens para onceavo tópico en conjunto reducido de datos.



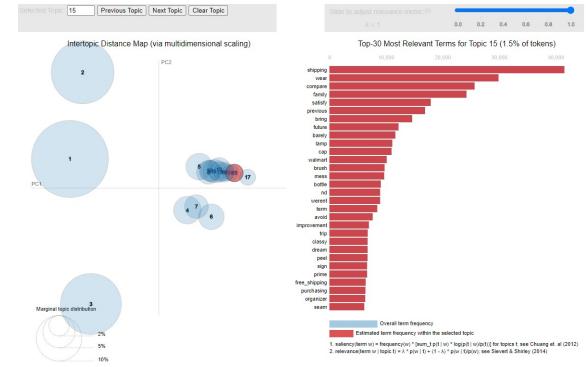
**Figura 64:** Tokens para doceavo tópico en conjunto reducido de datos.



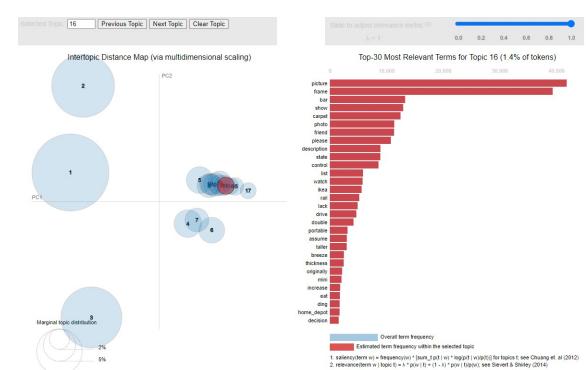
**Figura 65:** Tokens para treceavo tópico en conjunto reducido de datos.



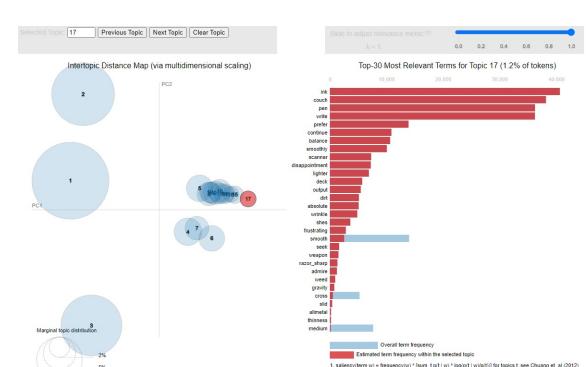
**Figura 66:** Tokens para catorceavo tópico en conjunto reducido de datos.



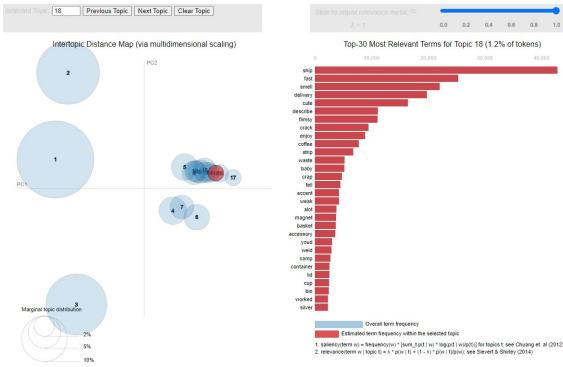
**Figura 67:** Tokens para quinceavo tópico en conjunto reducido de datos.



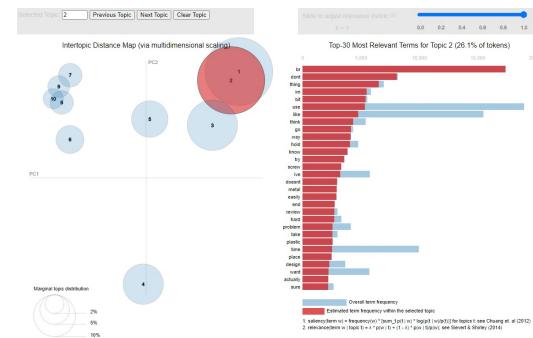
**Figura 68:** Tokens para dieciseisavo tópico en conjunto reducido de datos.



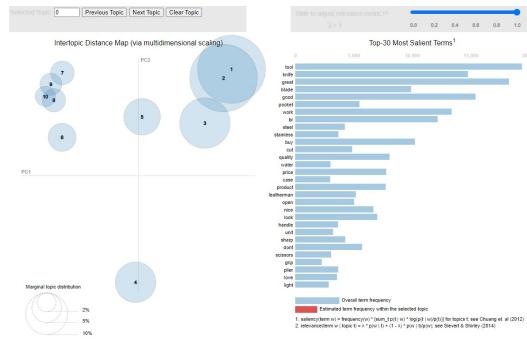
**Figura 69:** Tokens para diecisieteavo tópico en conjunto reducido de datos.



**Figura 70:** Tokens para dieciochoavo tópico en conjunto reducido de datos.

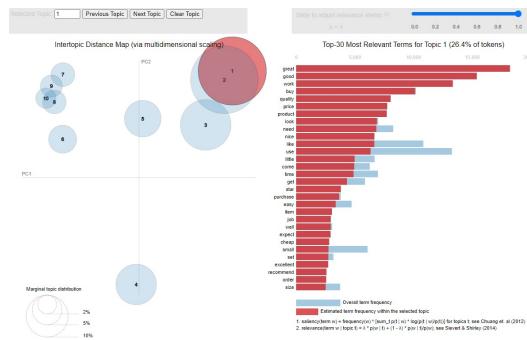


**Figura 73:** Tokens para segundo tópico en conjunto de datos de interés.

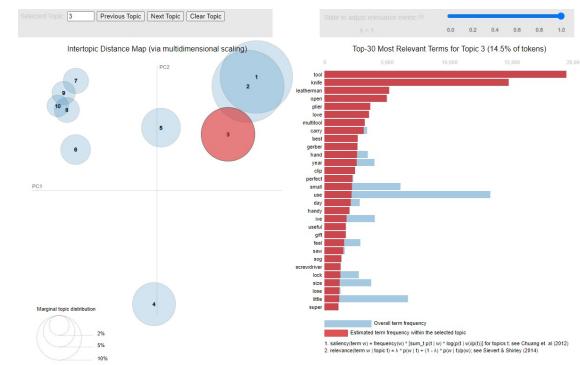


**Figura 71:** Términos más relevantes en conjunto de datos de interés.

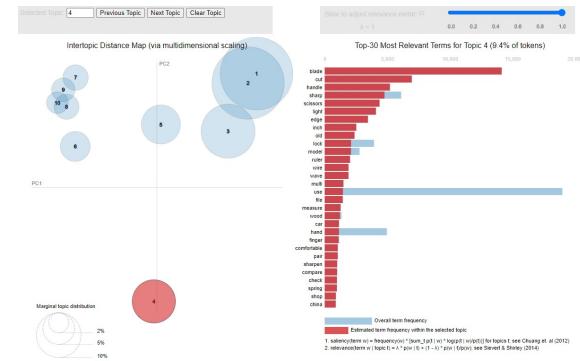
### 9.3. Tópicos del conjunto de datos de interés.



**Figura 72:** Tokens para primer tópico en conjunto de datos de interés.



**Figura 74:** Tokens para tercer tópico en conjunto de datos de interés.



**Figura 75:** Tokens para cuarto tópico en conjunto de datos de interés.

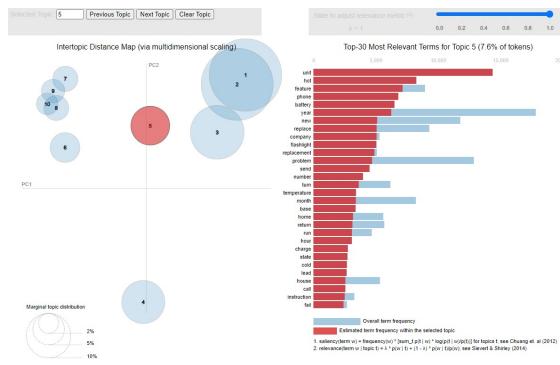


Figura 6: Tokens para quinto tópico en conjunto de datos de interés.

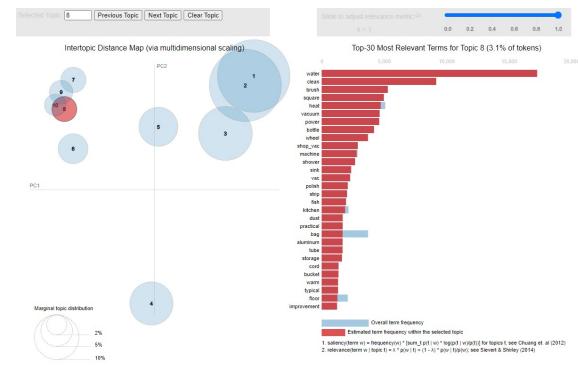


Figura 79: Tokens para octavo tópico en conjunto de datos de interés.

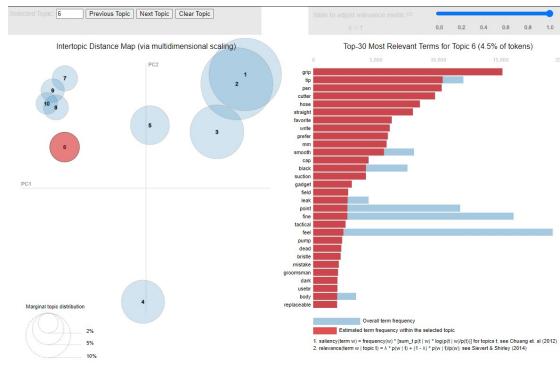


Figura 77: Tokens para sexto tópico en conjunto de datos de interés.

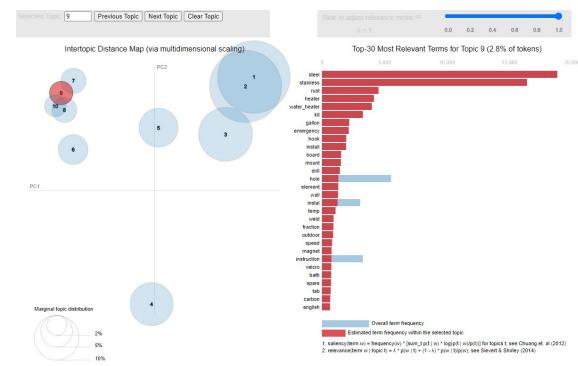


Figura 80: Tokens para noveno tópico en conjunto de datos de interés.

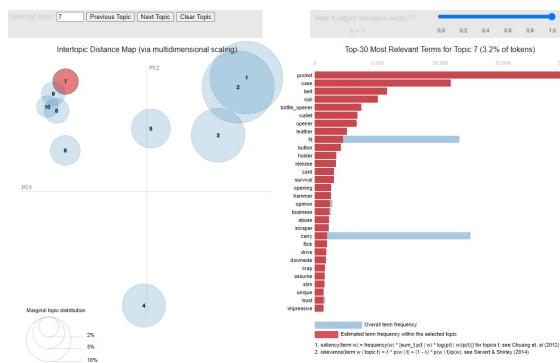


Figura 78: Tokens para séptimo tópico en conjunto de datos de interés.

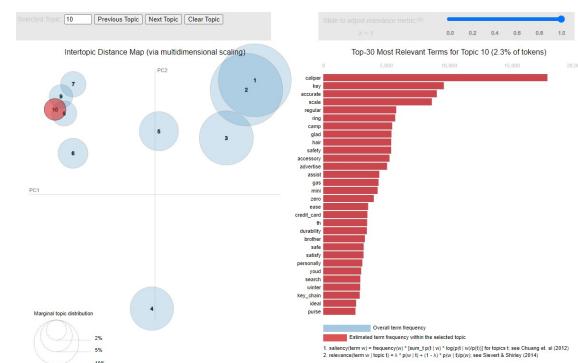


Figura 81: Tokens para décimo tópico en conjunto de datos de interés.