



Sistema de evaluación de características técnicas de los rastreadores de actividad física mediante el análisis de las reseñas de los clientes de Amazon

Gabriel Mantilla Saltos

Tutor(es):

Profesor Monica K. Villavicencio

Profesor Eduardo Cruz

Una tesis submitida para el grado de:
Magister en Ciencias de la Computación

Dedicatoria

Este trabajo está dedicado a mis queridos padres quienes me impulsaron emotivamente a seguir adelante con amor y su infinita sabiduría, a Dios por haberme permitido llegar hasta aquí, con salud y bondad para lograr este objetivo.

Declaracion Expresa

“Los derechos de titularidad y explotación, nos corresponde conforme al reglamento de propiedad intelectual de la institución: Gabriel Fernando Mantilla Saltos doy mi consentimiento para que la ESPOL realice la comunicación pública de la obra por cualquier medio con el fin de promover la consulta, difusión y uso público de la producción intelectual”

Gabriel Fernando Mantilla Saltos

Comité Evaluador

Mónica K. Villavicencio Cabezas
Profesor Tutor

Federico X. Dominguez Bonini
Profesor Evaluador

Eduardo S. Cruz Ramirez
Profesor Tutor

Reconocimiento

Mi más sincero agradecimiento a la Doctora Mónica Villavicencio, al Doctor Eduardo Cruz, y a la Doctora Parisa Eslambolchilar por brindarme sus conocimientos, correcciones e infinita paciencia en culminar con éxito este trabajo de investigación.

Resumen

Los rastreadores de actividad física alientan a las personas a ser más activas, a las que sufren de problemas de obesidad a perder peso monitoreando su dieta, a los adultos mayores a comprender su estado de salud midiendo su frecuencia cardiaca. En general, las empresas muestran publicidad para este tipo de productos y los describen como beneficiosos. Sin embargo, los usuarios buscan los productos que mejor se adapten a sus necesidades personales, para lo cual suelen revisar las opiniones de otros usuarios en plataformas de comercio electrónico como Amazon.

En esta investigación, se estudió la satisfacción de los usuarios tras publicar su experiencia de haber usado los rastreadores de actividad física. Se evaluó si las opiniones provistas están relacionadas a la calidad ofrecida por los fabricantes, mediante la utilización de técnicas de procesamiento de lenguaje natural. Se desarrollo un análisis de sentimientos en función a las características técnicas ofrecidas por los dispositivos, empleando el aprendizaje de transferencia de un modelo de lenguaje basado en Transformer (RoBERTa). Se reentrenó el modelo en un corpus más específico, adaptándolo a la terminología de las reseñas en base a las valoraciones de los usuarios.

Se desarrollo el entrenamiento del modelo en 2 módulos independientes, el primer módulo clasificó 20 características técnicas logrando un ajuste en el periodo de prueba del 93.48% de precisión, y el segundo modulo clasificó los sentimientos del usuario en un ajuste del 70.95%. Se realizó una comparación entre el sentimiento promedio del usuario vs el fabricante, obteniendo un 3.11 para el usuario y 3.99 para el fabricante.

Palabras Clave: BERT, análisis de sentimientos, NLP, aumento de datos.

Abstract

The use of Fitness trackers is increasingly by people to track their diet, measure their heart rate, control weight loss, count the steps, among other functions that help them monitor their health and increase their physical activity. The companies that manufacture these trackers advertise the benefits of this type of product, for which they include a description of their characteristics and benefits on their websites or on e-commerce platforms such as Amazon. On the other hand, users usually review the information provided by manufacturers and the opinions of other users. However, users look for products that better suit their needs.

In this research, user satisfaction was studied after publishing their experience of using physical activity trackers. It was evaluated whether the opinions provided are related to the quality offered by the manufacturers, through the use of natural language processing techniques.

A sentiment analysis was developed based on the technical characteristics offered by the devices, using transfer learning of a language model based on Transformer (RoBERTa). The model was retrained in a specific corpus, adapting it to the terminology of the reviews based on user ratings. The training of the model was developed in 2 independent modules, the first module classified the user's sentiment getting in the validation period a fit of 70.95% of accuracy, and the second module classified 20 technical characteristics getting a fit of 93.48% of accuracy. We obtained a comparison between the average sentiment of the user vs. the manufacturer of 3.11 and 3.99.

Keywords: Sentiment analysis, data augmentation, BERT, NLP.

Índice general

Resumen	II
Abstract	III
Índice de figuras	V
Índice de cuadros	VI
Lista de abreviaciones	VII
1 Introducción	1
1.1 Descripción del Problema.	3
1.2 Objetivos	3
1.2.1 Objetivo General	3
1.2.2 Objetivos Específicos.	3
2 Metodología	5
2.1 Módulo de Recolección de Datos.	6
2.2 Módulo de Preprocesamiento.	7
2.2.1 Normalización y Tokenización.	7
2.2.2 Segmentación en Cláusulas.	7
2.2.3 Lematización y Etiquetado Gramatical.	7
2.2.4 Eliminación de Palabras Vacías.	8
2.2.5 Módulo de Extracción de Características.	8
2.2.6 Módulo de Etiquetado de Sentimientos.	10
2.2.7 Módulo Aumento de Datos.	10
2.2.8 Módulos de Modelación.	11
2.2.9 Módulos de Visualización de Datos.	12
3 Resultados y Análisis	15
Conclusiones	24
Referencias	26
Appendices	28
Anexo A	29

Índice de figuras

2.1	Sistema Evaluación de Características Técnicas	5
2.2	Estructura de una reseña en plataforma de Amazon	6
2.3	Extracción de características, técnicas de las reseñas	9
2.4	Devlin J. (2019), Procedimientos Generales de preentrenamiento y ajustes para BERT (p. 4173)	12
2.5	Panel Resumen	13
3.1	Análisis de Sentimientos para periodo de Validación.	18
3.2	Análisis de Aspectos para periodo de Validación.	19
3.3	Comparativo del Sentimiento por Aspecto.	20
3.4	Comparativo del Sentimiento por Dispositivo.	21
3.5	Variación media del Sentimiento por Aspectos.	22
3.6	Variación media del Sentimiento por Dispositivos.	23

Índice de cuadros

2.1 Aspectos Técnicos.	8
3.1 Cantidad de reseñas de Dispositivo por Módulo.	15
3.2 Cantidad de reseñas después de aumento de datos.	16
3.3 Cantidad de reseñas por Sentimiento.	17

Lista de abreviaciones

BERT Bidirectional Encoder Representations from Transformers.

ESPOL Escuela Superior Politécnica del Litoral.

LSTM Long Short-Term Memory.

NLP Natural Language Processing.

RoBERTa Roberta a Robustly Optimized BERT Pretraining Approach.

1

Introducción

La actividad física se vincula estrechamente con los hábitos de vida de las personas, permitiéndoles mejorar su salud física y mental. Uno de cada cuatro adultos (1,400 millones de personas en el mundo) no realizan actividad física diaria de intensidad moderada o vigorosa [1]. La Organización Mundial de la Salud (OMS) indica que, la actividad física tiene muchos beneficios para la salud. La prevalencia actual y las tendencias de la actividad física insuficiente entre los adolescentes en edad escolar de 11 a 17 años mostraron una prevalencia mundial similar del 78.4% para los niños y del 84.4% para las niñas, utilizando información de encuestas escolares de 105 países. Reconociendo la importancia y urgencia de reducir los niveles globales de actividad física insuficiente.

Los objetivos de la Asamblea Mundial de la Salud en 2018 plantean disminuir la actividad física insuficiente en un 15% para 2030 [2]. Según esta asamblea, a nivel mundial, las mujeres son menos activas que los hombres (32% - 23%) y la actividad se reduce a mayores edades en la mayoría de los países. Además, las poblaciones más pobres y marginadas; así como las personas con discapacidad y con enfermedades crónicas tienen menores oportunidades de mantenerse más activas.

El sedentarismo es mayor en los países ricos y entre las mujeres. Una cuarta parte de la población mundial es sedentaria, y Latinoamérica es la región con más población con actividad física insuficiente, por lo que la inactividad en este contexto se ha convertido en un factor de riesgo para la salud pública. El sedentarismo en consecuencia multiplica la morbilidad y mortalidad, producto de las enfermedades degenerativas, cardiovasculares, metabólicas y ciertos tipos de cáncer [3].

Según el Canadian Fitness and Lifestyle Research Institute's (CFLRI), el monitor de actividad física del 2016 al 2018 afirma que solo el 25% de los canadienses mayores de 18 años, practica deporte [4]. En otro estudio, se encontró que aproximadamente 6 de cada 10 adultos pasaron más tiempo frente a pantallas o sentados durante la pandemia. Además, un 60% de los canadienses aumentaron su tiempo frente a la pantalla durante la pandemia, siendo un motivo de preocupación, pues se ha demostrado que los periodos prolongados de comportamiento sedentario están relacionados con resultados negativos para la salud. Con un mayor riesgo de desarrollar enfermedades como la obesidad, diabetes tipo 2, enfermedades cardiovasculares y mortalidad prematura [5].

Para los Estados Unidos, un estudio demostró que a nivel nacional el 22.9% de los adultos entre 18 y

64 años, cumplieron con las pautas de actividades aeróbicas y de fortalecimiento muscular, cuando lo recomendando es realizar actividad física regular para reducir el riesgo de muchas enfermedades crónicas, discapacidad y mortalidad [6]. Además, el 45.2% de los adultos con una discapacidad de movilidad, informaron que realizaban actividad física aeróbica, siendo caminar la actividad física más común [7].

Para Reino Unido, un estudio del 2019, mostró que los adultos de 16 o más años, el 68% de los hombres y el 60% de las mujeres, padecen de sobrepeso incluido obesidad. Según la organización Mundial de la Salud, más de una cuarta parte de la población adulta mundial no alcanza un nivel suficiente de actividad física.

La insuficiente actividad física aumentó en un 5% (es decir del 31.6% al 36.8%) en los países de ingresos altos en el período comprendido entre 2001 y 2016 [8]. Ren et al., 2018, mostró que la intervención de monitores de actividad física en trabajadores de oficina, permite mejorar la consciencia de estar físicamente más activos, además estimula el intercambio de conocimiento con otras personas, dependiendo de los mecanismos de juegos y actividades físicas que proporcione el dispositivo [9].

El ejercicio es más gratificante cuando pueden ver números diciéndoles qué tan bien lo están haciendo. Permitiéndoles entusiasmarse más por alcanzar un objetivo específico, con una cierta cantidad de pasos al día. Los rastreadores fitness están diseñados para fomentar este tipo de pensamiento. Mostrándote notificaciones para alegrarte cuando alcanzas un objetivo, indirectamente, poniendo en manifiesto el instinto competitivo. Los rastreadores pueden ayudar a motivarnos durante los entrenamientos, brindándonos información de nuestra rutina diaria o estado físico. En combinación con un teléfono inteligente sin requerir de registros o cálculos manuales potencialmente disruptivos. Estos dispositivos se están volviendo cada vez más populares en el cuidado de la salud personal, motivando a las personas a hacer más ejercicio sin necesidad de cambios en sus estilos de vida. Algunos usuarios lo usan para medir sus signos vitales, específicamente con fines médicos en adultos mayores.

Las diversas opciones en el mercado de los dispositivos también están aumentando, y los clientes buscan los productos que mejor se adapten a sus necesidades personales y a su economía. En este estudio, se realizó un levantamiento de información de 10 modelos de rastreadores de actividad física, que cuentan con las características básicas que tiene un rastreador. Los modelos estudiados fueron ANCwear, Amazfit Band 5, Fitbit Charge 5, Fitbit Charge 4, Garmin Vivofit 4, Fitbit Inspire 2, MorePro, HUAWEI Band 4 Pro, Xiaomi Mi Band 5 y Samsung Galaxy Fit 2. De los modelos mencionados, la información proviene de los países que más aportan con reseñas de estos productos (Estados Unidos, Canadá y Reino Unido).

En general, las empresas que producen estos dispositivos muestran publicidad que describe como beneficiosos a los rastreadores físicos, precisos y fáciles de usar. Sin embargo, no hay resultados de investigaciones sobre las opiniones de los usuarios que usan estos dispositivos para confirmar los beneficios que ofrecen los rastreadores. Por tanto, surge la necesidad de cuestionar la confiabilidad y precisión por la creciente tendencia del uso de estos dispositivos.

Shafae et al., 2014, introduce el enfoque de extracción de características técnicas en los rastreadores de actividad física utilizando el texto de las reseñas obtenidas de Amazon. Las características las definen como aspectos técnicos, e introducen la importancia de los aspectos evaluando la opinión pública a través de un análisis de sentimientos. El extractor de aspectos identifica las características del producto como el precio, la duración de la batería, entre otros. Luego, el extractor de opiniones clasifica la reseña como positiva, negativa o neutra [10].

La investigación de Shafae nos sirvió como base para realizar la extracción de los aspectos técnicos, para la clasificación del sentimiento. Se tagearon las reseñas observando el sentimiento expresado por

el usuario en el texto. En nuestro estudio, se propone comparar el sentimiento promedio del Usuario vs el sentimiento promedio del Fabricante. Se entrena un modelo, con una cantidad representativa de muestras, permitiéndonos estudiar si el modelo es capaz de identificar el sentimiento y el aspecto técnico en el texto.

En el periodo de validación con las 1000 muestras, se avaluó la capacidad de estimación del modelo a través del porcentaje de precisión y exactitud. Evaluando el promedio del sentimiento pronosticado del Usuario, comparado con el del Fabricante. Se puede aprovechar el ajuste del modelo, para continuar estimando nuevas reseñas.

1.1 Descripción del Problema.

En la actualidad, no es posible comprobar si las características técnicas brindados por los rastreadores de actividad física, cumplen con la expectativa de calidad de los usuarios. Por ello, es importante conocer la experiencia de los usuarios sobre el uso de los rastreadores. La plataforma de ventas en línea de Amazon posee una cantidad representativa de reseñas sobre estos dispositivos, de la cual se puede estudiar el sentimiento de los usuarios en función a sus características técnicas. Las reseñas brindan una fuente de información técnica concerniente a la experiencia de haber usado el dispositivo luego de la compra, frente a las alternativas de información general como encuestas, tweets, posts en redes sociales o blogs en internet. Una encuesta realizada por M. Y. Ali Salmony y A. Rasool Faridi en el 2021, mostró que el buen uso de las técnicas de preprocesamiento puede lograr resultados prometedores en conjuntos de datos no estructurados extraídos de reseñas en Amazon [11].

Por lo tanto, en esta tesis se propone realizar un análisis de sentimientos mediante el uso de algoritmos de Procesamiento de Lenguaje Natural (NLP por sus siglas en ingles), atreves de las opiniones provistas por los usuarios que han adquirido previamente los dispositivos. Y además, estimar el nivel de ajuste del algoritmo en un periodo de entrenamiento con 10,570 muestras, y en un periodo de prueba con 1,000 muestras.

Justificación del Problema

1.2 Objetivos

1.2.1 Objetivo General

Crear un modelo de evaluación de características técnicas de rastreadores de actividad física, utilizando técnicas de procesamiento de lenguaje natural (NLP), mediante el análisis de reseñas publicadas en la plataforma de comercio en línea Amazon.

1.2.2 Objetivos Específicos.

1. Implementar un módulo de recolección de datos, que permita obtener las reseñas de Amazon.
2. Implementar un módulo de procesamiento de datos, que permita incluir técnicas NLP, para identificar y filtrar oraciones que contengan opiniones sobre las características técnicas.
3. Implementar un módulo de aumentación de datos, que permita equilibrar aquellas características

técnicas que no contengan un número de muestras representativas para la modelación de los datos.

4. Implementar 2 módulos de modelación de datos, que permita reconocer y predecir las características técnicas del dispositivo contenidas en el comentario, y el cuantificar el sentimiento del usuario por cada característica técnica.
5. Implementar un modelo de visualización de datos, que permita el análisis interactivo, mediante gráficos, tablas, y métricas de tipo categóricas o continuas.

Marco Teórico.

En este estudio se utilizó como referencia dos metodologías que abarcan la problemática de una forma promisoría. Alaa Shafae et al., 2014 [10], presentó una metodología que evalúa la opinión pública de los rastreadores físicos provista por las reseñas de Amazon, mediante el análisis de sentimientos para los aspectos técnicos de estos dispositivos. Para extraer el sentimiento de cada oración, desarrollaron un sistema que combina los léxicos SentiWordNet y Liu's Opinion, para asignar una polaridad en un rango de -1 a 1, promediando la polaridad de todas las palabras de la oración, clasificando el comentario como positivo, negativo o neutral. Luego de que el sentimiento se agrega con el adecuado aspecto, los datos se resumen con pruebas estadísticas y textuales. Entrenaron algoritmos de aprendizaje supervisados (Máquinas de Soporte Vectorial y clasificador Naive Bayes), con la calificación provista por los usuarios (estrellas 1 y 2 como negativas, 3 neutral, 4 y 5 como positivas), logrando compararlos con su sistema. Obteniendo resultados prometedores frente a los enfoques supervisados. Wang et al. propusieron por primera vez la incrustación de aspectos e incorporaron el mecanismo de atención en LSTM para análisis de sentimientos. Los aspectos técnicos se extraen comúnmente usando sustantivos y frases frecuentes en las revisiones de texto. Además se aplicaron filtros, como la eliminación de palabras vacías y características redundantes [12].

Para esta investigación, se pre-entrenó un modelo NLP en dos módulos. El primero abarca el reconocimiento de las características técnicas del dispositivo, y el segundo la clasificación del sentimiento expresado en la opinión del usuario relacionado al aspecto, comparando el sentimiento del usuario y del fabricante.

2

Metodología

En este capítulo se detalla la metodología definida para creación del framework (ver figura 2.1), la cual incluye 7 módulos: 1) recolección de datos, 2) preprocesamiento, 3) extracción de características, 4) tagueo o clasificación de comentarios, 5) aumento de datos, 6) módulos de modelación (clasificación de aspectos, y de sentimientos) y 7) visualización de datos. Los módulos de modelación son una propuesta de utilizar técnicas de aprendizaje profundo para la obtención de resultados más precisos en comparación a los enfoques supervisados. A continuación se presenta los módulos del framework.

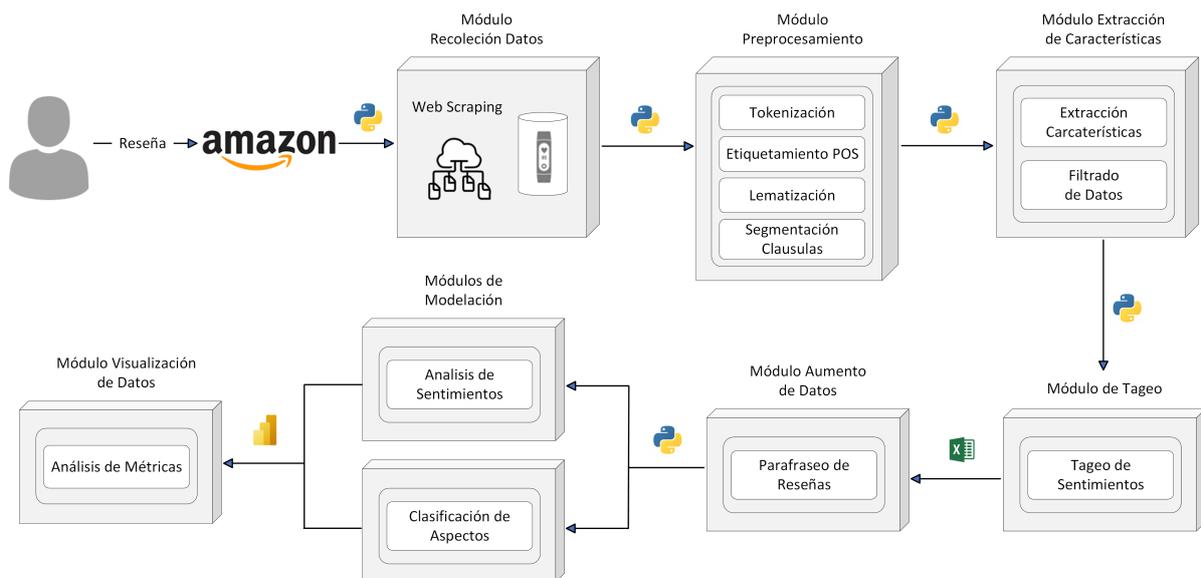


Figura 2.1: Sistema Evaluación de Características Técnicas

2.1 Módulo de Recolección de Datos.

En esta sección se describe el proceso de levantamiento de información, para el cual usamos la plataforma de comercio electrónico Amazon (ver figura 2.2), de donde se obtuvieron las reseñas de los usuarios de los dispositivos. La disponibilidad de los datos es de acceso libre, y bajo una serie de métodos, recolectamos los comentarios. En la página de cada dispositivo, se encuentran los comentarios de los usuarios y, la descripción del producto provista por el fabricante.

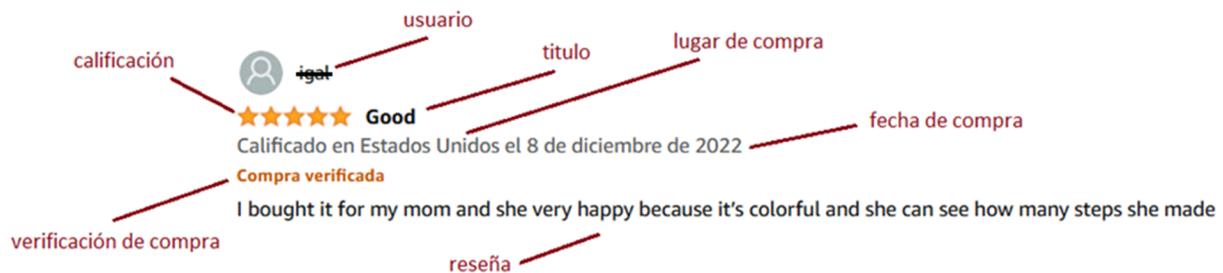


Figura 2.2: Estructura de una reseña en plataforma de Amazon

El usuario puede dar una calificación de hasta 5 estrellas, indicando el sentimiento general de la compra, que es diferente del sentimiento expresado en lo escrito en el comentario. La verificación de la compra certifica que el usuario realizó la compra y expuso su experiencia de haber usado el dispositivo. Dentro del cuerpo de la página web, también es posible obtener el nombre completo o modelo del dispositivo. La técnica usada en este trabajo de tesis para obtener datos de la web es conocida como web scraping, la cual permite extraer el contenido sin estructura de forma automática. Del código HTML se transforman los datos a otra fuente de almacenamiento como tablas o hojas de Excel. Al dirigimos a la sección de comentarios, podemos encontrar bloques de hasta 10 reseñas por página, siendo los primeros bloques ordenados en el idioma del país de comercio (inglés), luego aparecen los bloques de compradores que realizan sus reseñas en otro idioma diferente al inglés. Es decir que, el algoritmo á descarga bloques de 10 reseñas por página, hasta culminar con las reseñas disponibles. El usuario en su reseña da su opinión con respecto a las características técnicas que brinda el dispositivo, por lo que en una frase podemos encontrar ejemplos de más de 1 característica técnica. La mayoría de las veces, el usuario se expresa de modo general sobre el dispositivo; por tanto, es necesario un paso extra de filtrado de información. Para los comentarios de los Estados Unidos se utiliza el dominio www.amazon.com, para Reino Unido www.amazon.co.uk, y para Canadá www.amazon.ca. En Canadá, los usuarios acostumbran a escribir en idioma inglés y francés, por lo que se requiere de un algoritmo que reconozca el idioma a fin de filtrar solo las reseñas en inglés. De los 3 países se extraen las reseñas de cada dispositivo, y asegurándonos de obtener las mismas especificaciones por dispositivo. Los datos que se extrajeron por dispositivo fueron: el modelo, el título, la calificación obtenida, la verificación de compra, el nombre de usuario, el país donde se realiza la compra, la fecha de compra, y el comentario. Para la extracción se utilizó la librería BeautifulSoup que permite analizar documentos HTML. Para más información, favor

revisar el tutorial de descarga de comentarios de Amazon, en YouTube [13].

2.2 Módulo de Preprocesamiento.

El módulo de preprocesamiento depura y limpia el conjunto de datos para que estén listos para ser procesados en las siguientes etapas del procesamiento de texto. Se utilizó las librerías spaCy para procesamiento de lenguaje natural NLTK para el procesamiento de lenguaje simbólico y obtención de estadísticas, dfply para exploración y manipulación de datos, con funciones elegantes que permiten la manipulación de datos utilizando tuberías y langdetect, un algoritmo no determinístico de detección de idiomas en Python.

2.2.1 Normalización y Tokenización.

Esta etapa del proceso se la realiza en el siguiente orden, 1) convertir todo el cuerpo de cada reseña en un formato de palabras en minúsculas, 2) depurar la separación del texto en párrafos diferentes en 1 solo párrafo sin saltos, 3) eliminar signos de puntuación (es decir: signos de interrogación, de exclamación, y emojis, 4) transformar las palabras con apóstrofes que indican posesión, pluralización, y omisión en palabras sin apostrofes, expresiones de "isn't" se cambiaron a la forma "is not" para mejor interpretación, 5) eliminar menciones de nombres, enlaces y tokens numéricos identificados en la librería de una forma general, y 5) se realizó la tokenización que es el proceso de dividir una cadena de texto en un vector o lista de tokens. Los símbolos los reconoce como parte del cuerpo, por lo que este paso se realiza al final con la frase depurada.

2.2.2 Segmentación en Cláusulas.

Una cláusula incluye un objeto y un predicado asociado a un determinado dominio del tema. Se requiere dividir una oración en un conjunto de cláusulas. La segmentación se implementa a menudo con signos de puntuación o conjunciones en los límites de las oraciones. Se generó un conjunto de dos términos (bi-term) que resume cada cláusula, incluyendo un sustantivo que representa una característica, y una palabra de opinión (adverbio, verbo o adjetivo). El rendimiento del algoritmo depende de la calidad de las cláusulas, en referencia a la metodología de Jongho Ima et al., 2019, [14]. Con las oraciones podemos analizar mejor las características técnicas del dispositivo, clasificando aquellas oraciones que no están relacionadas con los aspectos con la clase Ninguna. En una reseña puede haber varias frases con los aspectos definidos. Las oraciones se separan en función de conjunciones como "y", signos de puntuación ".", ";", "...", ":", ":", "?". Comprobando si los tokens antes y después de la conjunción o signo de puntuación tienen diferente POS [15]. Dividiendo la reseña en varias oraciones definidas como cláusulas, por último, se tomaron y filtraron aquellas cláusulas que tienen un número de palabras de 5 a 60, para desear una mejor interpretación del modelo en la opinión analizada.

2.2.3 Lematización y Etiquetado Gramatical.

El proceso de tokenización, consiste en dividir el texto en oraciones, luego en palabras, las palabras del texto se ordenan en unidades llamadas tokens, del resultado del proceso de tokenización, se realizó etiquetamiento gramatical de la reseña, por sus siglas en inglés POS Tagging (part-of-speech tagging). Posteriormente se realiza la lematización, la lematización es un proceso que cambia la forma flexionada

de las palabras en un lema correspondiente, siendo la forma en la que se encuentra en un diccionario. Se realizó un análisis de las funcionalidades básicas que ofrece un rastreador físico, como la vida útil de la batería, el contador de pasos, entre otros. Los modelos modernos incluyen funcionalidades inteligentes específicas, como el reconocimiento del ejercicio que está realizando el usuario (correr, trotar, elíptica, máquina de remo, etc.). Algunos de los aspectos en su forma base son: *battery, heart rate, mobile, price, sleep, steps, stress, swim y water*.

2.2.4 Eliminación de Palabras Vacías.

En este paso se usó la lista de palabras provistas por la librería NLTK, que permite eliminar palabras comunes no significativas, como artículos y pronombres (por ejemplo, "as", "the", "shouldn't"). Este refinamiento permite limpiar aún más la oración, que en el proceso de reconocer el aspecto, permite obtener mejores resultados, aplicar técnicas de preprocesamiento es necesaria ya que afecta el desempeño de los clasificadores [16].

2.2.5 Módulo de Extracción de Características.

A partir de las guías técnicas, manuales e información de productos proporcionados en la página de Amazon, los aspectos técnicos se definieron como (ver cuadro 2.1). Las cláusulas que no presentan en el texto ninguno de los aspectos definidos en el análisis, se les asigna la clase None. Las cláusulas pueden tener más de 1 aspecto, seleccionando el primero que aparece en la lista de tokens, ver figura 2.3.

Alarm	Connect	None	Step Counter
App	Distance Tracking	Price	Swimproof + Swim tracking
Battery Life	Heart Rate Tracking	Reminders	Syncing
Calories Burned	Mobile App	Screen	Watch Face
Clock Face	Notifications	Sleep Tracker	Waterproof

Cuadro 2.1: Aspectos Técnicos.

```
'I was worried about all the reviews that mentioned endless vibrations, but so far no issues. I didnt realize I had my notifications like texts and calls turned off, so the only time it vibrated in the first few days of wearing it was when I hit 10k steps. I'd suggest reading up on how to set the correct settings for your goals and notifications before saying all it does is vibrate. Using the app to learn how to use the fitbit is also a breeze. It's stylish and comfortable, even to wear to sleep. It definitely motivates me to get moving and the app is great for tracking your progress and customization. Was easy to set up and sync with my phone as well. Overall, I'm impressed and very happy with my purchase.'
```

```
example.CLAUSE
```

```
'i didnt realize i had my notifications like texts and calls turned off, so the only time it vibrated in the first few days of wearing it was when i hit 10k steps'
```

```
example.PREPROCESS
```

```
['not',  
'realize',  
'notification',  
'like',  
'text',  
'call',  
'turn',  
'off',  
'so',  
'only',  
'time',  
'vibrate',  
'first',  
'few',  
'day',  
'wear',  
'hit',  
'10k',  
'step']
```

```
example.ASPECT
```

```
[('notification',), ('step',)]
```

Figura 2.3: Extracción de características, técnicas de las reseñas

2.2.6 Módulo de Etiquetado de Sentimientos.

Se etiquetaron 11,570 cláusulas, asignándoles un sentimiento y un aspecto técnico. De los cuales 10,570 cláusulas se eligieron aleatoriamente para el periodo de entrenamiento y 1,000 para el periodo de validación del modelo. Se etiquetaron las cláusulas que representan el aspecto None, siendo las muestras que no tienen ninguno de los 19 aspectos definidos. Las cláusulas que se etiquetaron con un valor de 1 presentan un sentimiento muy negativo, las cláusulas con una opinión de un aspecto no funcional se etiquetaron con un valor de 2, las cláusulas con un sentimiento neutral o indiferente con un valor de 3, las cláusulas con una opinión de un aspecto funcional con un valor de 4, y las cláusulas con sentimiento muy positivo con un valor de 5.

2.2.7 Módulo Aumento de Datos.

Para solventar la falta de ejemplos de entrenamiento para cada aspecto, se generaron nuevas reseñas parafraseando a partir de las reseñas originales de forma artificial. Para problemas de NLP, es más sensible utilizarlos debido a que esos cambios pueden afectar el significado semántico de la oración, pudiendo ser un problema para el clasificador. Para balancear el número de muestras que contienen los Aspectos, se desarrolló una función que parafrasea ó veces una cláusula, a partir de ó técnicas de aumento de datos. Por cada cláusula tendremos ó nuevas cláusulas parafraseadas manteniendo el mismo sentimiento.

Los aspectos con menos muestras y que reciben un aumento de datos son: Calories Burned, Alarm, Connect, Clock Face, Mobile App, Distance Tracking, Price, Notifications, Swimproof + Swim tracking, Reminders, Syncing, Waterproof y Watch Face. Las librerías usadas fueron textattack, nlpaug, textaugment. Para textattack se utilizaron los algoritmos WordNetAugmenter, EmbeddingAugmenter, EasyDataAugmenter; para nlpaug se utilizó BackTranslationAug, y para textaugment se utilizó EDA. WordNetAugmenter aumenta el texto reemplazando las palabras con sinónimos del diccionario de sinónimos de WordNet. EmbeddingAugmenter aumenta el texto reemplazando palabras similares en el espacio de incrustación ajustado, con un ajuste de similitud del coseno que garantiza al menos un 0.8 de similitud. EasyDataAugmenter aumenta el texto con combinaciones aleatorias de intercambios de posiciones de palabras, inserciones con sinónimos aleatorios, eliminación de palabras en el texto, y reemplazo aleatorio de palabras con sinónimos. BackTranslationAug aumenta el texto con retro-traducción; es decir, traduce la entrada en alemán y luego regresa con una traducción al inglés. EDA elige n palabras de la oración que no sean palabras vacías, luego reemplaza cada una de estas palabras con uno de sus sinónimos elegidos al azar. Para la frase "it's excellent value for its price and what it offers", se obtuvo los siguientes resultados de parafraseo según sus correspondientes librerías. De la librería textattack, algoritmo WordNetAugmenter se obtuvo "it's excellent value for its price and what it fling", del algoritmo EasyDataAugmenter se obtuvo "it's and value for its price excellent what it offers" y "it's excellent value for its price and splendid what it offers", y para el algoritmo EmbeddingAugmenter se obtuvo "it's marvellous value for its price and what textattack it offers". Mientras que para la librería textaugment del algoritmo EDA se tiene "it's excellent value for its price mary leontyne price and what it offers", y para la librería nlpaug de su algoritmo BackTranslationAug se obtuvo "it is excellent value for money and what it offers". Para el resultado de BackTranslationAug se cambia el aspecto Price por Money, y se decide eliminarlo debido a que introduce una oración sin el aspecto. Para EmbeddingAugmenter la cláusula pierde sentido de contexto al cambiar el orden de las palabras, siendo un error introducido que se elimina del resultado. Controlando estos errores se escogen las cláusulas para las muestras aumentadas, y con los aspectos originales se tiene en total 658 muestras por aspecto para el periodo de

entrenamiento.

2.2.8 Módulos de Modelación.

Para los siguientes módulos, se aplicó la misma arquitectura para ambos módulos. El módulo de Análisis de Sentimientos utilizó la variable dependiente SENTIMENT, creada en el módulo de etiquetado, donde se clasificó cada cláusula dependiendo del sentimiento observado. Para el módulo de Clasificación de Aspectos se usó la variable dependiente ASPECT, obtenida en el módulo de Extracción de Características. Se utilizaron las librerías torch y transformers, la librería torch de código abierto cuenta con una gama de algoritmos de aprendizaje profundo, mientras que transformers aborda problemas de NLP, y reemplaza a los enfoques supervisados más conocidos en la última década, a su vez las arquitecturas de redes neuronales recurrentes, como los basados en Long Short Term Memory (LSTM). Los transformers facilitan una mayor paralelización durante el entrenamiento, permitiendo el entrenamiento en conjunto de datos más grandes. El modelo utilizado en este análisis, es la arquitectura basada en “Bi-directional Encoder Representations from Transformers”, por sus siglas en inglés BERT o Representación de Codificador Bidireccional de Transformadores. BERT es una técnica de redes neuronales basada en Transformers para el preentrenamiento del procesamiento del lenguaje natural. BERT es una propuesta de enmascaramiento del lenguaje, que refuerza el aprendizaje bidireccional del texto, ocultando una palabra en una oración y obligando a usar bidireccionalmente palabras en cada lado de la palabra cubierta, prediciendo la palabra enmascarada. BERT no depende del pasado ni de la memoria del modelo, debido a que se estudian las correlaciones entre cada palabra de forma simultánea, permitiendo analizar frases grandes y pequeñas. La figura 2.4 se muestra la arquitectura de BERT [17].

Para el módulo de análisis de sentimientos se usaron 5 clases de sentimientos, mientras que para el módulo de clasificación de aspectos se utilizaron 20 clases. Empleamos el aprendizaje de transferencia ajustando el modelo de lenguaje basado en Transformer, RoBERTa (Robustly Optimized BERT Pretraining Approach, Liu et al., 2019) en el conjunto de datos obtenido de las reseñas de Amazon. RoBERTa fue entrenado usando el optimizador Adam con una tasa de aprendizaje de $2e-05$ para 20 épocas, cuenta con 125 millones de parámetros, $L = 12$ capas de autoatención, $H = 768$ número de características / capas ocultas, $A = 12$ cabezales de auto atención. Es una arquitectura basada en el modelo BERT lanzado por Google en 2018 por Yinhan Liu et al., quienes mejoran el modelo BERT, entrenando por más tiempo, con una mayor cantidad de datos y eliminando el objetivo de predicción de la siguiente oración. RoBERTa es una arquitectura sólidamente optimizada para entrenar sistemas de procesamiento de lenguaje natural [18]. Para el entrenamiento de RoBERTa, utilizaron un corpus que ha sido etiquetado específicamente para una tarea determinada, produciendo resultados de vanguardia ampliamente utilizados. Los cinco corpus utilizados en inglés, fueron BookCorpus (4.2 GigaBytes), Wikipedia (12 Gigabytes), CC-NEWS (76 Gigabytes), CommonCrawl News (76 Gigabytes) que contienen 63 millones de artículos de noticias en inglés rastreados desde septiembre de 2016 a febrero de 2019. OpenWebText (38 Gigabytes) contiene texto web extraído de URL compartidas en Reddit y contenido de Historias textos. Stories un conjunto de datos que contiene un subconjunto de datos de Common Crawl (31 Gigabytes). RoBERTa promueve los sistemas auto supervisados que se pueden desarrollar con menos dependencia del etiquetado de datos, el cual requiere mucho tiempo y recursos. Las modificaciones realizadas al modelo BERT fueron: 1) entrenar el modelo por más tiempo, con lotes más grandes y con más datos, 2) eliminar el objetivo de predicción de la siguiente oración, 3) realizar el entrenamiento con secuencias más largas y 4) cambiar la máscara estática por un patrón dinámico aplicado a los datos de entrenamiento. Para evitar el uso de la máscara estática única, se separa sus segmentos con

un token de separación, los datos de entrenamiento se duplican 10 veces para que cada secuencia se enmascare de 10 maneras diferentes durante las épocas de entrenamiento. Así, cada secuencia de entrenamiento se vio con la misma máscara donde se genera un patrón de enmascaramiento diferente, para cada vez que pasamos datos al modelo.

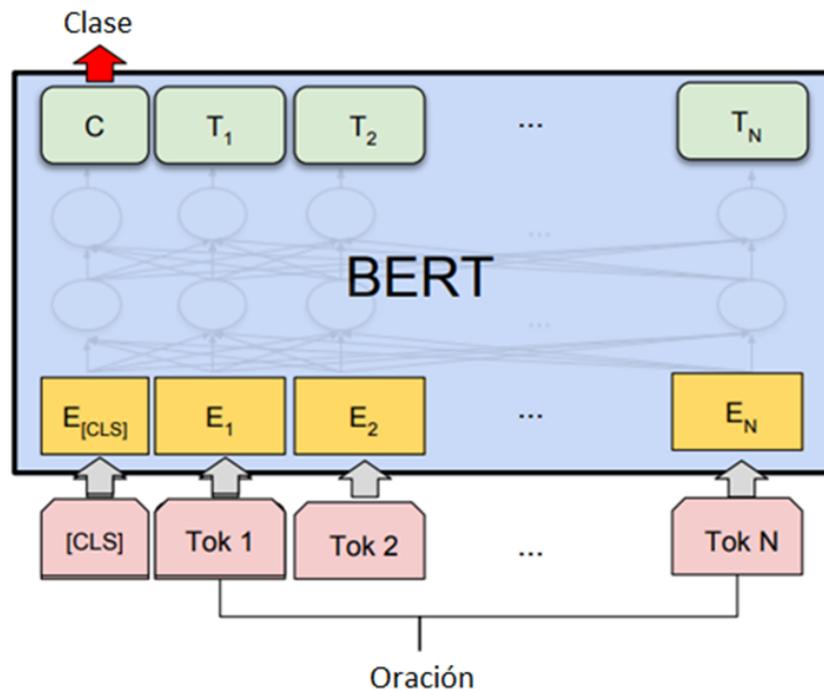


Figura 2.4: Devlin J. (2019), Procedimientos Generales de preentrenamiento y ajustes para BERT (p. 4173)

2.2.9 Módulos de Visualización de Datos.

Los resultados obtenidos de los módulos anteriores se pueden visualizar de una manera interactiva con el analista, a través de las variables que permiten utilizar filtros por sus categorías, dinamizando los gráficos, tablas o campos calculados. Los filtros que se apliquen como el país, el nombre del dispositivo, el periodo de entrenamiento y los aspectos, permiten interpretar detalladamente los resultados. De tal forma que, si se aplica el filtro de Reino Unido para País, Entrenamiento para el Periodo, No para Aumento de Datos, los gráficos y campos calculados muestran como interactúan los datos (ver figura 2.5). La plataforma de Power BI Desktop nos permite presentar los resultados del análisis en el último módulo. Power BI es una solución empresarial basado en la nube, que permite administrar diferentes fuentes de datos, analizarlos y presentarlos a través de paneles. Power Bi Desktop es una aplicación gratuita de escritorio que permite transformar, visualizar y modelar datos. El Back-End de Power Bi permite transformar los datos utilizando herramientas de código abierto como Python y software R. A continuación, se muestra el panel resumen del proyecto.

Para el análisis de las estadísticas en los resultados obtenidos del módulo de modelación, se utilizaron las siguientes métricas:

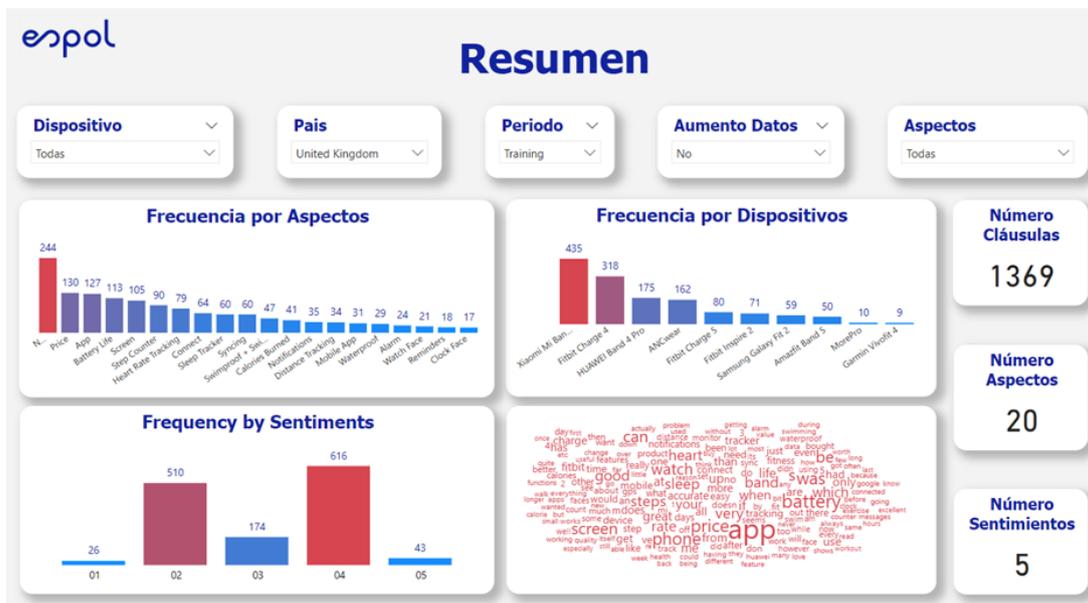


Figura 2.5: Panel Resumen

Matriz de Confusión.

Es una métrica de tipo continuo, que permite la visualización del ajuste del modelo de predicción con respecto a los valores observados. Cada columna de la matriz representa el número de predicciones de cada clase, y cada fila representa las clases observadas. Los valores que se presentan en el cuadro muestran la cuenta entre las clases predichas versus las observadas, así podemos cuantificar el desempeño entre una clase y otra. Los valores en la diagonal representan los aciertos, mientras que los que están por fuera representan los desaciertos o clasificaciones erróneas.

Exactitud.

Para conocer el ajuste del modelo, se requiere conocer la proporción que existe entre las predicciones realizadas correctamente, dividido para el total de predicciones realizadas.

Precisión.

Se conoce como el porcentaje de aciertos de clases detectadas, refiriéndose a la dispersión del conjunto de valores obtenidos, a partir de mediciones repetidas de una magnitud. A menor dispersión mayor es la precisión del modelo.

Distancia Euclidiana.

Para cuantificar que tan cerca o lejos están dos puntos, se utiliza el término de distancia. La distancia euclidiana es la que comúnmente usamos, y corresponde a la longitud de una línea recta entre dos vectores. Tomando el promedio de sentimiento del usuario y del sentimiento del fabricante, podemos conocer cuanto representa esa distancia [19].

Similitud del Coseno.

Esta medida cuantitativa, se calcula entre dos vectores distintos de cero, dentro del espacio interno del producto que mide el ángulo entre ellos. Esta medida es de utilidad en la recuperación de información en minería de texto. Se denota entre dos vectores de atributos A y B, $\cos \theta$, se representa mediante el producto escalar y la magnitud, según la ecuación 2.1.

$$\frac{A \cdot B}{\|A\| \cdot \|B\|} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}} \quad (2.1)$$

El valor del coseno varía de -1 a 1, cuanto más cercano a 1, más pequeño es el ángulo entre los dos vectores y más similar. En minería de textos se aplica la similitud de coseno con el objetivo de conocer la semejanza entre los textos [20]. Estas 5 medidas fueron utilizadas en el tablero de Power Bi, permitiendo estudiar los resultados de la predicción de los modelos, para las diferentes categorías representadas en los datos, estudiar la precisión para el periodo de entrenamiento y validación.

3

Resultados y Análisis

En esta sección, se expone el proceso por el que han ido pasando los datos, y el resultado obtenido. En el cuadro 3.1, se muestra como la cantidad de datos fue cambiando a medida que pasaba por cada módulo de procesamiento.

Dispositivo	Recolección de Datos	Extracción de Características	Aumento de Datos
Amazfit Band 5	3,270	843	1,532
Ancwear	935	552	546
Fitbit Charge 4	5,916	3,030	3,508
Fitbit Charge 5	1,859	1,659	1,731
Fitbit Inspire 2	5,685	1,109	2,394
Garmin Vivofit 4	760	668	818
Huawei Band 4 Pro	318	431	446
Morepro	2,193	1,516	1,474
Samsung Galaxy Fit 2	880	1,074	1,028
Xiaomi Mi Band 5	594	688	683
Total	22,410	11,570	14,160

Cuadro 3.1: Cantidad de reseñas de Dispositivo por Módulo.

En el primer módulo, se realizó el primer vistazo de los datos y se obtuvieron 22,410 reseñas de los tres países (Estados Unidos, Reino Unido y Canadá). Se puede observar que el dispositivo con un número mayor de muestras es Fitbit Charge 4, y el que menor muestras tuvo fue HUAWEI Band 4 Pro.

En el tercer módulo de Extracción de Características, cada comentario se dividió en cláusulas, y luego se seleccionaron solo las oraciones que tienen un aspecto, se incrementó y se filtró la base de datos; por tal motivo, se observa que en algunos dispositivos disminuyen las muestras y en otros aumenta. Por ejemplo, Amazfit Band 5 pasó de 3,270 a 843 muestras, y Samsung Galaxy Fit 2 pasó de 880 a 1,074. En el cuarto módulo se clasificaron 11,570 cláusulas obteniendo los sentimientos desde 1 como más negativo, 3 neutro y 5 más positivo. Que en el módulo de Aumento de Datos se alcanzó un total de

3 Resultados y Análisis

14,160 cláusulas (13,160 para entrenamiento y 1,000 para validación), preparando los datos para el módulo de modelación donde se entrena el modelo.

En el quinto módulo de Aumento de Datos, algunas cláusulas que hacen referencia a un aspecto, se eliminaron hasta obtener 658 muestras como regla. Debido a que los aspectos Clock Face y Watch Face que tenían un menor número de muestras (94 y 92), al parafrasear esas muestras se obtuvieron 564 y 566, logrando un total de 658. Mientras que otros aspectos como APP con 1,331 se eliminaron 673 alcanzando 658, balanceando el problema para cada Aspecto, para que no haya sesgo en el periodo de entrenamiento del modelo, ver cuadro 3.2.

Aspectos	Sin Aumento	Excluidas	Total	Parafraseadas	Con Aumento
Alarm	184	0	184	474	658
App	1,331	673	658	0	658
Battery Life	1,213	555	658	0	658
Calories Burned	263	0	263	395	658
Clock Face	94	0	94	564	658
Connect	352	0	352	306	658
Distance Tracking	178	0	178	480	658
Heart Rate Tracking	771	113	658	0	658
Mobile App	171	0	171	487	658
None	995	337	658	0	658
Notifications	341	0	341	317	658
Price	477	0	477	181	658
Reminders	118	0	118	540	658
Screen	689	31	658	0	658
Sleep Tracker	1,252	594	658	0	658
Step Counter	1,148	490	658	0	658
Swimproof + Swim tracking	196	0	196	462	658
Syncing	464	0	464	194	658
Watch Face	92	0	92	566	658
Waterproof	221	0	221	437	658
Total	10,550	2,793	7,757	5,403	13,160

Cuadro 3.2: Cantidad de reseñas después de aumento de datos.

El cuadro 3.1 muestra la cantidad de cláusulas que no tenían aumento de datos antes del proceso de aumento los datos, la cantidad de cláusulas excluidas o filtradas, la cantidad de cláusulas parafraseadas y la cantidad de cláusulas obtenidas con aumento de datos para el procesamiento en el siguiente modulo.

3 Resultados y Análisis

Las clases de los sentimientos se presentan en el cuadro 3.3, se puede observar que 6,394 muestras corresponden al sentimiento 4, que expresa un sentimiento positivo hacia el aspecto, y 176 muestras para el sentimiento 1, que expresa un sentimiento muy negativo hacia ese aspecto. Es posible balancear estas clases con una metodología que exprese sentimientos muy negativos o positivos hacia el aspecto, pero se desbalancearían las muestras que contienen un aspecto, y estaríamos introduciendo mayor error al análisis. El propósito de este análisis es comparar el sentimiento estimado del usuario con el del fabricante; de este modo estudiamos el sentimiento original presente en cada cláusula. Para el balanceo de las clases de los aspectos, recuerde que el algoritmo parafrasea 1 palabra en la cláusula sin cambiar el significado, reusando el sentimiento.

Sentimiento	Cuenta
1	176
2	5473
3	1702
4	6394
5	415
Total	14160

Cuadro 3.3: Cantidad de reseñas por Sentimiento.

3 Resultados y Análisis

Para el análisis de Sentimientos, tomando el periodo de validación, ver figura 3.1, se tomaron las muestras que no estuvieron en la fase de entrenamiento. Se obtuvo una exactitud de 81.6% y una precisión 70.95%. Esta diferencia se debe a dos motivos: 1) la métrica de precisión captura un valor diferente a la exactitud, debido a que es representativa la cantidad de muestras que están fuera de la diagonal, 2) por el desbalance de las 5 clases de sentimientos; la mayoría de las cláusulas se concentran en el sentimiento 2 y 4 con 361 y 460 muestras.

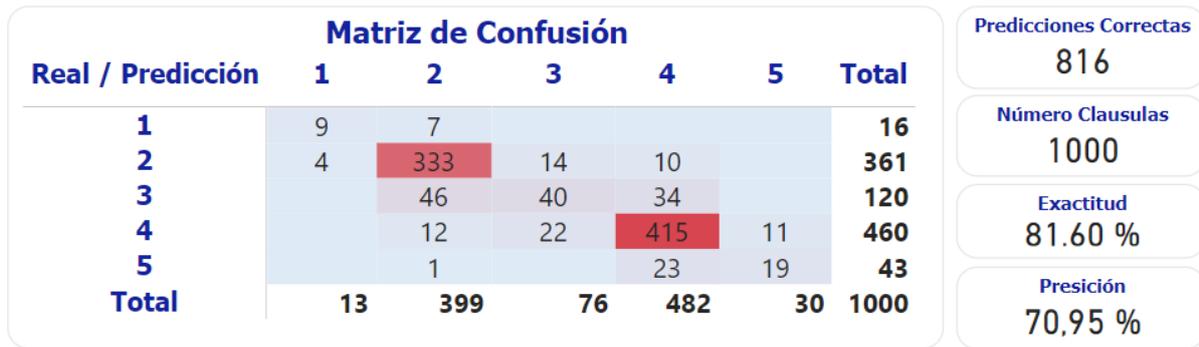


Figura 3.1: Análisis de Sentimientos para periodo de Validación.

3 Resultados y Análisis

Para el Análisis de Aspectos, tomando el periodo de validación, ver figura 3.2, se obtuvo una exactitud de 94.7% y una precisión de 93.48%. La matriz de confusión en la diagonal muestra la mayor cantidad de coincidencias entre las clases predichas versus las reales. Estas dos métricas son similares, debido a que nivelaron la cantidad de muestras por cada aspecto, en el periodo de entrenamiento.

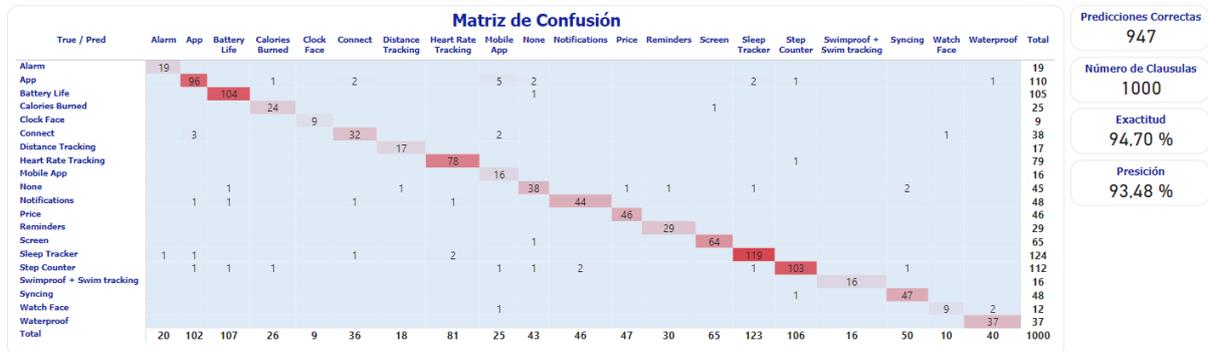


Figura 3.2: Análisis de Aspectos para periodo de Validación.

3 Resultados y Análisis

Para la comparativa del sentimiento de los Usuarios vs los Fabricantes en el periodo de validación, se agrupan los aspectos y luego se obtiene de ambos el sentimiento promedio de la predicción, ver figura 3.3. Las barras azules representan al sentimiento del fabricante, y las rojas el sentimiento de los usuarios. Se observa que el sentimiento del usuario es inferior al sentimiento del fabricante, obteniendo una Distancia Euclidiana de 14.98, un 0.97 de Similitud, y una Diferencia de Medias de 0.87. Ordenando las barras de mayor a menor, observamos que los usuarios se sienten mejor satisfechos en los aspectos de Price, Alarm, y Waterproof. Mientras una menor satisfacción hacia Syncing, Connect y Distance Tracking. La mayor diferencia de medias se observa para Mobile App, Clock Face y Distance Tracking (1.4).

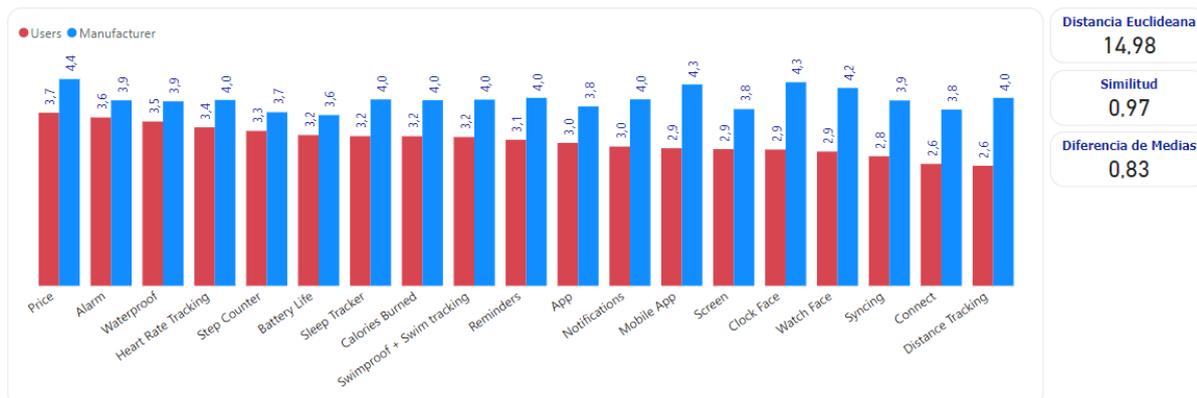


Figura 3.3: Comparativo del Sentimiento por Aspecto.

3 Resultados y Análisis

De la figura 3.4, podemos observar la comparativa del sentimiento agrupado por los Dispositivos. Teniendo el Fitbit Inspire 2 mayor aceptación que los demás competidores, mientras que Fitbit Charge 5 tiene la menor aceptación de los usuarios, en referencia a la Diferencia de Medias.

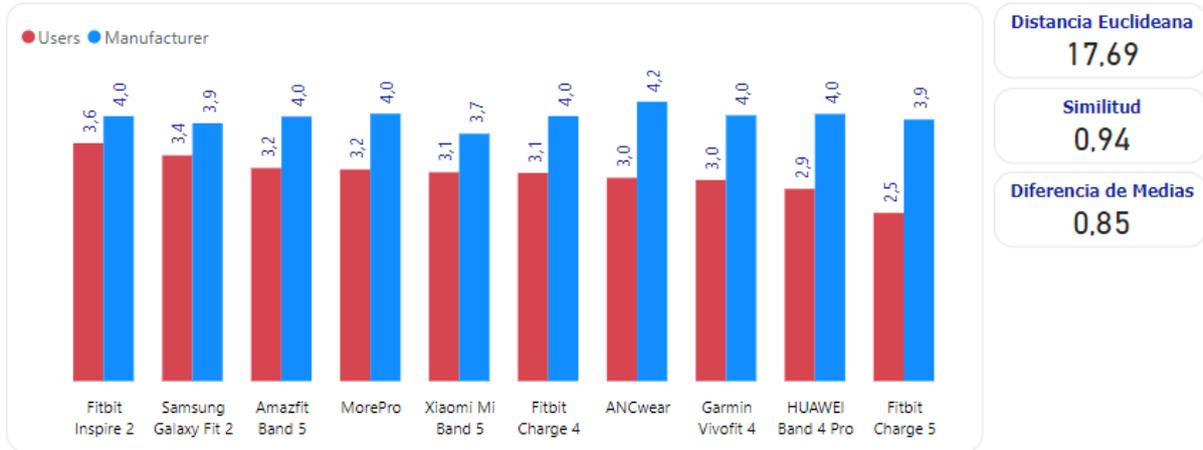


Figura 3.4: Comparativo del Sentimiento por Dispositivo.

3 Resultados y Análisis

De la figura 3.5, podemos observar la variación del sentimiento para Usuarios y Fabricantes, la mayor diferencia se da para Clock Face con 1.4, las cintas que bordean las líneas gruesas representan la desviación estándar de los datos. Para Battery Life se observa que los sentimientos expresados por los Usuarios y los Fabricantes se sobreponen, es decir que la mayor cantidad de opiniones son similares. Lo contrario ocurre para Mobile App, donde no se sobreponen las cintas.

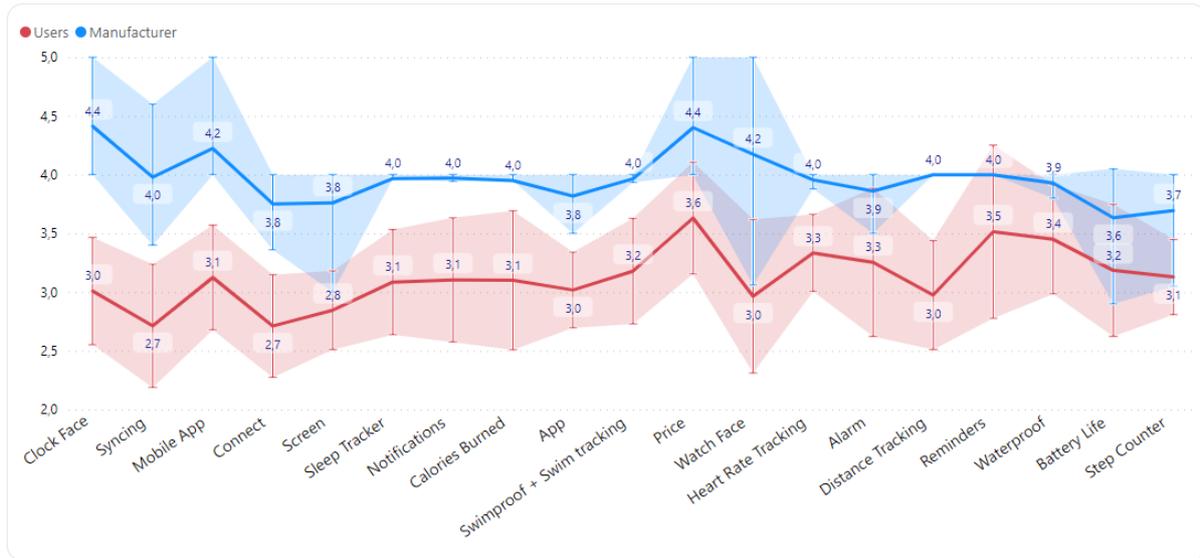


Figura 3.5: Variación media del Sentimiento por Aspectos.

3 Resultados y Análisis

De la figura 3.6, podemos observar la variación del sentimiento por dispositivos, la mayor diferencia de medias se da para Fitbit Charge 5 con 1.2. La menor diferencia se da para Xiaomi Mi Band 5, donde se sobreponen los sentimientos en un rango de 3.2 a 3.8, es decir que las opiniones son similares. Lo contrario ocurre para ANCwear, donde las cintas no se sobreponen.

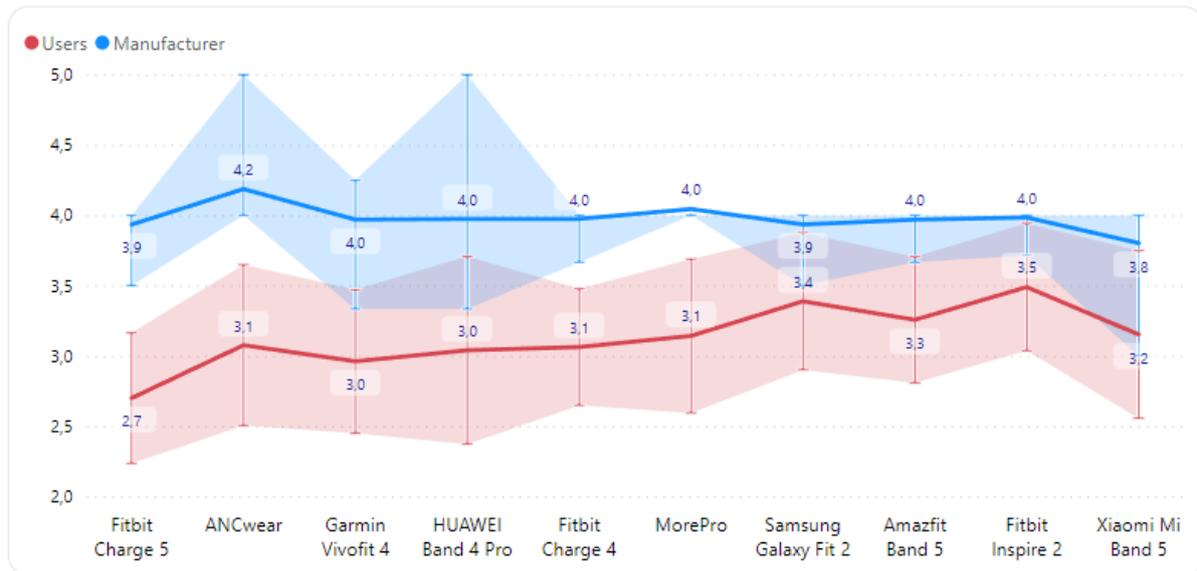


Figura 3.6: Variación media del Sentimiento por Dispositivos.

Conclusiones y Recomendaciones

El comercio en línea ha permitido que los usuarios compartan sus experiencias sobre las características ofrecidas por estos productos mediante comentarios textuales y de audio y video conocidas como reseña de productos. Además, pueden agregar una calificación de 1 a 5 estrellas, 1 como muy descontento y 5 como muy contento, como satisfacción general del producto. Con estas reseñas, otras personas que buscan conocer la satisfacción de otras pueden acceder a varios sitios web donde encontrar reseñas de todo tipo de productos. Por tanto, estas revisiones actúan como información complementaria al momento de tomar la decisión correcta del producto deseado. Las reseñas influyen en la decisión de los usuarios, confiando en los otros compradores que ya han usado el producto. La veracidad de las reseñas en los entornos de comercio electrónicos, es importante para que los usuarios puedan revisar las reseñas de productos similares. Por tanto, la reputación juega un papel fundamental en la compraventa de comercio electrónico. Elmurngi et al., 2018, indicaron que la competencia crea revisiones fantasmas para respaldar su reputación, y perjudicar a la competencia, afectando negativamente la vista de clientes y vendedores en términos de reputación. Los Investigadores presentaron un análisis de sentimiento que detecta polaridad negativa, neutral y positiva injusta en las reseñas. Utilizaron cuatro algoritmos de aprendizaje automático (Naïve Bayes, Árbol de decisión, regresión logística y Máquinas de Soporte Vectorial) para clasificar sentimientos. Los investigadores afirmaron que existe una tasa mayor de detección de críticas positivas injustas, que, de críticas neutrales o negativas, concluyendo que es más difícil para los compradores seleccionar vendedores creíbles [21]. Un estudio realizado en 2019 enfatiza la importancia de usar los votos de los usuarios como una importante fuente de información para los nuevos usuarios. Su contribución radica en un análisis estadístico del conjunto de datos públicos en la plataforma que contiene reseñas de Amazon, y la importancia de utilizar los de los usuarios en el estudio, demostraron que existe una relación inversa entre el número de reseñas y el porcentaje de votos. Los rastreadores fitness se catalogan mayormente dentro de las categorías de Health and Personal Care, Sports and Outdoors, y Cell Phones and Accessories. Con un total de 24 categorías representa el 12.21% (9.6 millones de reseñas para 79.44 millones en total), mientras que la categoría con más reseñas es la de Books, seguida de Electronics con un 28.33% y un 9.85%. Para la categoría Electronics con 7.8 millones de reseñas, el 58% de ellas que fueron alrededor de 4,5 millones, son representadas por la categoría con cero votos, mientras que la segunda clase de 1 a 5 votos representa cerca del 33% del total reseñas [22].

W. Liao et al., 2021, propusieron un diseño red neuronal basada en el aprendizaje multitarea para extraer las características del texto y emplear la representación de características para la clasificación multitarea en diferentes categorías de aspectos. Predice la polaridad del sentimiento de las diferentes categorías de aspectos. Aplicaron el mecanismo de atención cruzada para guiar al modelo en centrarse en la parte del texto más relacionada con la categoría de aspecto dada (RACSA). Para controlar desequilibrio de muestras de categorías pequeñas, integraron un factor de balance logarítmico para mejorar el peso de aprendizaje de las categorías de polaridad de sentimiento. Su modelo, logró mejores resultados en comparación con RoBERTa base y LSTM en aproximadamente un 2% de precisión [23].

M. Phan y P. Ogunbona, 2020, propusieron una solución acoplada similar, su modelo multicanal extrae los aspectos y clasifica los sentimientos de los aspectos en una sola solución de extremo a extremo. Las mejoras sustanciales resaltan el bajo rendimiento de los modelos de incrustación contextualizados recientes en la "comprensión" de las características sintácticas, proponiendo desarrollar más incrustaciones contextualizadas de aprendizaje de sintaxis. Este análisis exploró el aspecto gramatical de

la oración empleando un mecanismo de autoatención. También propusieron la distancia relativa sintáctica para restar importancia a los efectos adversos de las palabras no relacionadas, que tiene una conexión sintáctica débil con los aspectos, aumentando la precisión del clasificador de sentimientos de aspectos. En BERT, el contexto global se representa explícitamente como un par de textos que consta de una oración contextual y términos de aspectos. Cuando un token en el contexto global pertenece a un primer o segundo segmento del par de oraciones, su token de segmento se indexa como 1 o 2 respectivamente. Esta característica de predicción de la siguiente oración del modelo BERT permite que los modelos en Clasificación de Aspectos y Sentimientos basados en BERT capturan la relación semántica entre la oración contextual y el aspecto. Dado que RoBERTa elimina la tarea de predicción de la siguiente oración al entrenar el modelo, la representación de RoBERTa no es tan informativa como la representación de BERT para la tarea de Clasificación de Aspectos y Sentimientos [24].

Conclusiones.

Según nuestros resultados, el modelo es capaz de reconocer el aspecto en el texto de la cláusula, con una precisión del 93.48%, a su vez cuando no existe ningún aspecto, predice la clase None. Para el análisis de sentimientos, el modelo es capaz de asociar el sentimiento del comentario con una precisión del 70.95%. El comparativo entre el sentimiento del usuario versus el fabricante, muestra que la percepción general del usuario es menor que la oferta del fabricante, con una diferencia media de 0.87 (3.11 Usuario vs 3.99 Fabricante). Con relación al análisis de aspectos, Battery Life y Step Counter son los aspectos que menor diferencia media presentan (aproximadamente 0.6), situándose la mayor cantidad de sentimientos en el intervalo de 3.1 a 3.7. Mientras que Clock Face y Syncing presentan la mayor diferencia de medias (aproximadamente 1.4). Los fabricantes posicionan a Clock Face en un sentimiento promedio de 4.4 mientras que los usuarios en un 3. Para el análisis de los dispositivos, Xiaomi Band 5 y Fitbit Inspire 2 muestran menor diferencia media entre las opiniones de los usuarios y fabricantes (aproximadamente 0.6), en el intervalo 3.2 – 3.8. Además para Xiaomi Band 5, la variación de los sentimientos entre usuarios y fabricantes se sobreponen. Podemos concluir que el sentimiento promedio global se sitúa por encima del valor neutral con una similitud del 0.97, concluyendo que la calidad ofrecida por el fabricante cumple con la necesidad del usuario.

Recomendaciones.

Para el análisis de sentimientos, se requiere aumentar el número de muestras, para equilibrar las clases que tienen menor número de muestras, para nuestro caso el sentimiento de 1, 3 y 5. Así podremos aumentar la precisión del modelo, intentando no desequilibrar las clases de los aspectos. Este paso requiere aumentar a otros dispositivos que tengan las mismas características técnicas, y analizar si existen mayor número de reseñas con sentimientos muy negativos, para la clase 1, sentimientos neutros para la clase 3, y sentimientos muy positivos para la clase 5.

Referencias

- [1] R. Guthold, G. A. Stevens, L. M. Riley, and F. C. Bull, "Worldwide trends in insufficient physical activity from 2001 to 2016: a pooled analysis of 358 population-based surveys with 1.9 million participants," *The Lancet Child & Adolescent Health*, 2018.
- [2] —, "Global trends in insufficient physical activity among adolescents: A pooled analysis of 298 population – based surveys with 1.6 million participants," *The Lancet Child & Adolescent Health*, 2019.
- [3] W. B. Kannel, "Habitual level of physical activity and risk of coronary heart disease. the framingham study," *Canadian medical association Journal* 96, pp. 811–812, 1967.
- [4] C. F. . L. R. Institute, "Nature of sport participation, let's get active! pphysical activity in canadian communities, bolletin 8," 2019. [Online]. Available: <https://cflri.ca/sites/default/files/node/1687/files/CFLRI%20PAM%202016-2018%20Bulletin%208%20-%20Nature%20of%20sport%20participation.pdf>
- [5] —, "Monitoring & tracking the field: Changes in behavior due to the pandemic," 2022. [Online]. Available: <https://cflri.ca/sites/default/files/2022-03/CFLRI-TrackingTheField-Covid19-Changes-Behaviour-EN.pdf>
- [6] D. L. Blackwell and T. C. Clarke, "Monitoring & tracking the field: Changes in behavior due to the pandemic," 2018. [Online]. Available: <https://cflri.ca/sites/default/files/2022-03/CFLRI-TrackingTheField-Covid19-Changes-Behaviour-EN.pdf>
- [7] N. D. Hollis, Q. C. Zhang, A. C. Cyrus, E. Courtney-Long, and D. D. C. Ka. Watson, "Physical activity types among us adults with mobility disability, behavioral risk factor surveillance system, 2017," *Disability and Health Journal*, vol. 13, 2020.
- [8] O. M. de la Salud, "Actividad física," 2022. [Online]. Available: <https://www.who.int/es/news-room/fact-sheets/detail/physical-activity>
- [9] X. Ren, B. Yu, Y. Lu, and A. Brombacher, "Exploring cooperative fitness tracking to encourage physical activity among office workers." *Proceedings of the ACM on Human-Computer Interaction*, vol. 2, 2018.
- [10] A. Shafae, H. Issa, S. Agne, S. Baumann, and A. Dengel, "Aspect-based sentiment analysis of amazon reviews for fitness tracking devices," *Springer International Publishing Switzerland*, 2014.
- [11] M. Y. A. Salmony and A. R. Faridi, "Supervised sentiment analysis on amazon product reviews," *A survey. 2nd International Conference on Intelligent Engineering and Management (ICIEM)*, pp. 132–138, 2021.
- [12] Y. Wang, M. Huang, L. Zhao, and X. Zhu, "Attention-based lstm for aspect-level sentiment classification," *Proceedings of the 2016 conference on empirical methods in natural language processing*, pp. 606—615, 2016.
- [13] J. W. Rooney, "How i scrape amazon reviews using python, requests & beautifulsoup," 2021. [Online]. Available: <https://www.youtube.com/watch?v=DIT8rwyPEns&t=4>

- [14] Im, Jongho, Song, Taikgun, Lee, Youngsu, Kim, and Jewoo, "Confirmatory aspect-based opinion mining processes," *arXiv preprint arXiv: 1907*, 2019.
- [15] D. M. Koupaei, T. Song, K. S. Cetin, and J. Im, "An assessment of opinions and perceptions of smart thermostats using aspect-based sentiment analysis of online reviews," *Building and Environment*, vol. 170, 2020.
- [16] E. Haddi, X. Liu, and Y. Shi, "The role of text pre-processing in sentiment analysis," *Procedia Comput. Sci.*, vol. 17, p. 26–32, 2013.
- [17] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *Proceedings of NAACL-HLT 2019. Minneapolis, Minnesota*, 2019.
- [18] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach. computation and language," *arXiv preprint arXiv: 1907.11692*, 2019.
- [19] N. Bourbaki, "Topological vector spaces," *Springer*, 1987.
- [20] P. N. Tan, M. Steinbach, and V. Kumar, "Introduction to data mining," *Addison-Wesley*, 2005.
- [21] E. Elmurngi, E. I., and A. Gherbi, "Unfair reviews detection on amazon reviews using sentiment analysis with supervised learning techniques," *Journal of Computer Science*, 2018.
- [22] S. AlZu'bi, A. Alsmadiv, S. AlQatawneh, M. Al-Ayyoub, B. Hawashin, and Y. Jararweh, "A brief analysis of amazon online reviews," *Sixth International Conference on Social Networks Analysis, Management and Security*, 2019.
- [23] W. Liao, B. Zeng, X. Yin, and P. Wei, "An improved aspect-category sentiment analysis model for text sentiment analysis based on roberta. applied intelligence," *Appl Intell* 57, pp. 3522–3533, 2021.
- [24] M. H. Phan and P. Ogunbona, "Modelling context and syntactical features for aspect-based sentiment analysis," *School of Computer Science and Software Engineering*, 2020.

Appendices

Anexo A

Se comparte el github público del proyecto:

<https://github.com/gabofreq/FITNESS-TRACKERS-ANALYSIS>

Donde se encuentran los archivos planos de Excel, con los datos de cada módulo de datos, también se encuentran los script del código de python, con la respectiva explicación de cada módulo.