ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL

Facultad de Ingeniería en Electricidad y Computación

Evaluación de efectividad de monitores de actividad física mediante el análisis de reseñas en tiendas en línea

PROYECTO INTEGRADOR

Previo la obtención del Título de:

Ingeniera en Ciencias de la Computación

Presentado por:
Eunice Alexandra Gálvez Nan
Adriana Lourdes Riofrío Silva

GUAYAQUIL - ECUADOR
Año: 2022

DEDICATORIA

El presente proyecto le dedico a mi familia, a mis padres por apoyarme en cada aspecto de mi vida y siempre impulsar una mejor versión de mí misma; a mi hermano Daniel por su incondicional guía y apoyo; asimismo a mis docentes y amigos que me han brindado sus conocimientos, soporte y amistad y han hecho tan amena esta etapa universitaria.

Eunice Alexandra Gálvez Nan

DEDICATORIA

El presente trabajo lo dedico a mi familia: a mis padres, por su constante apoyo y guía; y a mi hermano, Andrés, por estar siempre a mi lado y ayudarme en toda mi vida universitaria.

Adriana Lourdes Riofrío Silva

AGRADECIMIENTOS

Agradezco a todos los docentes que Dios me ha puesto a lo largo de mi vida. Gracias por inculcarme su pasión por la asignatura que imparten, pero sobre todo por enseñarme más que conceptos académicos, enseñarme a ser buena ciudadana, persona. Gracias a mis padres, hermano, amigos que también han sido profesores en mi vida y han ayudado a formar la persona que soy ahora; finalmente agradezco también a mi país por brindarme educación de excelencia y gratuita.

Eunice Alexandra Gálvez Nan

AGRADECIMIENTOS

Mi más sincero agradecimiento a mi familia, por acompañarme y guiarme a lo largo de este trayecto. A nuestros tutores, por sus consejos y apoyo para la realización de este proyecto, y finalmente a ESPOL, mi alma máter.

Adriana Lourdes Riofrío Silva

DECLARACIÓN EXPRESA

“Los derechos de titularidad y explotación, nos corresponde conforme al reglamento de propiedad intelectual de la institución; *Eunice Alexandra Gálvez Nan* y *Adriana Lourdes Riofrío Silva* damos nuestro 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"

Eunice Alexandra
Gálvez Nan

Adriana Lourdes
Riofrío Silva

EVALUADORES

Firmado electrónicamente por:
BORIS XAVIER VINTIMILLA BURGOS

Boris Vintimilla

PROFESOR DE LA MATERIA

Eduardo Cruz

PROFESOR TUTOR

RESUMEN

Los monitores de actividad física son dispositivos cuya popularidad está en aumento como instrumento para mejorar la salud y el estilo de vida de una persona. Los usuarios a menudo usan las descripciones provistas por los fabricantes respecto a las caracțerísticas funcionales del dispositivo al comprarlo, que prometen efectos positivos; sin embargo, no hay resultados significativos que comprueben la veracidad de estas afirmaciones. En este proyecto se propuso un *framework* para evaluar la eficiencia de los monitores de actividad física usando reseñas de Amazon. Este consistió en cuatro módulos que incluyeron la recolección de datos, preprocesamiento de datos, extracción de características y análisis de sentimientos. Para este último se usaron dos métodos, uno con un enfoque léxico con SentiWordNet y otro con un enfoque basado en Inteligencia Artificial con el modelo BERT. A partir de los resultados se encontró que el *score* del modelo BERT para clasificar el sentimiento de las características de los dispositivos como positivo, negativo o neutral, fue de 81% en precisión y 72% en exactitud, en comparación con la precisión del 55% y exactitud del 66% obtenidas con SentiWordNet. En base a los resultados obtenidos con ambas metodologías se concluye que la efectividad de los monitores de actividad física es sobreestimada por parte de los fabricantes. Los usuarios perciben que el rendimiento de las características ofrecidas en los dispositivos es menor al publicitado.

Palabras Clave: Análisis de sentimientos, monitores de actividad física, BERT, SentiWordNet

ABSTRACT

Activity trackers are devices that are growing in popularity as a tool to improve a person's health and lifestyle. Users often use the descriptions provided by the manufacturers regarding the functional features of the device when buying it, which promise positive effects; however, there are no significant results that prove the veracity of these statements. This project proposes a framework to evaluate the efficiency of activity trackers using Amazon reviews. It has of four modules that included data collection, data preprocessing, feature extraction, and sentiment analysis. For the latter, two methods were used, one with a lexical approach using SentiWordNet and the other with an approach based on Artificial Intelligence using BERT model. Based on the results, the precision and accuracy of the BERT model to obtain the score and classify the sentiment of the features of the devices as positive, negative, or neutral, was 81% and 72%, respectively, compared to the precision of 55% and accuracy of 66% obtained with SentiWordNet. It is concluded that the effectiveness of physical activity monitors is overestimated by manufacturers. Users perceive that the performance of the features offered on the devices is lower than advertised.

Keywords: Sentiment analysis, activity trackers, SentiWordNet, BERT

ÍNDICE GENERAL

ABREVIATURAS

ALD	Asignación Latente de Dirichlet
BERT	*Bidirectional Encoder Representations from Transformers*
ESPOL	Escuela Superior Politécnica del Litoral
LSTM	*Long-Short Term Memory*
MAF	Monitor/Monitores de Actividad Física
ML	*Machine Learning*
NLTK	*Natural Language Toolkit*
PLN	Procesamiento de Lenguaje Natural
POS	*Part Of Speech*
SVM	*Support-Vector Machine*

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

CAPÍTULO 1

1. Introducción

En el sector salud, la tecnología corporal promete efectos positivos para el cuidado personal. El uso de este tipo de dispositivos ha sido una tendencia en crecimiento desde 2016 [1], en este contexto, su popularidad —junto con la de los teléfonos inteligentes— ha aumentado especialmente en países desarrollados. Por ello, se espera que el número de dispositivos activos a nivel mundial pase de 526 millones en 2017 a aproximadamente 1.1 billones en 2022 [2].

Específicamente, los monitores de actividad física (MAF) son dispositivos que están ganando popularidad debido a sus beneficios, ya que proveen información acerca de las rutinas diarias de los usuarios sin requerir mediciones potencialmente disruptivas. Investigaciones muestran que pueden ayudar a prevenir el riesgo de desarrollar enfermedades cardiacas y sirven para diagnosticar problemas médicos rápidamente [3], [4]. Además, tienen potencial para incrementar la actividad física y la motivación [5] [6]. Esto permite que las personas puedan ejercitarse más sin implementar cambios significativos a su estilo de vida y que controlen su estado de salud a lo largo del tiempo [7].

Los monitores de actividad física son el tema principal de este proyecto. En este capítulo, determinamos el alcance del proyecto, que se centra en la necesidad de medir la efectividad de los MAF para verificar que las especificaciones publicitadas por sus fabricantes se cumplen. En la siguiente sección, establecemos la motivación para este trabajo y tanto las limitaciones como los enfoques actuales sobre la evaluación de los MAF. Asimismo, se presenta la relevancia de una solución que beneficie tanto a los usuarios como a los fabricantes.

También definimos nuestro objetivo principal como la creación de un *framework* que evalúa la efectividad de los MAF y formulamos cuatro objetivos específicos que describen los pasos para realizarlo. Finalmente, en el marco teórico presentamos una revisión del estado del arte sobre el análisis de sentimientos basado en características, que se usa para resolver el problema planteado.

1.1 Descripción del problema

En los últimos años, la cantidad de monitores de actividad física disponibles en el mercado ha aumentado y por ello los consumidores buscan productos que se ajusten mejor a sus necesidades. Los fabricantes de los MAF los publicitan como beneficiosos, amigables con los usuarios y precisos. Sin embargo, no hay resultados significativos que comprueben la veracidad de estas afirmaciones [8]. Los MAF están relacionados al bienestar de los usuarios, por lo que elegir uno es una decisión importante que si se realiza de forma inadecuada puede llegar a “representar un riesgo para la salud” [8].

Los usuarios utilizan las descripciones provistas por los fabricantes al momento de comprar un MAF y tienen interés en su efectividad. Por lo tanto, existe la necesidad de contrastar la información que proveen los diseñadores de estos dispositivos y las perspectivas de sus usuarios.

1.2 Justificación

Hay varios estudios en los que se ha investigado la efectividad de los monitores de actividad física en diferentes contextos, que se han realizado con enfoques experimentales y cualitativos [9] [10]. Aquellos con enfoques experimentales se centran en aspectos tecnológicos; usan grupos de participantes en ambientes controlados para probar la precisión y confiabilidad de los dispositivos, pero proveen una “perspectiva limitada sobre las interacciones del usuario con el sistema” [9]. Por otro lado, los que se enfocan en aspectos cualitativos recolectan datos mediante encuestas o entrevistas y evalúan características de usabilidad en base a la experiencia de los usuarios después de utilizar el dispositivo [5].

A pesar de que está información es valiosa, no está disponible ni es accesible para el público general y tiene limitaciones como el número de características evaluadas. De manera que, para los usuarios, la manera más fácil de conseguir esta información sobre la efectividad de los dispositivos es a través de los comentarios de consumidores en tiendas en línea. Esta fuente de datos es, por lo tanto, una alternativa para evaluar los MAF.

Así que, un modelo que involucra a las características del dispositivo —tal como el fabricante las publicita— con las percepciones directas de los usuarios es útil para facilitar la información sobre la relación entre las expectativas de estos dispositivos y su impacto real de forma más transparente y accesible. Además, aborda las limitaciones mencionadas anteriormente y permite incluir tanto los aspectos positivos como negativos asociados a los dispositivos. Esto ayuda también para tomar decisiones durante la compra de un MAF o durante su proceso de diseño.

1.3 Objetivos

1.3.1 Objetivo general

Crear un *framework* de evaluación de monitores de actividad física, mediante el análisis de reseñas publicadas en tiendas en línea, para medir la efectividad de estos productos.

1.3.2 Objetivos específicos

1. Definir criterios de selección de MAF disponibles en el mercado para la elección de cuatro dispositivos en tiendas en línea con sus respectivas características.

2. Recolectar reseñas de los MAF seleccionados para la identificación de las características más comunes de estos dispositivos según los usuarios.

3. Implementar un modelo de análisis de sentimientos para la clasificación de sentimientos asociados con cada característica.

4. Reportar los resultados obtenidos mediante visualizaciones estadísticas y métricas para comparar la opinión de los usuarios sobre las características de cada uno de los MAF seleccionados con las especificaciones provistas por su fabricante correspondiente.

1.4 Marco teórico

Existen investigaciones relacionadas sobre el análisis de reseñas de usuarios acerca de productos o servicios. Para ello, usan modelos de análisis de sentimientos. Algunas se enfocan en teléfonos móviles [11] [12], termostatos inteligentes [13], y aplicaciones móviles [14], [15]. Sin embargo, hay pocos estudios específicos sobre monitores de actividad física, como el presentado en [16].

1.4.1 Análisis de sentimientos

El análisis de sentimientos, también conocido como minería de opinión, es una técnica de Procesamiento de Lenguaje Natural (PLN) usada para analizar texto y extraer opiniones sobre un tópico. Con ello se determina si la opinión es positiva, negativa o neutral. Según la literatura, existen dos tipos de análisis de sentimientos: basado en características y no basado en características [13]. El enfoque basado en características se centra en extraer las características a partir de un texto, que son características representativas del tópico general, para identificar el sentimiento asociado a cada una. Este tipo de análisis es el que se describe a continuación, puesto que se ajusta a los objetivos planteados para este proyecto. Se puede dividir en una serie de pasos que incluyen: preprocesamiento de texto, extracción de características, clasificación de sentimientos y detección de polaridad, y evaluación [17] [18].

1.4.1.1 Preprocesamiento de texto

Esta fase es importante para mejorar el resultado del análisis de reseñas. En [13], algunas de las técnicas empleadas incluyen la remoción de palabras vacías (como preposiciones), emojis, signos de puntuación innecesarios y la conversión de palabras con apostrofe. También se usa *stemming*, un procedimiento que reduce una palabra a su raíz [18], ya que es necesario cuando el algoritmo de minería de opinión elegido requiere PLN y etiquetado gramatical *Part of Speech* (POS); con POS se asigna a cada palabra su categoría gramatical correspondiente, como se muestra en la Figura 1.1.

También se utiliza para reducir el número de características en textos para las técnicas de *Machine Learning* (ML) aplicadas en [15].

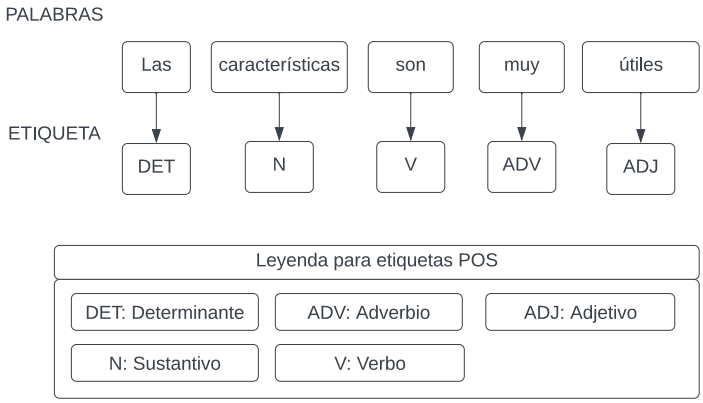

Figura 1.1. Proceso básico de POS *tagging*

Yiran y Srivastava [11] eliminan símbolos problemáticos como “-” y “/”, entre otros, además de las palabras vacías. Mientras que Samy et al. [14] aplican *tokenization*, *stemming* y lematización; lematización es un procedimiento que cambia la forma flexionada de las palabras a lema correspondiente, que es la forma en la que se encontraría en un diccionario. En otro enfoque, M et al. [12] hacen una corrección ortográfica, expansión de jerga y conversión a letras minúsculas como parte del preprocesamiento.

1.4.1.2 Extracción de características

Para este paso hay dos enfoques: predefinir características de los dispositivos o extraerlos a partir de las reseñas. Para el enfoque en el que se predefinir las características, Malekpour Koupaei *et al.* [13] usan una técnica de minería de opinión confirmatoria basada en características en la que se establecen las categorías que se buscan en las reseñas. Este algoritmo, llamado DiSSBUS, es una propuesta de Im *et al.* [19] y tiene seis pasos: *disintegrating*, *summarizing*, *straining*, *bagging*, *upcycling* y *scoring*. En el primer paso, se usa POS para separar el texto en cláusulas; considera las palabras antes y después de una conjunción para determinar si la oración hace referencia a una misma característica. Si el POS de las palabras no es el mismo, la

oración puede ser dividida en dos cláusulas. Para el siguiente paso, se genera un conjunto de dos términos denominado *bi-term* que resume cada cláusula; incluye un sustantivo, que representa una característica, y una palabra de opinión (verbo, adverbio o adjetivo).

Para el paso de *straining*, cada *bi-term* se clasifica como común y no común en base a la frecuencia con la que aparece en el conjunto de reseñas. En el paso de *bagging*, los *bi-terms* comunes se asignan a las categorías predefinidas, mientras que en el paso de *upcycling* los no comunes se colocan en una o varias de las categorías predefinidas que más se relacionen con estos. Este algoritmo es útil porque uno de los desafíos del análisis a nivel de oraciones es el análisis de oraciones que contengan más de un tópico. Sin embargo, su rendimiento depende de la calidad de las cláusulas, por lo que no funciona bien con oraciones largas y detalladas [19].

Por otro lado, en [16] se predefinen características de monitores de actividad física. Las características se clasifican como “dinámicas” si las características de los dispositivos difieren de las de otros, mientras que se consideran “estáticos” si se trata de características que comparten todos los dispositivos analizados. Para obtener estas características se usan dos conjuntos de palabras. El primero incluye un conjunto de sustantivos simples y compuestos que se usan en las características de los productos o en sus descripciones. En tanto que el segundo conjunto incluye sustantivos que se utilizan frecuentemente con verbos (por ej., *sleep tracking*, *tracks steps*). Además, realizó una revisión manual para mantener las palabras más generales y se agruparon características similares en base a “sinónimos”, “derivaciones” e “hipónimos” en WordNet.

En [14], [11] y [20] se emplea Asignación Latente de Dirichlet (ALD), una técnica de modelamiento de tópicos, para extraer las características a partir de las reseñas, empleada por su “simplicidad y efectividad en aplicaciones prácticas” [20]. En [11], también se considera que una oración puede describir más de una característica por lo que las reseñas se dividen por puntuación y conjunciones. Similarmente, Hong y Wang [20] proponen un *framework* para la extracción de características y resumen de

opiniones. Emplean reglas gramaticales con POS y ALD para encontrar características y después usan un modelo de red neuronal recurrente, *Long-Short Term Memory* (LSTM), para resumir cada característica con sus respectivos sentimientos; este método funcionó para reseñas cortas y largas.

1.4.1.3 Clasificación de sentimientos y detección de polaridad

La clasificación de sentimientos y detección de polaridad es un proceso en el que una opinión expresada en un texto se clasifica como positiva, negativa o neutral, dependiendo de las palabras usadas en una oración. Para esta fase existen enfoques basados en ML, léxicos e híbridos. En [13], para esta fase que recae en el último paso del algoritmo DiSSBUS, se emplea el enfoque basado en léxicos para determinar la polaridad de los sentimientos y asignar un score. Para ello, se usa el diccionario WordNet, con el que se hace la clasificación en base a los valores de sustantivos, adjetivos y verbos. Yiran y Srivastava [11] aplican SentiWordNet, una extensión de WordNet [18], para el etiquetamiento de sentimientos. Este se describe en la Figura 1.2.

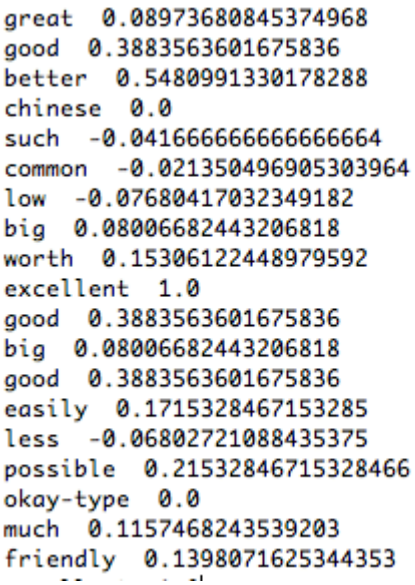

Figura 1.2. Ejemplo del diccionario SentiWordNet

Por otra parte, en [12], [21] se usa *Support-Vector Machine* (SVM), una técnica de ML utilizado para problemas de clasificación como el análisis de sentimiento. En [12] se utiliza 400 000 reseñas recolectadas en Amazon para clasificar tres características en

dos categorías: positivo y negativo. Así mismo, en [21] se utiliza 2000 reseñas de películas de IMDb.com y 400 opiniones de productos como películas, libros, autos y teléfonos para clasificar dichos datos en positivo y negativo. Por otro lado, Panichella *et al.* [15] aplican *Naive Bayes* para predecir el sentimiento de 2090 reseñas recolectadas de la App Store y Google Play. Además, [15] para reducir el ruido usa la distribución ji al cuadrado x^2 a fin de seleccionar las palabras que se consideran más importantes para determinar un sentimiento. Las oraciones se etiquetan manualmente y el modelo se entrena para obtener un valor entero en el rango [-1,1].

Por otro lado, [22], [23] evaluaron diversos modelos de inteligencia artificial para identificar el sentimiento de reseñas y obtuvieron los mejores resultados con el modelo BERT, *Bidirectional Encoder Representations from Transformers*, modelo diseñado para la comprensión del lenguaje, el cual está entrenado con una ingente cantidad de datos no etiquetados en profundas representaciones bidireccionales a fin de que los *tokens* puedan ser comprendidos en contextos de ambos sentidos y ayudar a la desambiguación de las palabras. Debido a su arquitectura y su vasto preentrenamiento, el modelo guarda ya un conocimiento del lenguaje, por lo que permite resolver tareas de PLN añadiendo una capa más al modelo, evitando el uso de tiempo y recursos computacionales que se hubiesen requerido para entrenar un modelo específico desde cero. Entre las tareas que puede realizar se encuentran: análisis de sentimientos, reconocimiento de entidades, traducción, entre otros [24].

1.4.1.4 Evaluación

La evaluación del modelo de análisis de sentimientos depende de las técnicas elegidas en los pasos previos. M. *et al.* [12] miden la eficiencia del clasificador SVM con cinco métricas: *recall*, matriz de confusión, *accuracy*, *precision* y *F-score*. Similarmente, Yiran y S. Srivastava [11] evaluaron el rendimiento con matrices de confusion, *recall* y *F-score*; para este paso etiquetaron las oraciones manualmente [11]. Adicionalmente, Panichella *et al.* [15] etiquetaron manualmente todas las oraciones para cada categoría y evaluaron el conjunto de verdad generado por humanos con la clasificación que se generó automáticamente, mediante *precision*, *recall* y *F-score*.

En base a esta revisión de literatura, se decidió aplicar dos técnicas de análisis de sentimientos para asignar un *score* al sentimiento de cada característica de los dispositivos. La primera es SentiWordNet, que se basa en un enfoque léxico, mientras que la segunda es el modelo BERT, que usa un enfoque basado en Inteligencia Artificial. En la fase de preprocesamiento, para ambos enfoques, se descartan las reseñas en idiomas diferentes al inglés; Luego, para SentiWordNet se eligió la eliminación de puntuación, emojis y *stopwords*; y la aplicación de *tokenization* y lematización. En cambio, para el modelo BERT se escogió un *tokenizer* para realizar la conversión de texto a datos numéricos para que estos datos sean compatibles con el modelo. En la extracción de características, para ambas técnicas, se decidió que las características de los MAF serían predefinidas y que esta fase se realizaría aplicando la metodología propuesta en [16]. En el siguiente capítulo se describe en detalle la metodología y el *pipeline* del proyecto.

CAPÍTULO 2

2. Metodología

En este capítulo se detalla la metodología definida para crear el *framework* de evaluación de MAF. Se dividió en cuatro módulos de acuerdo con las fases identificadas para el proceso de análisis de sentimientos. Estos consisten en: 1) recolección de datos, 2) preprocesamiento de datos, 3) extracción de características y 4) análisis de sentimientos. Posteriormente se resumieron los resultados finales de cada dispositivo en un *dashboard* creado con la plataforma Power BI, como se indica en la Figura 2.1.

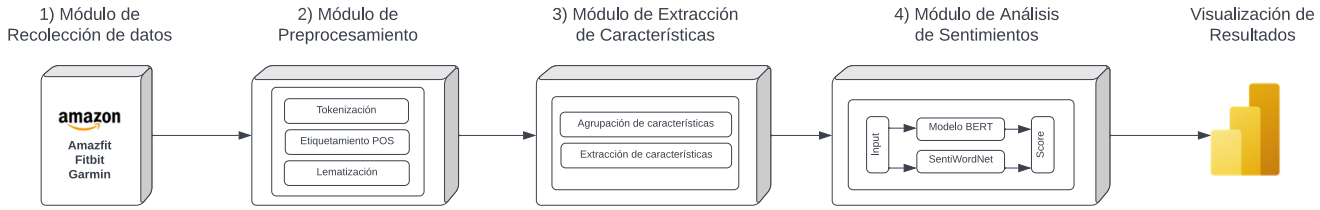

Figura 2.1 Módulos del *framework* de evaluación de MAF

Los módulos de extracción de características y de análisis de sentimientos se dividen en dos fases: una para fabricantes y otra para las reseñas escritas por los usuarios, como se muestra en la Figura 2.2; los mismos pasos presentados en los módulos en la Figura 2.1 se aplicaron para cada una de las fases en ambos módulos y el rendimiento de cada modelo se comparó a partir de métricas de evaluación para determinar cuál generaba los mejores resultados.

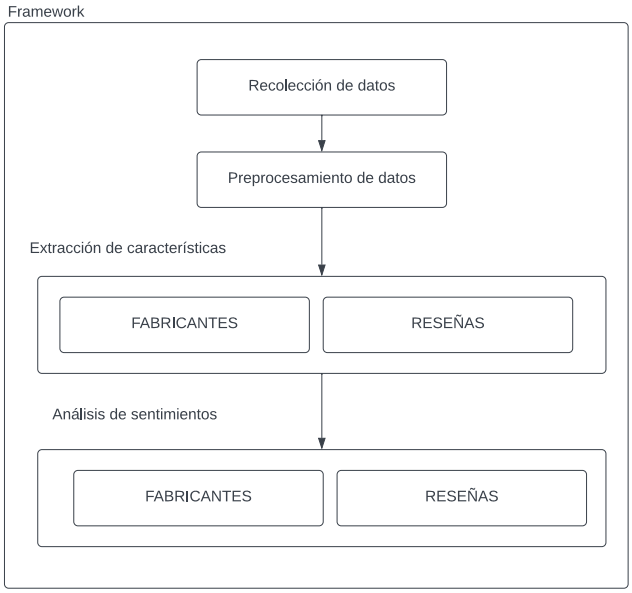

Figura 2.2 División de fases para módulos del *framework* evaluación de MAF

2.1 Recolección de datos

En esta etapa se realizaron tres tareas, que se centraron en la definición de criterios para la selección de tres MAF y la recolección de reseñas de cada dispositivo. Previo a esto, se decidió que se usaría la tienda de comercio en línea Amazon para obtener los datos debido a que contiene una gran variedad de productos disponibles, están organizados en categorías y contienen información provista por los fabricantes. Además, la estructura de las reseñas en Amazon es útil para extraer la información necesaria para el análisis, ya que entre sus características tiene: calificación numérica de la reseña; utilidad de la reseña; e información sobre la compra, es decir, si la compra es verificada [25].

2.1.1 Definición de criterios de selección de monitores de actividad física

Existen diversas marcas que constan entre las más dominantes en el mercado de los MAF, entre ellas Fitbit, Polar, Garmin, Misfit, Jawbone y Coros [1]. Entre las características de los MAF que se consideran antes de elegir modelos para su análisis están la variedad de lugares en los que se pueden usar, que pueden ser distintas partes del cuerpo como la cintura o la muñeca; la variedad de tipos de monitores de actividad,

como pedómetros, acelerómetros y aquellos que funcionan con aplicaciones de teléfonos inteligentes; y la variedad de actividades a las que hace seguimiento [26].

Los MAF que se usan en la muñeca y que permiten conectarse a teléfonos inteligentes han sido los más populares entre los usuarios [27], principalmente debido a que analizan datos biométricos y presentan esta información de manera simple, además de distinguirse por sus usos multipropósito gracias a elementos “utilitarios y estéticos” [28].

Por lo tanto, en base los resultados encontrados en la revisión de literatura y a los requerimientos para este proyecto, se eligieron los siguientes criterios para elegir un MAF: ser parte de la lista de productos “más vendidos a nivel internacional” (“*International Best Sellers*”) provista por Amazon, el MAF debe usarse en la muñeca y tiene que conectarse a teléfonos inteligentes.

2.1.2 Elección de tres MAF en base criterios de selección definidos

En base a los criterios definidos previamente y a una búsqueda exhaustiva en la página de productos “más vendidos a nivel internacional” de Amazon —en la categoría de “Electrónicos y Dispositivos”—, se eligieron tres dispositivos que cumplieron con estos requisitos, de las marcas Fitbit, Amazfit y Garmin respectivamente: “Fitbit Inspire 2 Health & Fitness Tracker with a Free 1-Year Fitbit Premium Trial, 24/7 Heart Rate, Black/Black, One Size (S & L Bands Included)”, “Amazfit Band 5 Activity Fitness Tracker with Alexa Built-in, 15-Day Battery Life, Blood Oxygen, Heart Rate, Sleep & Stress Monitoring, 5 ATM Water Resistant, Fitness Watch for Men Women Kids, Black” y “Garmin vívofit 4 activity tracker with 1+ year battery life and color display. Small/Medium, Black. 010-01847-00, 0.61 inches”.

2.1.3 Recolección de reseñas y refinamiento del *dataset*

Para este proyecto se trabajó con reseñas en idioma inglés, extraídas de tres países, para tener más datos para complementar el análisis. Se seleccionaron los tres siguientes países: Estados Unidos, Reino Unido y Canadá. Primero se consideró a

Estados Unidos porque es el principal mercado para la tecnología corporal; culturalmente, hay una mayor influencia para adoptar otras formas de hacer ejercicio además de las tradicionales, como ir al gimnasio, debido a las facilidades brindadas por esta tecnología [29]. Similarmente, se incluyó a Canadá ya que la tecnología corporal es la cuarta tecnología en aumento en este país [30]. Además, de manera general, el uso de dispositivos corporales es la segunda tendencia más popular en Norteamérica [31]. Por otro lado, en Reino Unido también hay una tendencia considerable en los usuarios para usar un MAF, con un 15% de usuarios en este país en 2020 [32].

Las reseñas se extrajeron para cada dispositivo en amazon.com para Estados Unidos, amazon.co.uk para Reino Unido y amazon.ca para Canadá. Además, se filtraron con la opción de "Compra verificada" de Amazon para recolectar únicamente aquellas que han sido escritas por personas que han adquirido el producto. Después se realizó un refinamiento del *dataset* eliminando las reseñas en idiomas diferentes al inglés. En la Tabla 2.1 se resume la cantidad de reseñas usadas finalmente para cada dispositivo.

Tabla 2.1 Resumen de reseñas para monitores de actividad seleccionados

Dispositivo	Reseñas (EE. UU)	Reseñas (Canadá)	Reseñas (Reino Unido)	Total
Fitbit Inspire 2	4667	30	989	5686
Amazfit Band 5	2963	41	266	3270
Garmin vívofit 4	550	196	14	760

2.2 Preprocesamiento de datos

Como parte de esta fase se realizaron tres tareas para limpiar los *datasets* recopilados y refinarlos para mejorar el análisis. Para ello se usaron las librerías spaCy y *Natural Language Toolkit* (NLTK) en Python.

2.2.1 Normalización y *Tokenization*

Con la normalización se convirtieron las reseñas a un único formato mediante la transformación del texto a letras minúsculas. También se eliminaron signos de puntuación que no eran significativos, como signos de interrogación y de exclamación, y

signos de puntuación usados para representar emoticones. Asimismo, las palabras con apóstrofes que indican posesión, omisión y pluralización se transformaron a palabras sin apóstrofe, es decir, expresiones como "*isn't*" se cambiaron a la forma "*is not*". Se removieron menciones de nombres, enlaces y *tokens* numéricos. Finalmente se realizó *tokenization* para separar las palabras de una reseña. Para realizar esta etapa se utilizaron las herramientas de la librería spaCy.

2.2.2 POS *Tagging* y Lematización

Con los resultados obtenidos tras realizar *tokenization*, se realizó etiquetamiento POS a cada *token* de la reseña, es decir, se asignó a cada palabra su correspondiente categoría gramatical. Posteriormente se realizó la lematización, con la librería spaCy, a cada *token*, es decir que, se transforman las palabras a su forma base usando las reglas definidas en las etiquetas POS. Por ejemplo, palabras como "*steps*" o "*counted*" se redujeron a "*step*" y "*count*", respectivamente.

2.2.3 Eliminación de *stop words*

Para esta etapa se usó la lista de *stop words* provista por la librería NLTK para eliminar palabras comunes que no son significativas, como artículos (por ej., "*a*", "*the*") y pronombres. También se hizo un refinamiento manual para evitar la eliminación de palabras que podrían ser útiles para el análisis de sentimientos, como las palabras con negaciones, por ejemplo: "*shouldn't*". Cabe recalcar que esta remoción se hizo únicamente para el análisis con SentiWordNet.

2.3 Extracción de características

Esta etapa consiste en la extracción de características de los dispositivos para ello se realizaron dos tareas para la identificación de características en las reseñas. La primera consistió en la extracción y clasificación de características de acuerdo con las descripciones provistas por el fabricante, que se realizó de forma manual. Mientras que la segunda consistió en la extracción de características a partir de las reseñas recolectadas por cada dispositivo.

2.3.1 Extracción de características establecidas por el fabricante

En esta sección, se establecieron 2 conjuntos de características candidatas para obtener las características de los dispositivos según lo definido en [33]; el primero conjunto (conjunto A) consiste en todos los sustantivos y sustantivos compuestos de las secciones “Sobre este artículo”, “Descripción del producto” y “Del fabricante”, en la página de Amazon de cada dispositivo; y el segundo conjunto (conjunto B), incluye a los sustantivos que por lo general van acompañados de verbos específicos, obtenidos de la misma sección de Amazon del conjunto A. Algunas de las características candidatas recolectadas se ejemplifican en la Tabla 2.2, entre ellas: número de pasos, monitoreo de frecuencia cardíaca, calorías quemadas, entre otras y en la Tabla 2.3 se presentan ejemplos de características candidatas para cada uno de los conjuntos.

Tabla 2.2 Ejemplos de características candidatas

Fitbit Inspire 2	Amazfit Band 5	Garmin vívofit 4
Heart rate tracking	*Battery*	*Step tracking*
Step counter	*Speed tracking*	*Alarm*

Tabla 2.3 Ejemplos de los conjuntos obtenidos

Conjunto A	Conjunto B
Step counter	*Calories burned*
Calories tracker	*Sleep monitoring*

Por otra parte, dadas las distintas variaciones en que una característica puede aparecer en un texto (por ej., *calories tracker* y *calories burned*), también se hizo un refinamiento agrupando manualmente todas las posibles variaciones de una misma característica.

2.3.2 Extracción de características en las reseñas

Para extraer las características de los dispositivos en las reseñas se utilizaron los conjuntos de características obtenidos en el paso anterior y una segmentación por cláusulas para evitar que en la medida de lo posible un mismo texto haga referencia a varias características.

Para la segmentación de cláusulas, primero, los comentarios se separaron en oraciones en base a los signos de puntuación ".", ";", "…" y después, se identificaron las conjunciones como "*and"* y otros signos de puntuación como ",", "?"; para verificar las siguientes dos condiciones, la primera condición se cumple si los *tokens* antes y después de la conjunción o signo de puntuación tienen diferentes POS [13]. La segunda se cumple si la primera potencial cláusula de la oración ya tiene un aspecto y la segunda potencial cláusula tiene otro aspecto, también se puede cumplir si la primera potencial cláusula ya tiene más de un aspecto.

Finalmente se clasificó cada cláusula en una o más características dependiendo de las derivaciones identificadas en la cláusula. Por lo que, como resultado final de este módulo, se generó un conjunto de cláusulas para cada característica.

2.4 Análisis de sentimientos

En esta etapa se realizó la extracción de sentimientos y detección de polaridad tanto para las características de los dispositivos, a partir de las descripciones de los fabricantes, como para la opinión sobre las características, a partir de las reseñas. También se evaluó el rendimiento del *framework* con cuatro métricas. Para ambas fases de análisis de sentimientos se usó primero SentiWordNet y posteriormente el modelo BERT.

SentiWordNet es un recurso léxico que asigna una puntuación a cada palabra. Contiene *synsets*, que son "conjuntos de sinónimos que representan un concepto" [34], con tres puntajes numéricos que indican el valor positivo, negativo y objetivo del sentimiento expresado —sus valores se sitúan en el rango [0.0, 1.0]— y la suma de los tres es igual a 1. Para el análisis se consideró a los verbos, sustantivos, adjetivos y adverbio como palabras que expresan un sentimiento. SentiWordNet usa etiquetamiento POS, por lo que se utilizó la librería spaCy. En la Tabla 2.4 se presentan las equivalencias utilizadas entre POS y SentiWordNet.

Tabla 2.4 Equivalencias entre POS y SentiWordNet

Nombre	Abreviación POS	Abreviación SentiWordNet
Sustantivo	NN	n
Verbo	VB	v
Adjetivo	JJ	a
Adverbio	RB	r

Para el modelo BERT se empleó una versión multilingüe especializada en el análisis de sentimientos de reseñas de productos, *nlptown/bert-base-multilingual-uncased-sentiment* [35]. Retorna una predicción de *ranking* del 1 al 5 de la reseña ingresada. La versión utilizada fue entrenada con 69 millones de reseñas de Amazon obtenidas desde 1995 hasta 2015 [36].

2.4.1 Análisis de sentimientos de características según fabricantes

Para esta fase, se recolectaron manualmente palabras asociadas que describan a cada una de las características identificadas previamente para cada dispositivo y se creó un *dataset* con los datos recopilados. Estas se tomaron de las mismas secciones de Amazon que se usaron en el módulo de extracción de características.

Para SentiWordNet, se realizó el etiquetamiento POS a todas las palabras asociadas a cada característica y después se obtuvieron todos los *synsets* que provee SentiWordNet para esa palabra con esa etiqueta POS. Para cada *synset* se obtuvo su puntaje positivo y negativo y se calculó su correspondiente *score* (que se encuentra en el rango [-1.0, 1.0]) con la ecuación 2.1 [34]. Después se determinó el *score* de la palabra a partir del promedio de los *scores* de todos sus *synsets* asociados con la ecuación 2.2 y, finalmente, se calculó el *score* de la característica promediando los *scores* de todas sus palabras asociadas mediante la ecuación 2.3.

$$Score_{(synset)} = [posScore] - [negScore] \quad (2.1)$$

$$Score_{(word)} = \frac{1}{n}\sum_{i=1}^{n} Score_{(synset)_i} \quad (2.2)$$

$$Score_{(feature)} = \frac{1}{n}\sum_{i=1}^{n} Score_{(word)_i} \quad (2.3)$$

En el caso del modelo BERT, se usaron las oraciones completas de las descripciones de los fabricantes de cada uno de los grupos de aspectos previamente identificados y se obtuvo el *ranking* de sentimiento de la descripción.

2.4.2 Análisis de sentimientos de características según reseñas de usuarios

Los pasos de esta fase se resumen en la Figura 2.3. Este proceso se realizó para cada una de las cláusulas asociadas a cada característica; cada cláusula se analizó también en base las palabras que la componen.

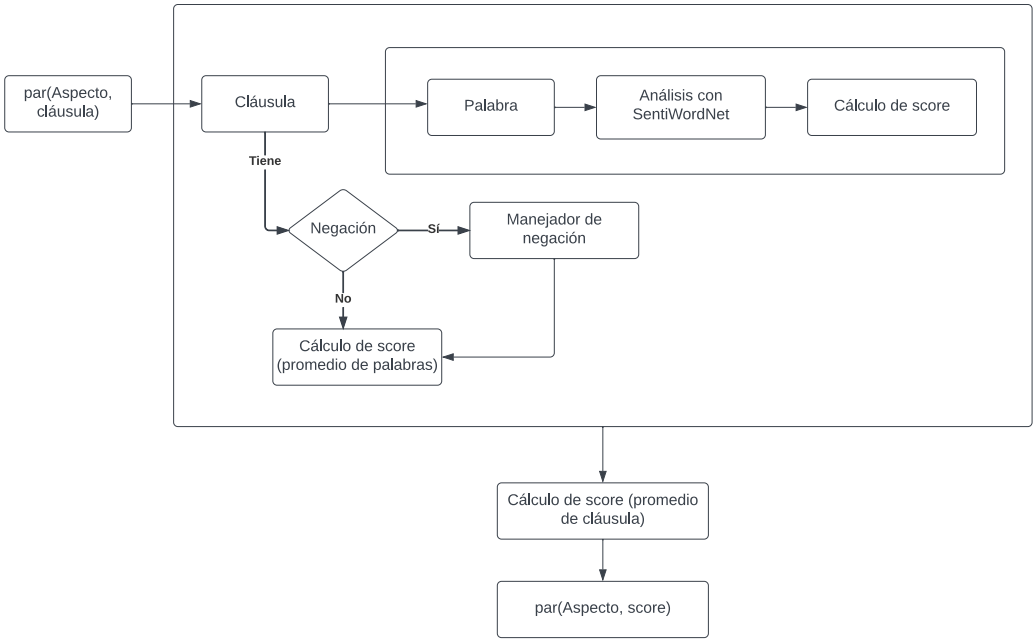

Figura 2.3 Pasos para el análisis de sentimientos de reseñas

Como *input* de esta fase se tiene conjuntos de cláusulas para cada característica que se identificó en las reseñas. Similar a los pasos del análisis de las características según los fabricantes, se obtuvieron todos los *synsets* de una palabra y para cada uno se obtuvo su *score* mediante la ecuación 2.2. Después se obtuvo el *score* de la palabra promediando los *scores* de todos sus *synsets* con la ecuación 2.3.

Asimismo, se incluyó una etapa de detección y manejo de negación; en cada cláusula se comprobó si tenían palabras que indicaran negación (por ej., “*not*”, “*never*”,

"*no*"). Para ello, se empleó el árbol de dependencias de la cláusula para detectar las relaciones semánticas y determinar el alcance de la negación. Dependiendo de este alcance, se invirtió la polaridad de las palabras que aparecían después de la negación [33].

Después se calculó el *score* de toda la cláusula promediando el *score* de todas sus palabras, con la ecuación 2.3. Finalmente, se obtuvo el *score* de cada característica como el puntaje promedio de los *scores* de todas las cláusulas con la ecuación 2.4.

$$ScoreFinal_{(feature)} = \frac{1}{n}\sum_{i=1}^{n} Score_{(clause)_i} \quad (2.4)$$

Para cada dispositivo se obtuvo un conjunto de pares característica-*score* con todas las características identificadas en las reseñas. En base a la evaluación de tres *thresholds* para definir polaridad de las cláusulas [37], se decidió utilizar la distribución de todos los *scores* obtenidos por característica. Los rangos son los siguientes: [-1,-0,01] indica un sentimiento negativo, (-0.01, 0.01) indica un sentimiento neutral y [0.01, 1] indica un sentimiento positivo.

En el caso del modelo BERT, el *score* de la característica fue obtenido por el promedio del *ranking* sentimental que se obtiene con el modelo en cada una de las cláusulas donde se encontró dicho aspecto. Para todas las cláusulas que ingresaron al modelo no se realizó eliminación de *stop words* ni lematizado. La ecuación 2.5 representa el score que el modelo BERT proporciona a una cláusula cuyo rango va desde el 1 al 5. La ecuación 2.6 hace referencia al score de la característica *i* de un dispositivo el cual es el promedio de todas las cláusulas donde se encontraba esa característica en ese dispositivo, *n* representa el número total de cláusulas con característica *i*.

$$ScoreBERT_{(clause)} = m;\ m \in [1,5] \quad (2.5)$$

$$ScoreFinalBERT_{(feature)} = \frac{1}{n}\sum_{i=1}^{n} ScoreBERT_{(clause)_i} \quad (2.6)$$

Asimismo, para cada dispositivo se obtuvo un conjunto de pares característica-*score* con todas las características identificadas en las reseñas. Se utilizaron los rangos definidos en [33] para definir la polaridad de las cláusulas: [1,3) indica un sentimiento negativo, igual a 3 un sentimiento neutral y (3, 5] un sentimiento positivo.

2.4.3 Evaluación

Para validar la clasificación de polaridad de las reseñas se escogieron las métricas de evaluación *precision*, *recall*, *accuracy* y F1 *score*. *Precision* hace referencia a la proporción de reseñas que correctamente se las clasificó con una etiqueta x de todas las reseñas que se clasificó con la etiqueta *x*; donde *x* puede ser positivo, negativo o neutral; matemáticamente se lo representa en la ecuación 2.5. Por otro lado, *recall* hace referencia a la proporción de reseñas que correctamente se clasificó con la etiqueta x de todas las reseñas que realmente poseían esa etiqueta *x*, matemáticamente se lo representa en la ecuación 2.6. *Accuracy* indica el porcentaje de reseñas acertadas por el modelo y se obtiene con la fórmula en la ecuación 2.7. F1 *score* combina las métricas de *precision* y *recall*, como se muestra en la ecuación 2.8. Estas métricas se obtuvieron por cada una de las polaridades de las reseñas (positivo, negativo y neutral).

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \tag{2.5}$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ \ Negative} \tag{2.6}$$

$$Accuracy\ = \frac{True\ Positive + True\ Negative}{True\ Positive + True\ Negative + False\ \ Negative + False\ Positive} \tag{2.7}$$

$$F1\ score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{2.8}$$

Para la evaluación se utilizó un conjunto de prueba de quinientas cláusulas para cada uno de los tres dispositivos. Además, se verificó que éstas estuvieran equitativamente distribuidas por sentimiento según los resultados previamente obtenidos

por SentiWordNet: 200 cláusulas positivas, 200 negativas y 100 neutrales. Se utilizó el primer y tercer cuartil de la distribución de *score* de SentiWordNet para realizar dicha clasificación. Sólo para Garmin se utilizó el segundo cuartil debido a que el número de cláusulas era menor. Además, dichas cláusulas fueron equitativamente distribuidas por su longitud entre cortas y largas. Se utilizó el segundo cuartil del número de *tokens* para delimitar esta clasificación. Finalmente, el conjunto de prueba se etiquetó manualmente en base a una puntuación de 1 a 5.

Tras aplicar esta metodología para la creación del *framework* de evaluación de MAF, los resultados se resumieron gráficamente usando la plataforma Power BI. En el siguiente capítulo se presentan los resultados obtenidos.

CAPÍTULO 3

3. Resultados y análisis

En este capítulo se muestran los resultados obtenidos a partir de la implementación del *framework* de evaluación del rendimiento de monitores de actividad física y se hace un análisis, tanto de estos como de los métodos utilizados para conseguirlos. Primero se resumen las características obtenidas a partir de la recolección de datos y posteriormente se muestran los datos finales del análisis de sentimientos, tanto de los datos de los fabricantes como de las reseñas de los usuarios.

3.1 Análisis de resultados

Los resultados presentados en la Tabla 3.1 se obtuvieron como parte del módulo de recolección de datos. Se muestran las características encontradas para los dispositivos seleccionados de las marcas Amazfit, Fitbit y Garmin, respectivamente.

Tabla 3.1 Aspectos recolectados por dispositivo

Dispositivo	Aspectos	Total
Amazfit	*Notifications, Battery life, Sleep monitoring, Women's health tracking, Water resistance & swim tracking, Blood oxygen monitoring, Calories burned, Heart rate monitoring, Breathing exercise,* Alexa*, Stress monitoring, Watch faces, Screen, Color display, Step counter, Distance tracking, Health Assessment System, Speed tracking, App, Alarm*	20
Fitbit	*Daily Readiness Score, Heart rate tracking, Calories burned, Step counter, Sleep tracker, App, Breathing rate, Battery life, Reminders, Stress management, Active Zone Minutes, Distance tracking, Tile, Notifications, Menstrual Health Tracking, Timer & Stopwatch, Activity tracking, Water resistance & swim tracking, Badges, Challenges, Alarm, Screen, Watch faces, Breathing sessions*	24
Garmin	*Battery life, Screen, Timer,* Move IQ*, Alarm, Syncing, Distance traveled, Calories burned, Step counter, Inactivity*	16

	tracking, Sleep monitoring, Activity tracking, App, Step goal, Water resistance & swim tracking, Weather widget	

A partir de esto, se determinó que todos los dispositivos compartían algunas características. Por ello, se decidió hacer una clasificación en dos grupos para el análisis: “Aspectos Estáticos” y “Aspectos Dinámicos”. Los aspectos estáticos son aquellos que son compartidos por todos los dispositivos y se resumen en la Tabla 3.2. Por otro lado, los aspectos dinámicos son aquellos que son diferentes para cada dispositivo y que no son compartidos por todos; para Amazfit se encontró un total de 12; para Fitbit, un total de 16; y para Garmin, un total de 8.

Tabla 3.2 Aspectos estáticos

Aspecto
Alarm
App
Battery Life
Calories burned
Distance traveled
Sleep monitoring
Water resistance & swim tracking
Screen

Para la fase de análisis de sentimientos, primero se obtuvieron resultados preliminares con SentiWordNet. Inicialmente se propuso un modelo de cálculo del *score* que usaba únicamente el significado más común de una palabra; sin embargo, este enfoque se reemplazó debido a que se limitaba el posible significado positivo o negativo que podría tener en una cláusula, y con ello afectaba al *score*. En su lugar, se consideraron todos los posibles significados que tuviera una palabra en SentiWordNet. Los resultados del etiquetamiento de las cláusulas encontradas para cada característica en las tres categorías definidas (Positivo, Negativo y Neutral), para cada dispositivo, se resumen en la Tabla 3.3, Tabla 3.4 y Tabla 3.5, para Amazfit, Fitbit y Garmin, respectivamente.

Se encontró que los aspectos en la categoría “Estáticos” aparecieron en una mayor cantidad de cláusulas, a diferencia de aquellos denominados como “Dinámicos”, y tenían una mayor proporción de cláusulas clasificadas como positivas; excepto por *Battery life* en Garmin, que tenía más cláusulas negativas.

Tabla 3.3 Número de cláusulas clasificadas por sentimiento para Amazfit

Aspecto	Positivo	Negativo	Neutral
Notifications	87	33	45
Battery life	293	168	120
Sleep monitoring	554	250	193
Water resistance & swim tracking	55	13	10
Blood oxygen monitoring	45	22	27
Calories burned	51	33	23
Heart rate monitoring	334	105	97
Alexa	174	106	95
Stress monitoring	44	30	37
Watch faces	79	37	21
Screen	141	93	78
Step counter	475	248	178
Distance tracking	30	26	16
Health Assessment System	31	14	11
App	575	328	278
Alarm	34	14	10
Women’s health tracking	1	1	4
Breathing exercise	2	3	2
Color display	2	1	0
Speed tracking	4	5	2

Tabla 3.4 Número de cláusulas clasificadas por sentimiento para Fitbit

Aspecto	Positivo	Negativo	Neutral
Daily Readiness Score	1	0	0

Heart rate tracking	610	202	157
Calories burned	172	69	62
Step counter	950	323	320
Sleep tracker	1166	426	284
App	698	288	216
Battery life	323	186	82
Reminders	74	20	24
Distance tracking	35	17	12
Notifications	110	81	74
Activity tracking	18	10	6
Water resistance & swim tracking	65	21	20
Challenges	40	11	9
Alarm	26	22	10
Screen	235	217	117
Watch faces	23	23	14
Timer & Stopwatch	24	7	8
Breathing rate	0	2	0
Stress management	12	1	5
Active Zone Minutes	8	4	6
Tile	9	4	2
Menstrual Health Tracking	9	2	2
Badges	3	0	0
Breathing sessions	3	1	0

Tabla 3.5 Número de cláusulas clasificadas por sentimiento para Garmin

Aspecto	Positivo	Negativo	Neutral
Battery life	112	188	55
Screen	36	36	22
Timer	2	0	1
Move IQ	0	1	0
Alarm	2	2	0

Syncing	41	46	46
Distance traveled	8	6	6
Calories burned	4	2	9
Step counter	107	42	50
Inactivity tracking	0	1	0
Sleep monitoring	42	24	17
Activity tracking	19	8	2
App	67	24	23
Step goal	1	0	0
Water resistance & swim tracking	42	14	6
Weather widget	6	0	5

De manera general, la distribución del número de cláusulas por sentimiento y aspecto en cada uno de los dispositivos —según los resultados obtenidos con SentiWordNet— fueron en la mayoría positivos. Esta distribución se muestra en la Figura 3.1 para Amazfit, Figura 3.2 para Fitbit y Figura 3.3 para Garmin.

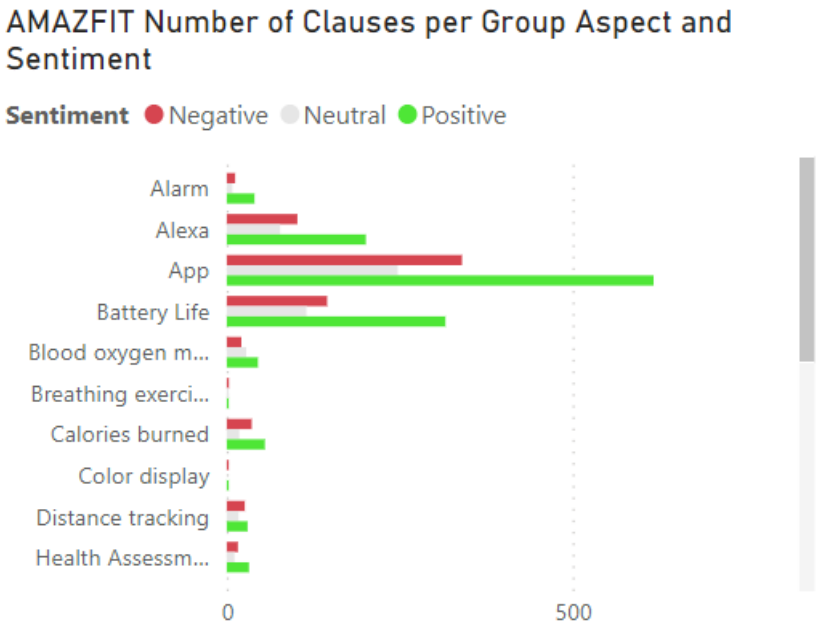

Figura 3.1 Número de cláusulas por sentimiento y aspecto - Amazfit (SentiWordNet)

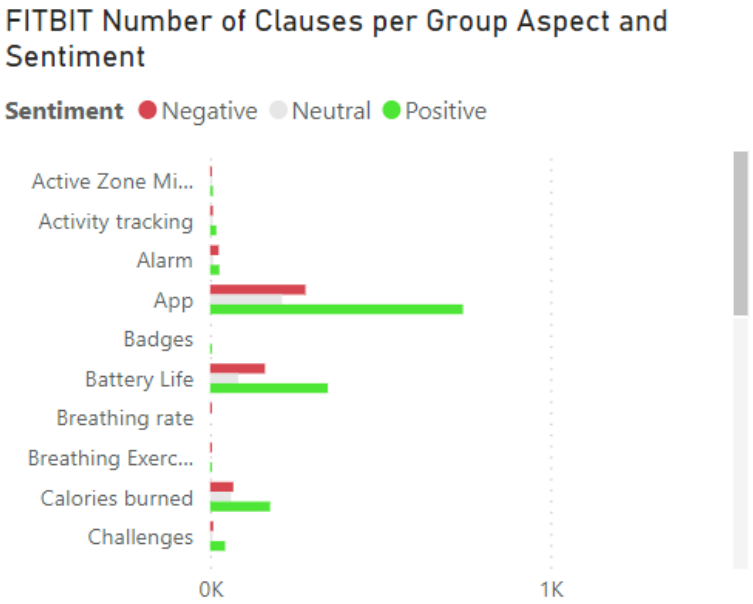

Figura 3.2 Número de cláusulas por sentimiento y aspecto - Fitbit (SentiWordNet)

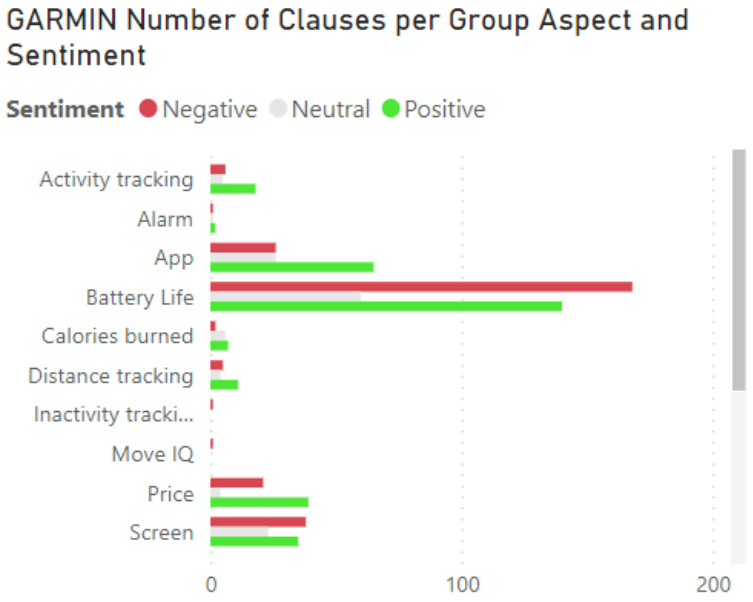

Figura 3.3 Número de cláusulas por sentimiento y aspecto - Garmin (SentiWordNet)

En la comparación entre la opinión de los fabricantes y la opinión de los usuarios a partir de las reseñas, se observó que para Amazfit el promedio de la diferencia entre fabricantes y usuarios era de 0.072; para Fitbit fue de 0.065; y para Garmin, fue de 0.046. Los valores obtenidos se resumen en la Tabla 3.6 para Amazfit, en la Tabla 3.7 para Fitbit y en la Tabla 3.8 para Garmin.

Tabla 3.6 *Score* de fabricantes y de reseñas para Amazfit (SentiWordNet)

Aspecto	***Score* Fabricante**	***Score* reseñas**
Notifications	0,265625	0,030375
Battery life	0,135417	0,044656
Women's health tracking	0,025417	0,042631
Water resistance & swim tracking	0,208333	0,103366
Blood oxygen monitoring	0	0,029059
Calories burned	0,0625	0,025231
Alexa	0,097222	0,043297
Watch faces	0	0,029609
Screen	0,230235	0,030713
Color display	0,09375	0,040548
Step counter	0,0625	0,034186
Distance tracking	0,0625	0,017755
Health Assessment System	0,20375	0,047939
Speed tracking	0,0625	0,003569
App	0	0,036288
Alarm	0	0,024776
Breathing exercises	0,02461	-0,0017
Sleep monitoring	0,14205	0,031252
Heart rate monitoring	0,05	0,026754
Stress monitoring	0,390625	0,010509

Tabla 3.7 *Score* de fabricantes y de reseñas para Fitbit (SentiWordNet)

Aspecto	*Score* Fabricante	*Score* reseñas
Active Zone Minutes	0,08056	0,01028
Activity tracking	0,05903	0,03353
Alarm	0,07813	0,02296
App	0,21875	0,05932
Badges	0,06250	0,10090
Battery Life	0,45833	0,04701
Breathing rate	0,09875	-0,03681
Calories burned	0,18837	0,02736
Challenges	0,06250	0,07339
Daily Readiness Score	0,04419	0,01190
Distance tracking	0,05903	0,02129
Heart rate tracking	0,05000	0,03427
Notifications	0,00000	0,01095
Reminders	0,09427	0,04841
Screen	0,08523	0,00356
Sleep tracker	0,14205	0,04550
Step counter	0,05903	0,03310
Stress management	0,39063	0,04370
Tile	0,04545	0,04023
Timer & Stopwatch	0,02903	0,02826
Watch faces	0,05000	0,02342
Water resistance & swim tracking	0,20833	0,07856
Women's health tracking	0,02542	0,05827

Tabla 3.8 *Score* de fabricantes y de reseñas para Garmin (SentiWordNet)

Aspecto	*Score* Fabricante	*Score* reseñas
Battery life	0	-0,000570

Screen	0	-0,002097
Timer	0	0,107025
Move IQ	0,125	-0,019271
Alarm	0,20625	0,069118
Syncing	0,25	0,018176
Distance traveled	0,059028	0,006496
Calories burned	0,041667	0,021762
Step counter	0,041667	0,035918
Inactivity tracking	0,041667	0,091700
Sleep monitoring	0,142045	0,028809
Activity tracking	0,041667	0,050917
App	0,105114	0,051738
Step goal	0,304688	0,047018
Water resistance & swim tracking	0,083333	0,150375
Weather widget	0	0,043717

El *score* de las características de los dispositivos según SentiWordNet con respecto a los fabricantes es significativamente mayor en los tres dispositivos, tal como se puede observar en la Figura 3.4 para Amazfit, en la Figura 3.5 para Fitbit y en la Figura 3.6 para Garmin.

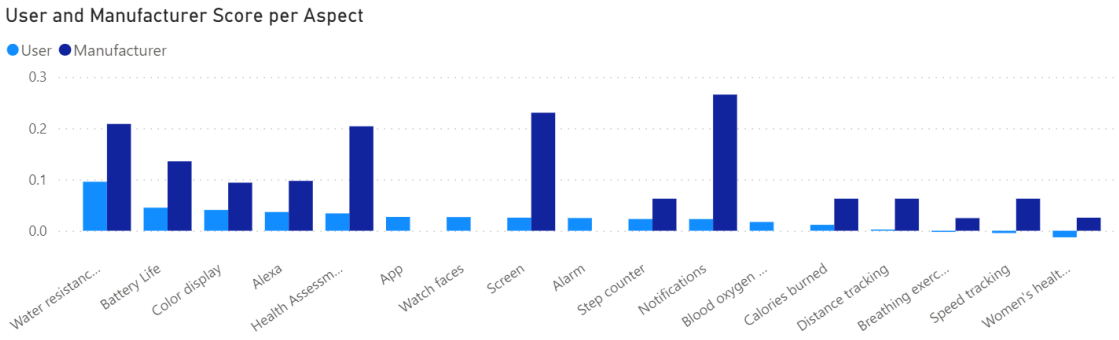

Figura 3.4 Diferencia de *score* de fabricantes y reseñas - Amazfit (SentiWordNet)

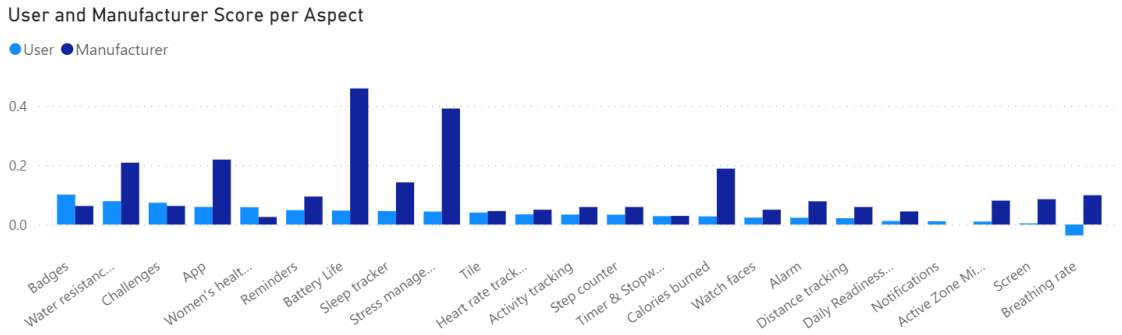

Figura 3.5 Diferencia de *score* de fabricantes y reseñas - Fitbit (SentiWordNet)

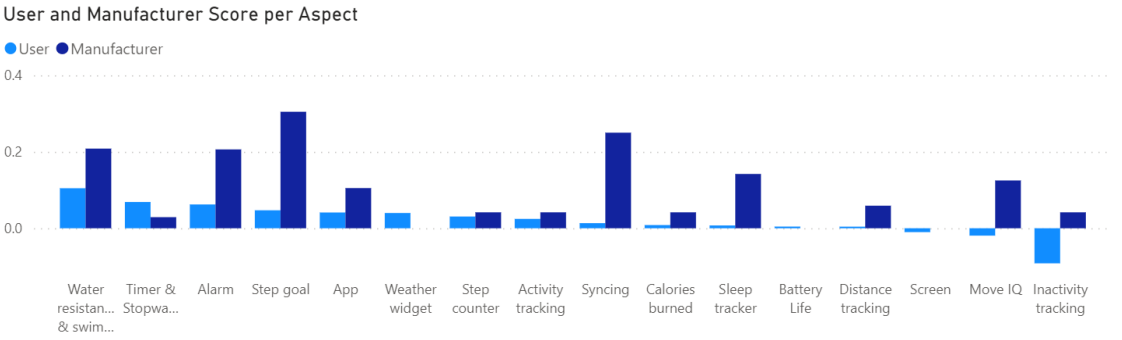

Figura 3.6 Diferencia de *score* de fabricantes y reseñas - Garmin (SentiWordNet)

Para validar estos resultados y el rendimiento del *framework* con SentiWordNet, se obtuvieron cuatro métricas de evaluación. El *accuracy* fue de 0.66, lo que indicó que el modelo predijo correctamente el sentimiento del 66% del total de los datos. Similarmente, el *recall* fue de 0.66, es decir que, se identifica correctamente el sentimiento de los datos en un 66%. Por otro lado, el valor de *precision* fue de 0.55; esto indicó que cuando se clasifica una cláusula con una de las tres etiquetas de sentimiento (positivo, negativo y neutral), la predicción es correcta el 55% de las veces. Estas métricas se resumen en la Tabla 3.9. En conjunto, se determinó que el rendimiento con SentiWordNet es significativo para la predicción del sentimiento de reseñas; no obstante, debido a las limitaciones asociadas a un diccionario de uso general, el rendimiento fue menor al 70%.

Tabla 3.9 Métricas de evaluación de SentiWordNet

Métrica	Valor
Accuracy	0.66
Recall	0.66
Precision	0.55
F1-score	0.60

Para mejorar los resultados obtenidos con SentiWordNet y obtener un mejor rendimiento para el *framework*, se decidió utilizar otra técnica para la fase de análisis de sentimientos. Se usó una red neuronal entrenada con reseñas de productos de Amazon; específicamente el modelo BERT. En la Tabla 3.10 se resumen los resultados del número de cláusulas positivas, negativas y neutrales para Amazfit, en la Tabla 3.11, para Fitbit y en la Tabla 3.12 para Garmin.

Tabla 3.10 Número de cláusulas clasificadas por sentimiento para Amazfit (BERT)

Aspecto	Positivo	Negativo	Neutral
Alarm	46	11	3
Alexa	217	141	28
App	633	448	132
Battery life	399	154	39
Blood oxygen monitoring	62	25	7
Breathing exercises	6	1	0
Calories burned	46	51	13
Color display	3	0	0
Distance tracking	41	16	16
Health Assessment System	35	16	8
Heart Rate Tracking	352	136	59
Notifications	109	40	21
Screen	141	126	52
Sleep tracker	700	256	112
Speed tracking	7	4	0

Step counter	566	260	103
Stress management	86	23	6
Watch faces	95	29	16
Water resistance & swim tracking	60	17	4
Women's health tracking	4	2	0

Tabla 3.11 Número de cláusulas clasificadas por sentimiento para Fitbit (BERT)

Aspecto	Positivo	Negativo	Neutral
Daily Readiness Score	1	0	0
Heart rate tracking	697	215	82
Calories burned	235	47	24
Step counter	1220	275	156
Sleep tracker	1431	382	172
App	881	258	105
Battery life	475	116	30
Reminders	98	16	5
Distance tracking	45	12	8
Notifications	127	112	30
Activity tracking	25	5	3
Water resistance & swim tracking	82	19	9
Challenges	51	5	4
Alarm	41	14	6
Screen	215	277	87
Watch faces	31	20	10
Timer & Stopwatch	29	8	3
Breathing rate	2	0	0
Stress management	16	1	1
Active Zone Minutes	13	3	2
Tile	11	1	3
Women's health tracking	0	1	3
Badges	3	0	0

Breathing exercises	3	1	0

Tabla 3.12 Número de cláusulas clasificadas por sentimiento para Garmin (BERT)

Aspecto	Positivo	Negativo	Neutral
Battery life	180	167	21
Screen	25	60	11
Timer	3	0	0
Move IQ	0	0	1
Alarm	3	1	0
Syncing	51	76	8
Distance traveled	12	5	3
Calories burned	12	1	2
Step counter	130	46	28
Inactivity tracking	1	0	0
Sleep monitoring	58	17	14
Activity tracking	19	8	2
App	59	47	11
Step goal	1	0	0
Water resistance & swim tracking	45	15	5
Weather widget	7	3	1

De manera general, la distribución del número de cláusulas por sentimiento y aspecto en cada uno de los dispositivos —según los resultados obtenidos con el modelo BERT— fueron en la mayoría positivos. Esta distribución se muestra en la Figura 3.7 para Amazfit, Figura 3.8 para Fitbit y Figura 3.9 para Garmin.

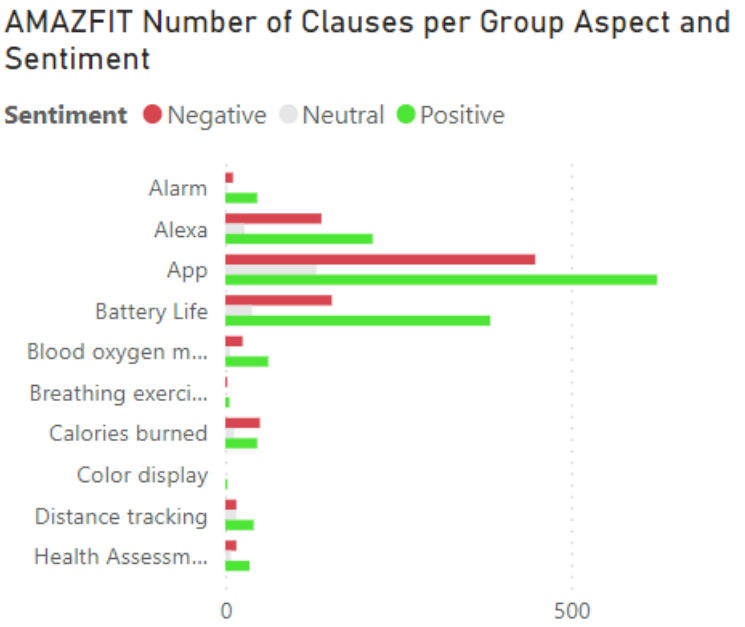

Figura 3.7 Número de cláusulas por sentimiento y aspecto - Amazfit (BERT)

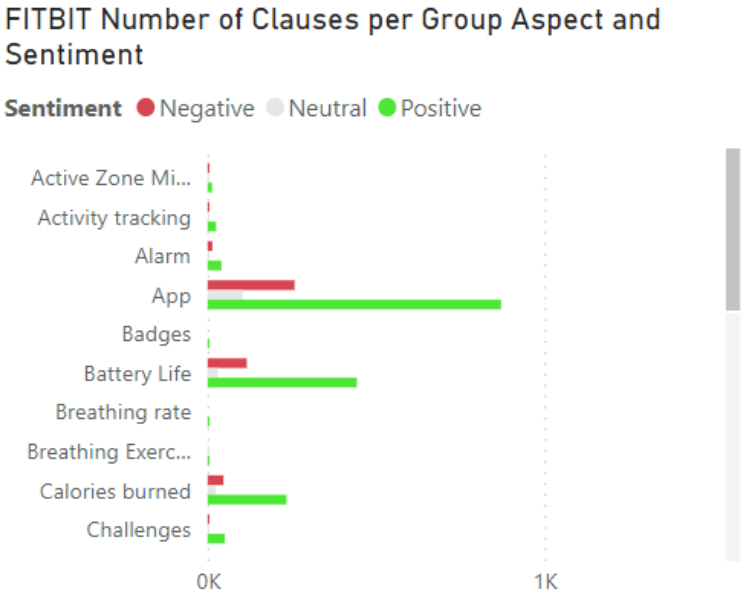

Figura 3.8 Número de cláusulas por sentimiento y aspecto - Fitbit (BERT)

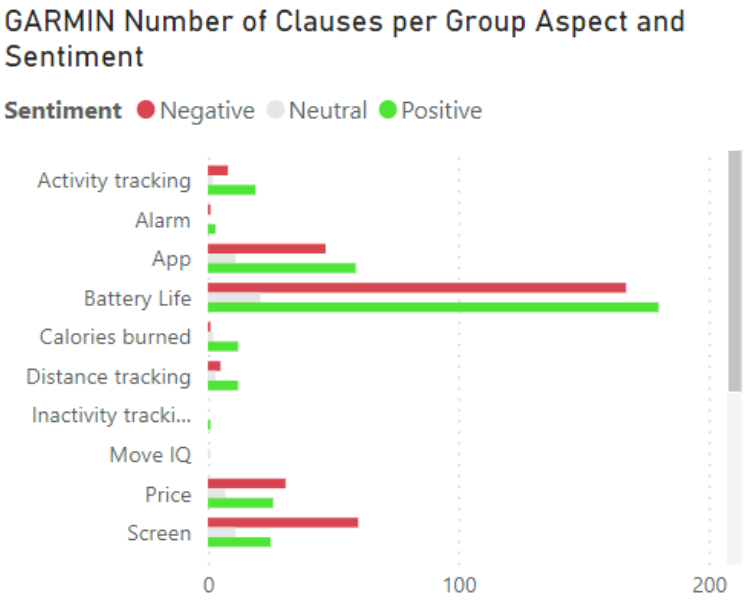

Figura 3.9 Número de cláusulas por sentimiento y aspecto - Garmin (BERT)

En la comparación entre la opinión de los fabricantes y la opinión de los usuarios a partir de las reseñas con el modelo BERT, se obtuvo que para Amazfit el promedio de la diferencia entre usuarios y fabricantes era de 0.629; para Fitbit fue de 0.447; y para Garmin fue de 0.760. Los valores obtenidos se resumen en la Tabla 3.13 para Amazfit, en la Tabla 3.14 para Fitbit y en la Tabla 3.15 para Garmin.

Tabla 3.13 *Score* de fabricantes y de reseñas para Amazfit (BERT)

Aspecto	*Score* Fabricante	*Score* reseñas
Alarm	1	0,53333
Alexa	1	0,17617
App	1	0,13891
Battery Life	1	0,39865
Blood oxygen monitoring	1	0,37234
Breathing exercises	1	0,71429
Calories burned	1	-0,04091

Color display	1	0,66667
Distance tracking	1	0,30822
Health Assessment System	1	0,33051
Notifications	1	0,37647
Screen	1	0,06192
Sleep tracker	1	0,39841
Speed tracking	1	0,59091
Step counter	1	0,30301
Watch faces	1	0,43537
Water resistance & swim tracking	1	0,48148
Women's health tracking	1	0,41667

Tabla 3.14 *Score* de fabricantes y de reseñas para Fitbit (BERT)

Aspecto	***Score* Fabricante**	***Score* reseñas**
Active Zone Minutes	1	0,55556
Activity tracking	1	0,54545
Alarm	1	0,44262
App	1	0,47709
Badges	1	1,00000
Battery Life	1	0,56763
Breathing rate	1	1,00000
Calories burned	1	0,59967
Challenges	1	0,75000
Daily Readiness Score	1	1,00000
Distance tracking	1	0,50000
Heart rate tracking	1	0,47736
Notifications	1	0,05019
Reminders	1	0,62605
Screen	1	-0,10708
Sleep tracker	1	0,51259

Step counter	1	0,55421
Stress management	1	0,83333
Tile	1	0,53333
Timer & Stopwatch	1	0,51250
Watch faces	1	0,18852
Women's health tracking	1	0,53846

Tabla 3.15 *Score* de fabricantes y de reseñas para Garmin (BERT)

Aspecto	***Score* Fabricante**	***Score* reseñas**
Activity tracking	1	0,39655
Alarm	1	0,50000
App	1	0,11111
Battery Life	1	0,04620
Calories burned	1	0,70000
Distance tracking	1	0,27500
Inactivity tracking	1	1,00000
Move IQ	1	0,00000
Screen	1	-0,34375
Sleep tracker	1	0,39888
Step counter	1	0,40441
Step goal	1	-1,00000
Syncing	1	-0,17037
Timer & Stopwatch	1	0,66667
Water resistance & swim tracking	1	0,43077
Weather widget	1	0,40909

Al evaluar el rendimiento del modelo BERT, se obtuvieron los resultados mostrados en la Tabla 3.16. El *accuracy* fue de 0.72, esto indica que el modelo predijo correctamente el sentimiento del 72% del total de los datos. Asimismo, el valor de

precision fue de 0.81; esto indicó que cuando se clasifica una cláusula con una de las tres etiquetas de sentimiento (positivo, negativo y neutral), la predicción es correcta el 81% de las veces. Por lo tanto, en comparación con los resultados obtenidos con SentiWordNet, se mejoró el rendimiento del *framework* para obtener el *score* más ajustado a los datos reales.

Tabla 3.16 Métricas de evaluación del modelo BERT

Métrica	**Valor**
Accuracy	0.72
Recall	0.60
Precision	0.81
F1-score	0.69

El *score* de las características de los dispositivos según el modelo BERT con respecto a los fabricantes es negativo para algunas características. Para Amazfit, en la Figura 3.10 se observa que *calories burned* tiene un *score* menor a cero. Similarmente, para Fitbit, en la Figura 3.11, *screen* tiene un valor negativo. Finalmente, para Garmin, en la Figura 3.12, el *score* de *screen, step goal* y *syncing* es menor a cero.

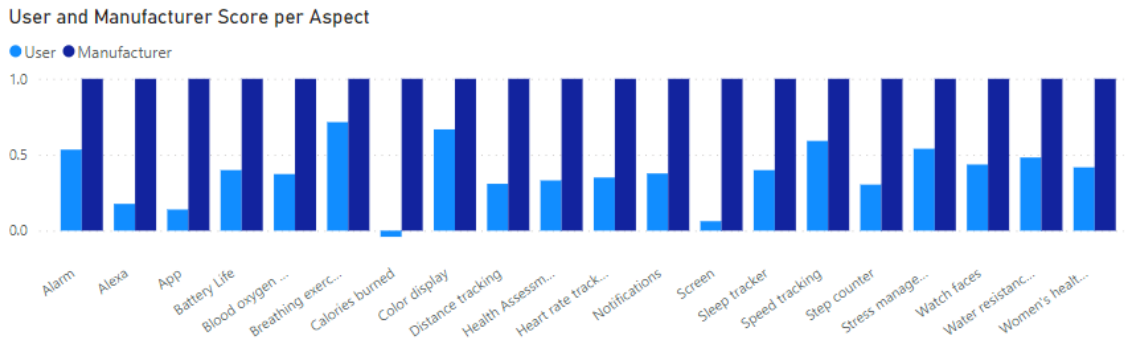

Figura 3.10 Diferencia de *score* de fabricantes y reseñas - Amazfit (BERT)

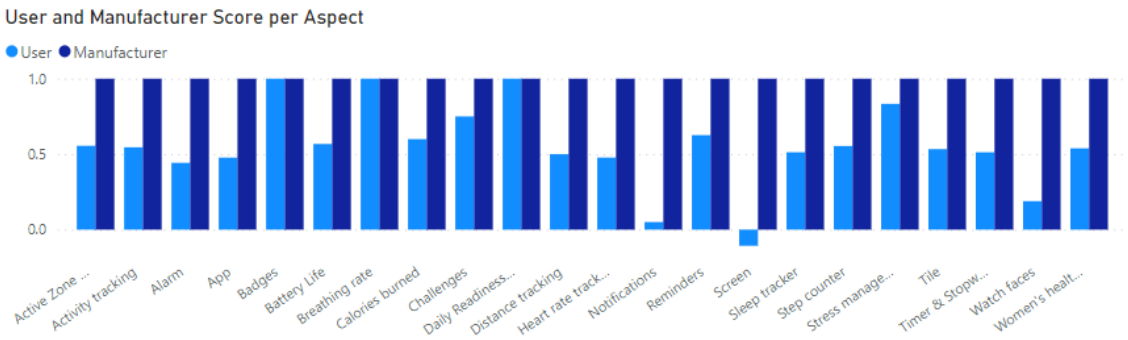

Figura 3.11 Diferencia de *score* de fabricantes y reseñas - Fitbit (BERT)

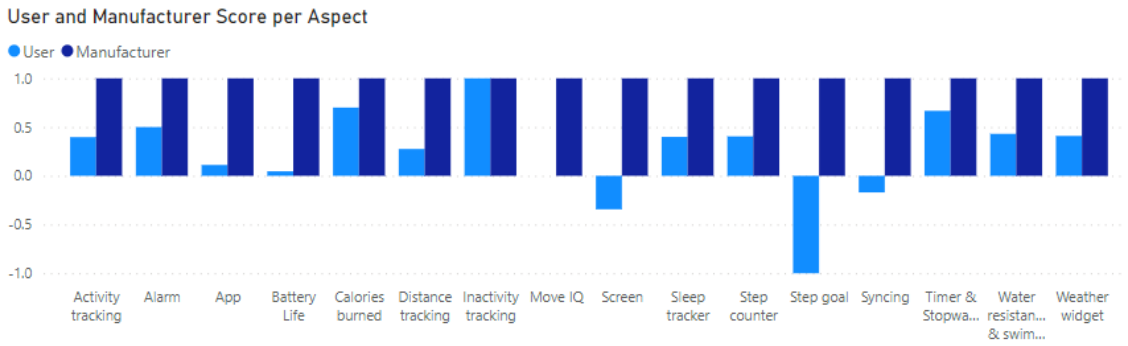

Figura 3.12 Diferencia de *score* de fabricantes y reseñas - Garmin (BERT)

En el siguiente capítulo se presentan las conclusiones extraídas a partir de los resultados presentados en este capítulo, así como las recomendaciones para la mejora y uso del presente proyecto y los trabajos futuros que se pueden derivar a partir de esta propuesta.

CAPÍTULO 4

4. Conclusiones y recomendaciones

En este capítulo se presentan las conclusiones, recomendaciones y trabajos futuros que se pueden extraer a partir de este proyecto. La propuesta de este proyecto se centró en el análisis de sentimientos basado en aspectos para evaluar la efectividad de monitores de actividad. Con el diseño del *framework* propuesto se logró establecer una comparación que permitió evaluar la diferencia entre las descripciones provistas por los fabricantes y las reseñas que los usuarios publican en tiendas en línea. Respecto a trabajos previos, el proyecto complementa el análisis al considerar la opinión de los fabricantes de los MAF.

4.1 Conclusiones

Para el análisis de reseñas, se determinó que una parte importante del diseño del *framework* es la extracción de características. Se concluyó que un procesamiento a nivel de cláusulas es útil para hacer análisis de sentimientos basado en aspectos ya que la calidad de estas influye directamente en los resultados que se obtengan tanto usando técnicas basadas en enfoques léxicos como SentiWordNet o en ML como el modelo BERT. Con los resultados obtenidos con el *framework* se considera que se puede medir la satisfacción de los usuarios y la opinión de los fabricantes para identificar la efectividad de los monitores de actividad física.

Para la fase de análisis de sentimientos, se determinó que el *framework* con el modelo BERT tiene un mejor rendimiento que la versión con un diccionario como SentiWordNet. Con BERT se obtuvo un *precision* alto, de 0.81, en comparación con el valor de 0.55 obtenido con SentiWordNet. De esta manera, se cumplió con el objetivo principal de crear un *framework* que produzca buenos resultados para la medición de la efectividad del rendimiento de monitores de actividad física a partir de la comparación entre la opinión recolectada de los fabricantes y a partir de reseñas escritas por usuarios.

Por otro lado, al comparar los resultados obtenidos se determinó que hay una mayor cantidad de menciones de los aspectos clasificados como “Estáticos” en las reseñas, estos son: alarma, aplicación, batería, calorías quemadas, distancia viajada, monitoreo del sueño, resistencia al agua y la pantalla del MAF. Por lo que corresponden a las características con más alta demanda entre los MAF.

Se determinó que, en general, la efectividad de los monitores de actividad física es sobreestimada por parte de los fabricantes. Los usuarios perciben que el rendimiento de las características ofrecidas en los dispositivos es menor al publicitado. Esta diferencia es mayor para las características de los dispositivos que no son compartidas por todos los dispositivos, clasificadas como aspectos “Dinámicos”.

4.2 Recomendaciones

Tomando en consideración el dispositivo a analizar y sus características específicas, es necesaria una recolección amplia de derivaciones, es decir, formas en las que se hace referencia a las características tanto de parte del fabricante como de los usuarios. De esta manera se encontrará más información sobre los aspectos, especialmente aquellos clasificados como “Dinámicos” y se obtendrá un mayor alcance durante la fase de la extracción de características.

Para el análisis de sentimientos, al usar un enfoque léxico como SentiWordNet se puede ampliar su rendimiento con un corpus específico del dominio, que en este caso corresponde a los monitores de actividad física. De esta manera se complementa el score calculado para las palabras que expresan un sentimiento.

4.3 Trabajos a futuro

La propuesta de este proyecto tiene posibilidades de mejora y propuestas para ampliar su alcance y automatizar algunos de los pasos de las fases de recolección de datos y extracción de características. Dado que el proyecto original se delimitó al análisis de tres dispositivos seleccionados a partir de criterios de selección, una vez consolidados

los módulos del *framework*, se puede extender a un modelo generalizado que sirva como base para un sistema de referencias para analizar reseñas.

Asimismo, para alimentar el modelo y mejorar las predicciones, se puede tomar como base los resultados y la metodología de SentiWordNet para mejorar el etiquetamiento de los *datasets* para el entrenamiento del modelo BERT o de otras técnicas basadas en ML. Por otro lado, el uso de este *framework* puede extenderse para realizar análisis de reseñas de otros tipos de dispositivos además de los MAF.

BIBLIOGRAFÍA

[1] W. R. Thompson, «Worldwide Survey of Fitness Trends for 2022,» *ACSM'S Health & Fitness Journal,* vol. 26, nº 1, pp. 11-20, 2022.

[2] F. Laricchia, «Statista,» 31 de marzo de 2022. [En línea]. Available: https://www.statista.com/topics/1556/wearable-technology/#dossierKeyfigures. [Último acceso: 26 de mayo de 2022].

[3] L. A. Bove, «Increasing Patient Engagement Through the Use of Wearable Technology,» *The Journal for Nurse Practitioners,* vol. 15, nº 8, pp. 535-539, 2019.

[4] N. H. Nguyen, N. T. Hadgraft, M. M. Moore, D. E. Rosenberg, C. Lynch, M. M. Reeves y B. M. Lynch, «A qualitative evaluation of breast cancer survivors' acceptance of and preferences for consumer wearable technology activity trackers,» *Supportive Care in Cancer,* vol. 25, nº 11, p. 3375–3384, 2017.

[5] K. Burford, N. Golaszewski y J. Bartholomew, «"I shy away from them because they are very identifiable": A qualitative study exploring user and non-user's perceptions of wearable activity trackers,» *Digital Health,* vol. 7, 2021.

[6] K.-L. Edward, L. Garvey y M. Aziz Rahman, «Wearable activity trackers and health awareness: Nursing implications,» *International Journal of Nursing Sciences,* vol. 7, nº 2, pp. 179-183, 2020.

[7] Y. Meng, W. Speier, C. Shufelt, S. Joung, J. E Van Eyk, C. N. Bairey Merz, M. Lopez, B. Spiegel y C. W. Arnold, «A Machine Learning Approach to Classifying Self-Reported Health Status in a Cohort of Patients With Heart Disease Using Activity Tracker Data,» *IEEE Journal of Biomedical and Health Informatics,* vol. 24, nº 3, pp. 878-884, 2020.

[8] J. Lee y J. Finkelstein, «Activity Trackers: A Critical Review,» *e-Health - For Continuity of Care - Proceedings of MIE 2014,* vol. 205, pp. 558-562, 2014.

[9] G. Shin, M. H. Jarrahi, Y. Fei, A. Karami, N. Gafinowitz, A. Byun y X. Lu, «Wearable activity trackers, accuracy, adoption, acceptance and health impact: A systematic literature review,» *Journal of Biomedical Informatics,* vol. 93, 2019.

[10] C. Maher, J. Ryan, C. Ambrosi y S. Edney, «Users' experiences of wearable activity trackers: a cross-sectional study,» *BMC Public Health,* vol. 17, nº 1, 2017.

[11] Y. Yiran y S. Srivastava, «Aspect-based Sentiment Analysis on mobile phone reviews with LDA,» *Proceedings of the 2019 4th International Conference on Machine Learning Technologies,* p. 101–105, 2019.

[12] R. M., V. R. Hulipalled, K. Venugopal y L. Patnaik, «Consumer insight mining: Aspect based Twitter opinion mining of mobile phone reviews,» *Applied Soft Computing,* vol. 68, pp. 765-773, 2018.

[13] D. Malekpour Koupaei, T. Song, K. S. Cetin y 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.

[14] H. Samy, A. Helmy y N. Ramadan, «Aspect-based Sentiment Analysis of Mobile Apps Reviews using Class Association Rules and LDA,» *2021 Tenth International Conference on Intelligent Computing and Information Systems (ICICIS),* 2021.

[15] S. Panichella, A. Di Sorbo, E. Guzman, C. A. Visaggio, G. Canfora y H. C. Gall, «How Can I Improve My App? Classifying User Reviews for Software Maintenance and Evolution,» de *2015 IEEE International Conference on Software Maintenance and Evolution (ICSME)*, Bremen, 2015.

[16] H. Issa, A. Shafaee, S. Agne, S. Baumann y A. Dengel, «User-sentiment based Evaluation for Market Fitness Trackers - Evaluation of Fitbit One, Jawbone Up and Nike+ Fuelband based on Amazon.com Customer Reviews,» de *International Conference on Information and Communication Technologies for Ageing Well and e-Health*, Kaiserslautern, Germany, 2015.

[17] H. Sankar y V. Subramaniyaswamy, «Investigating sentiment analysis using machine learning approach,» *2017 International Conference on Intelligent Sustainable Systems (ICISS),* pp. 87-92, 2017.

[18] K. Amarouche, H. Benbrahim y I. Kassou, «Product Opinion Mining for Competitive Intelligence,» *Procedia Computer Science,* vol. 73, pp. 358-365, 2015.

[19] J. Im, T. Song, Y. Lee y J. Kim, «Confirmatory aspect-level opinion mining processes for tourism and hospitality research: a proposal of DiSSBUS,» *Current Issues in Tourism,* vol. 25, nº 12, 2021.

[20] M. Hong y H. Wang, «Research on customer opinion summarization using topic mining and deep neural network,» *Mathematics and Computers in Simulation,* vol. 185, pp. 88-114, 2021.

[21] N. Zainuddin y A. Selamat, «Sentiment Analysis Using Support Vector Machine,» de *IEEE 2014 International Conference on Computer, Communication, and Control Technology (I4CT 2014)*, 2014.

[22] A. S. M. AlQahtani, «Product Sentiment Analysis for Amazon Reviews,» *International Journal of Computer Science & Information Technology (IJCSIT),* vol. 13, nº 3, pp. 15-30, 2021.

[23] M. Geetha y D. Karthika Renuka, «Improving the performance of aspect based sentiment analysis using fine-tuned Bert Base Uncased model,» *International Journal of Intelligent Networks,* vol. 2, pp. 64-69, 2021.

[24] H. Xu, B. Liu, L. Shu y P. Yu, «BERT Post-Training for Review Reading Comprehension and Aspect-based Sentiment Analysis,» *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers),* p. 2324–2335, 2019.

[25] S. Wladislav, Z. Johannes, W. Christian, K. Andre y F. Madjid, «Sentilyzer: Aspect-Oriented Sentiment Analysis of Product Reviews,» *2018 International Conference on Computational Science and Computational Intelligence (CSCI),* pp. 270-273, 2018.

[26] E. Beekman, S. M. Braun, D. Ummels, K. van Vijven, A. Moser y A. J. Beurskens, «Validity, reliability and feasibility of commercially available activity trackers in physical therapy for people with a chronic disease: a study protocol of a mixed methods research,» *Pilot and Feasibility Studies,* vol. 3, nº 1, p. 3, 2017.

[27] A. Rubin y J. Ophoff, «Investigating Adoption Factors of Wearable Technology in Health and Fitness,» *2018 Open Innovations Conference (OI),* 2018.

[28] E. C. Nelson, T. Verhagen y M. L. Noordzij, «Health empowerment through activity trackers: An empirical smart wristband study,» *Computers in Human Behavior,* vol. 62, pp. 364-374, 2016.

[29] V. M. Kercher, K. Kercher, T. Bennion, P. Levy, C. Alexander, P. C. Amaral, Y.-M. Li, J. Han, Y. Liu, R. Wang, H.-Y. Huang, B.-H. Gao, A. Batrakoulis, L. F. J. Gómez Chávez, J. L. Haro, A. R. P. Zavalza, L. E. A. Rodríguez, O. L. Veiga, M. Valcarce-Torrente y A. Romero-Caballero, «2022 Fitness Trends from Around the Globe,» *ACSM'S Health & Fitness Journal,* vol. 26, nº 1, pp. 21-37, 2022.

[30] J. Mason, F. Brundisini, S. Hill, D. Kumar y T. Rader, «2022 Health Technology Trends to Watch: Top 10 List,» *Canadian Journal of Health Technologies,* vol. 2, nº 3, p. 10, 2022.

[31] V. Kercher, Y. Feito y B. Yates, «Regional Comparisons: The Worldwide Survey of Fitness Trends,» *ACSM'S Health & Fitness Journal,* vol. 23, nº 6, pp. 41-48, 2019.

[32] KANTAR, «Global Wearable Tech Trends 2020,» 2020.

[33] A. Shafaee, H. Issa, S. Agne, S. Baumann y A. Dengel, «Aspect-Based Sentiment Analysis of Amazon Reviews for Fitness Tracking Devices,» *Lecture Notes in Computer Science,* pp. 50-61, 2014.

[34] D. C. Cavalcanti, R. B. C. Prudêncio, S. S. Pradhan, J. Y. Shah y R. S. Pietrobon, «Good to be Bad? Distinguishing between Positive and Negative Citations in Scientific Impact,» *2011 IEEE 23rd International Conference on Tools with Artificial Intelligence,* 2011.

[35] «nlptown/bert-base-multilingual-uncased-sentiment,» [En línea]. Available: https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment. [Último acceso: 29 de agosto de 2022].

[36] «LiYuan/amazon-review-sentiment-analysis,» [En línea]. Available: https://huggingface.co/LiYuan/amazon-review-sentiment-analysis. [Último acceso: 29 de agosto de 2022].

[37] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng y C. Potts, «Learning Word Vectors for Sentiment Analysis,» *Proceedings of the 49th Annual Meeting of the*

Association for Computational Linguistics: Human Language Technologies, pp. 142-150, 2011.