Extracción y clasificación automática de opiniones relacionadas con el proceso de evaluación de usabilidad mediante la técnica think-aloud en productos software

Cesar Bahena-Ríos
Departamento de Ciencias Computacionales
TecNM/CENIDET
Cuernavaca, México
m21ce054@cenidet.tecnm.mx

Gabriel González-Serna
Departamento de Ciencias Computacionales
TecNM/CENIDET
Cuernavaca, México
gabriel.gs@cenidet.tecnm.mx

Noé Castro-Sánchez
Departamento de Ciencias Computacionales
TecNM/CENIDET
Cuernavaca, México
noe.cs@cenidet.tecnm.mx

Nimrod González-Franco
Departamento de Ciencias Computacionales
TecNM/CENIDET
Cuernavaca, México
nimrod.gf@cenidet.tecnm.mx

Máximo López-Sánchez
Departamento de Ciencias Computacionales
TecNM/CENIDET
Cuernavaca, México
maximo.ls@cenidet.tecnm.mx

Juan Gómez-Ramírez
Departamento de Ciencias Básicas
TecNM/ITAcapulco
Acapulco, México
juan.gr@acapulco.tecnm.mx

***Resumen*— Este trabajo de investigación presenta un enfoque innovador para la evaluación de usabilidad y de la experiencia del usuario (UX) en productos software mediante la técnica think-aloud protocol (TAP) y la aplicación de técnicas de Procesamiento de Lenguaje Natural (PLN), minería de opinión y análisis de emociones. Un grupo de usuarios evaluaron prototipos de productos software aplicando la técnica TAP, la cual requiere que el usuario verbalice en voz alta sus pensamientos mientras interactúa con la interfaz digital de un prototipo software, de todas las interacciones se registró audio y video. Los audios se procesan para transcribir el texto (STT), se le aplican técnicas de PLN para la extracción de opiniones. Posteriormente se clasifica la polaridad de cada opinión, como positiva o negativa, finalmente cada opinión se asigna a una de tres categorías para obtener únicamente las opiniones relevantes para el evaluador de usabilidad y de la UX. Los resultados muestran la efectividad de este enfoque en el proceso de evaluación de la usabilidad y la UX en productos software, lo que tendrá un impacto significativo en la detección de problemas de usabilidad en etapas tempranas del proceso de desarrollo de software.**

Keywords*— *Análisis de emociones, PLN, think-aloud, usabilidad, U.

INTRODUCCIÓN

La evaluación de usabilidad en productos digitales es un proceso crucial en el desarrollo de software, sobre todo en etapas temprana. Una característica de este proceso es que puede resultar costoso en términos de tiempo y recursos, en particular en lo que respecta a la extracción de los comentarios de los usuarios, el cual es un proceso manual. En esta investigación presentamos una solución innovadora que permite reducir la complejidad en el proceso de extracción de opiniones de usabilidad; combina la técnica think-aloud protocol (TAP) con el Procesamiento de Lenguaje Natural (PLN), minería de opinión y análisis de emociones. El enfoque propuesto permite la extracción y clasificación automática de opiniones relevantes relacionadas con problemas de usabilidad. Para validar el método se utilizaron prototipos de media fidelidad que fueron evaluados por un grupo de diez usuarios utilizando la técnica TAP, que implica verbalizar sus pensamientos y emociones mientras exploran el prototipo. Posteriormente se procesa de manera individual cada audio para transcribir el texto de forma automática (STT), el texto de salida se procesa con herramientas de PLN para extraer las opiniones, se aplican técnicas de análisis de emociones para clasificar la polaridad de cada opinión. Para mejorar la efectividad del enfoque propuesto, se utilizó una bolsa de léxico emocional como filtro para obtener solo las opiniones relevantes para el evaluador de usabilidad. Los resultados de esta investigación muestran la efectividad del método en la detección temprana de problemas de diseño e interacción en productos software, lo que podría tener un impacto significativo en el proceso de desarrollo de software para mejorar la experiencia del usuario final.

II. MÉTODO

A. Materiales y métodos

Para el desarrollo de las pruebas realizadas en esta investigación se describen los detalles de los materiales y métodos utilizados, así como el porqué de cada uno.

B. Prototipo software

Para el diseño de prototipo de media fidelidad se utilizó el software Balsamiq (Balsamiq. Rapid, effective and fun wireframing software, n.d.) (Fig. 1) el cual cuenta con aspectos básicos como hipervínculos, botones, cajas de selección, entre otros elementos de interacción, este prototipo se utilizó en el proceso de evaluación de usabilidad y de UX.

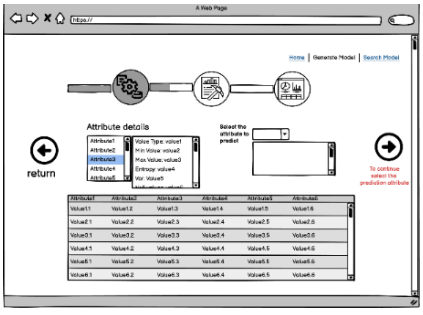

Fig. 1. Página prototipo

C. Protocolo think-aloud

Es un protocolo de análisis cualitativo, existen dos variantes Think-Aloud Protocol (TAP) y Co-Operative Evaluation Think-Aloud (CETA), para esta investigación se

aplicó el TAP, es la variante más común, consiste en pedirle al usuario que exprese en voz alta lo que está pensando mientras interactúa con el producto o prototipo software. Durante la interacción, el usuario es observado, se graba audio y video mientras realiza una serie de tareas en el prototipo.

A medida que el usuario realiza las tareas, debe verbalizar sus pensamientos, lo que permite al evaluador comprender la forma en que el usuario interactúa con el prototipo y las posibles dificultades de usabilidad que enfrenta. El objetivo de la técnica TAP es obtener información detallada sobre la experiencia del usuario mientras interactua con un producto o prototipo software, incluyendo las dificultades, los errores, las expectativas y las impresiones generales, positivas y negativas. Posteriormente el evaluador escucha cada uno de los audios para extraer manualmente las opiniones. Estas pruebas se caracterizan por su fácil implementación ya que no requiere gran esfuerzo monetario o logístico, al no necesitar equipos costosos. Una de las ventajas destacables de esta prueba son las reacciones espontaneas de los participantes al no tener mucho tiempo para pensar y así verbalizar su opinión sincera.

Sin embargo, hay dos problemas asociados a la técnica TAP, l) el sesgo de aquiescencia, que es la tendencia de las personas a estar de acuerdo en todo y 2) el sesgo de deseabilidad social, que es el deseo de informar puntos de vista que otros considerarán favorablemente.

D. Muestra de datos

Para conformar el conjunto de datos para validad el método propuesto, se conformó un grupo de diez usuarios que evaluaron un prototipo software de media fidelidad mediante exploración e interacción, aplicando la técnica TAP, los usuarios realizaron tareas simples como selección de elementos, simulación de carga y descarga de archivos, entre otras acciones, en prototipos de media fidelidad se simula la interacción y navegación, carecen de diseño visual y funcionalidad real.

E. Pysentimiento

La extracción de opiniones provenientes de diversos textos ha sido de gran interés en los últimos años ya que permite obtener información valiosa sobre diversas temáticas, sin embargo, un problema que se distingue en este ámbito es la falta de herramientas para minería de opinión que sean multilenguaje o que trabajen específicamente con idioma español, por ello, se desarrolló Pysentimiento (Pérez et al., 2021), un conjunto de herramientas para el análisis de sentimientos y procesamiento del lenguaje natural, con la ventaja de procesar directamente el problema antes mencionado, permitiendo trabajar en idioma español, además de ser una librería de código abierto para permitir a los investigadores acceder de manera más fácil y controlada.

Con el uso de esta librería se implementó el módulo para clasificar automáticamente la polaridad de los comentarios extraídos del texto que se trascribió de manera automática de los usuarios participantes.

Las métricas de evaluación que se aplicaron fueron Precision (1), Recall (2), F1-Score (3) y dos más, Micro-F1 (4) y Macro-F1 (5), y son obtenidas a través de la matriz de confusión (Fig. 2)

$$Precision = \frac{TP}{TP + FP} \quad (1)$$

$$Recall = \frac{TP}{TP + FP} \quad (2)$$

$$F1 - Score = \frac{Precision * Recall}{Precision + Recall} \quad (3)$$

$$Micro\ F1 = \frac{2 * (Precision * Recall)}{Precision * Recall} \quad (4)$$

$$Macro\ F1 = \frac{Sum(F1\ scores)}{Number\ of\ classes} \quad (5)$$

Fig. 2 Matriz de confusión (Arce, 2019)

Los resultados obtenidos con la herramienta Pysentimiento se presentan en la Tabla 1, se destacan los obtenidos en el proceso de análisis de sentimientos en español con las diferentes bases de datos utilizadas, siendo *Beto* la que obtuvo los mejores resultados 0.672 y 0.667 para Micro-F1 y Macro-F1 respectivamente.

TABLA. 1. MÉTRICAS DE EVALUACIÓN DE PYSENTIMIENTO

	Sentiment		Emotion	
Model	Micro f1	Macro f1	Micro f1	Macro f1
distilbert	0.649	0.642	0.503	0.383
mbert	0.645	0.643	0.516	0.394
en bert	0.686	0.684	0.559	0.439
roberta	0.686	0.684	0.563	0.445
bertweet	**0.697**	**0.696**	**0.584**	**0.476**
distilbert	0.602	0.599	0.600	0.463
es mbert	0.609	0.604	0.610	0.474
beto	**0.672**	**0.667**	**0.688**	**0.548**

F. Léxico emocional

La minería de opinión y el análisis de sentimientos se ha convertido en una rama de interés para el ámbito del procesamiento del lenguaje natural, para lograr tareas que busquen estos objetivos se han implementado diversos recursos léxicos, que en su mayoría son para el idioma inglés, muy limitados para el español o incluso recursos inexistentes para algunos casos. Para tratar este problema, algunos autores desarrollaron trabajos que conforman las principales contribuciones, en esta investigación se enfocó principalmente en dos, en (Castro-Sánchez, 2015) se conformó una bolsa de léxico afectivo con diversos recursos, que incluyen obras basadas en teorías psicológicas para identificar palabras asociadas a emociones, por otra parte, en (Diaz Rangel, 2014) se presenta un método para la creación de diccionarios etiquetados, conformando un léxico emocional etiquetado con seis emociones relacionadas.

Derivado de los trabajos identificados en la revisión de la literatura, se procedió a descargar los corpus de palabras para

analizarlos, en este proceso encontramos que el primer corpus contenía 3344 palabras emocionales y el segundo 2036, cada uno con una estructura similar y orientada al trabajo que presentaron los autores, se procedió a extraer solo las palabras para conformar un solo corpus de léxico emocional (LE) verificando que no hubiera palabras duplicadas y ordenándolas en orden alfabético, así mismo, se comprobó que todas las palabra estuvieran lematizadas, con el objetivo de tomar en cuenta las variaciones con las que cuentan las palabras en sí, por ejemplo, "gusta" y "gustan" son variaciones del lema "gustar", el resultado fue un corpus de LE conformado por 4400 palabras, la (Fig. 3) muestra un fragmento de dicho corpus.

```
listaEmocionalFinal=["abajamiento", "abajo ", "abandonado", "
"abochornar", "abofetear", "abolir", "abollado", "abolladura
", "abrasivo", "abrazar", "abrazo", "abrigado", "abrigar", "
"abundante", "abundantemente", "abundar", "aburrido", "aburri
"accidente", "acción", "acechar", "acecho", "aceptable", "ac
"acogedor", "acogedoramente", "acoger", "acojonante", "acomet
"acongojado", "acongojar", "aconsejable", "aconsejar", "acopi
"acuchillar", "acuciar", "acucioso", "acuerdo", "acuitadament
"adherente", "adhesión", "adicción", "adicto", "adivinanza",
```

Fig. 3 Fragmento del léxico emocional

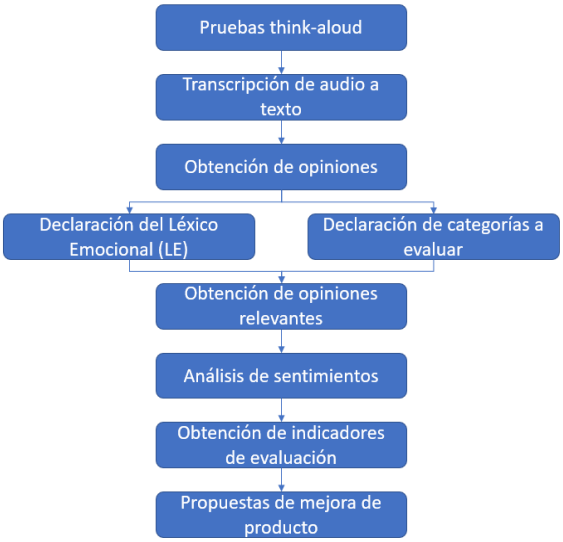

Fig. 4 Diagrama de flujo del método completo

G. Whisper

Whisper (Radford et al., 2022) es un modelo de reconocimiento del habla para propósitos diversos, está entrenado con un conjunto de datos de audio de diversa procedencia, también se puede utilizar para el reconocimiento de voz multilingüe, traducción de voz e identificación de idiomas, en la revisión de la literatura fue la herramienta con los mejores resultados en el proceso de transcribir un audio a texto (STT), llegando incluso a tener tan solo una tasa de error de palabras (WER) del 3% para el español, métrica aplicada para evaluar sistemas de reconocimiento automático del habla (Chen, 2021), lo cual, mejora incluso el error humano que ronda el 4% aproximadamente.

H. Categorías de evaluación

Se llevó a cabo la clasificación en tres categorías de evaluación de aspectos encontrados en pruebas TAP realizadas con anterioridad a este trabajo, para ello, se dio lectura a todos los textos traducidos automáticamente por Whisper, para extraer todos los atributos que los usuarios mencionaron en cada una de las pruebas, se identificaron una diversidad de palabras relacionadas con tres categorías principales, 1) elementos visuales, 2) atributos de diseño, y 3) atributos de funcionalidad, sin embargo dependiendo del producto que se esté evaluando o los objetivos específicos de cada investigador podrían modificarse y obtener variantes más específicas o generales, por ejemplo, categorías como calidad del sistema, calidad de la información y calidad de la interfaz.

I. Diagrama general del método

La fig. 4 muestra el diagrama de flujo del método desarrollado en esta investigación para el análisis de opiniones, se pueden clasificar en tres grupos más generales, la adquisición de opiniones relevantes que va desde las pruebas TAP hasta la obtención final de una lista completa de opiniones que se clasificaron como relevantes, posteriormente está el proceso de cálculo estadístico del sistema de indicadores propuesto, compuesto por el análisis de sentimientos o polarización de opinión y el cálculo de los indicadores, finalmente las propuestas de mejora de los productos evaluados.

J. Opiniones relevantes

La adquisición de opiniones relevantes es el proceso principal del método, nos centramos en obtener solo opiniones que fueran relevantes para el equipo de diseño del producto digital que se están evaluando, ya que, dentro del corpus de opiniones se encuentran muchas frases que no presentan ninguna carga emocional y tampoco van dirigidas hacia ninguna parte de interés del producto, frases como: "doy clic aquí", "este es el principio", "ok", son ejemplos de éstas. Por ello, se implementó un filtro en el cual todas y cada una de las opiniones deberan incluir dos tipos de palabras específicas, por una parte, alguna palabra del léxico emocional previamente establecido y por otra alguna palabra relaciona con las categorías a evaluar, de esta manera se asegura que casi cualquier frase que pase el filtro es de utilidad y puede considerarse una opinión relevante, por ejemplo, "Me gusta mucho el botón home" es una frase con polaridad positiva ya que expresa una opinión favorable hacia un componente que en este caso cae dentro de la categoría de elementos visuales, la palabra "gusta" denota una emoción que al ser lematizada a "gustar" se encuentra dentro del léxico emocional definido, dicho esto, la frase cumpliría con el filtro propuesto y se guardaría como una opinión relevante.

K. *Indicadores de evaluación*

En esta investigación se propone un sistema de indicadores para la evaluación de los productos digitales en los cuales se hace uso del análisis de sentimientos y la asociación con cada atributo de categoría que se está evaluando, las cuales pueden modificarse para evaluar más o menos aspectos, incluso modificarse por completo dependiendo fel objetivo del investigador, de esta forma se presentan tres indicadores, 1) Satisfacción de usuario (SU), 2) Atención de usuario (AU) y 3) Prioridad de mejora (PM), cada uno de estos indicadores es calculado para cada categoría a evaluar.

La satisfacción de usuario (SU) de cada categoría es calculada por la ecuación 6, donde ComenPos, ComenNeu y ComenNeg representan el total de comentarios positivos, neutrales y negativos respectivamente clasificados como tal por la herramienta de análisis. Un resultado de satisfacción

alto o cercano a 1 significa que los usuarios están generalmente satisfechos con los atributos de esa categoría.

$$SU = \frac{ComenPos + (0.5 * ComenNeu)}{ComenPos + ComenNeu + ComenNeg} \quad (6)$$

Para entender el siguiente indicador, es necesario comentar que cuanto más cercano a 1 significará que los usuarios suelen mencionar mucho en su comentarios atributos relacionados con una categoría, lo que significa que se preocupan mucho por ésta, lo cual no implica que la preocupación sea mala o buena, dicho esto, la atención del usuario es calculada por la ecuación 7, donde se añade un factor, la sumatoria de los comentarios relevantes total, los cuales, como se mencionó anteriormente, son los que pasaron el filtro propuesto para aportar información relevante.

$$AU = \frac{ComenPos + ComenNeu + ComenNeg}{ComenReleTotal} \quad (7)$$

Finalmente, el indicador de prioridad de mejora, que identifica la prioridad en la que debe ser atendido y mejorado algún concepto evaluado, que en este caso, son las categorías de evaluación establecidas, éste es calculado por la ecuación 8, es importante resaltar que cuanto menor sea la satisfacción del usuario y mayor sea la atención dada, resultará en una prioridad de mejora más alta, por ejemplo, en una prueba realizada, el 50% de los usuarios de un total de 100 hicisron comentarios relacionados con la categoría de diseño, y se mostró una satisfacción del 90%, por el contrario, un número reducido de usuarios 20% comentaron acerca de la funcionalidad, con una satisfacción del 10%, realizando el cálculo del indicador de prioridad de mejora por categoría tenemos un resultado de 0.05 para la primera y 0.18 para la segunda, por lo que el valor más alto indicaría que la funcionalidad es algo de lo cual el equipo de diseño debería revisar y corregir, ya que, a pesar de que un número reducido de usuarios opinaron sobre este atributo, por lo general se tiene una insatisfacción con la misma. Con esto se demuestra que este sistema de indicadores es efectivo.

$$PM = (1 - SU) * AU \quad (8)$$

Una vez obtenidos los indicadores de evaluación, es necesario que el equipo de diseño del producto software realice un análisis apoyado de estos resultados para la mejor estrategia que se pueda plantear, tomando en cuenta aspectos que cada equipo decidirá, por ejemplo, el costo humano de mejora de cada categoría.

III. Resultados, Discusión y Conclusiones

En esta sección se lleva a cabo la experimentación del método con un caso de estudio del prototipo software, las secciones se listan desde la adquisición de las opiniones generales hasta los resultados obtenidos.

A. Pruebas think-aloud Protocol (TAP)

Se registró la participación de diez usuarios que aplicaron la técnica think-aloud Protocol (TAP), los participantes evaluaron un prototipo de media fidelidad, se utilizó una herramienta denominada QUXBox, para grabar video y audio de las interacciones de los usuarios participantes, se obtuvieron diez archivos de audio con toda la verbalización que los usuarios realizaron, al mismo tiempo, se grabó video del rostro de los usuarios y se registraron datos de movimiento ocular (eye-traking) para su posterior análisis. Como se mencionó anteriormente, no fue necesario adquirir equipo de cómputo especializado en esta investigación, se adaptó un aula con las condiciones ergonómicas adecuadas para realizar las pruebas y se utilizó un equipo de cómputo con un micrófono direccional como se muestra en la (Fig 5).

Fig. 5 Setup de pruebas think-aloud + QUXBox

B. Transcripción de audios y obtención de opiniones

Una vez que finalizaron las pruebas, se utilizó la herramienta Whisper (Radford et al., 2022) para realizar el proceso de transcribir los audios a texto (STT), en la (Fig. 6) se muestra el resultado de una transcripción de audio.

Audio1

tengo que hablar ok, los logos no van donde corresponden el título pues está monto nada encima de los logos en cuanto a los links de home y todo, pues están en el lugar correcto la descripción corta o resumen, no sé cómo lo quieren decir pues yo creo que no, está un poco arriba y el texto, los encabezados son muy chiquitos y el texto no se ve nada en cuanto a los logos, se requieren que sean un poco más grandes, más vistos se requiere que la página sea más vistosa con colores y cosas así de ese tipo pues no sé qué más necesidad de indicar si aquí en el título no tiene nada las letras no se ven, en los logos yo creo que están si extendemos el logo un poquito más hacia la izquierda y centrado que no esté aquí todo, tiene mucho espacio tiene mucho espacio en esta parte, en esta y en esta que debería de utilizar el título aquí en alto, que se vea negro o sea, que se distinga, que tenga algo de colores, estas cosas no sé qué sea un parágrafo, es una persona en

Fig. 6 Fragmento de texto transcrito

Un problema común en las herramientas de STT se relaciona con los signos de puntuación(coma, punto y coma, punto, etc.), normalmente presentan fallas en esta tarea, al poner u omitir signos, debido a diversos factores como las pausas que hace el hablante, el uso de muletillas, etc., sin embargo, Whisper a diferencia de otras herramientas, demostró que es capaz de escribir puntuación casi siempre en el lugar correcto, por ello, es que se decidió realizar la separación del texto de salida a través de puntos y comas, para extraer comentarios pequeños y realizar el proceso para identificar la polaridad correspondiente de cada uno de los comentarios de forma manual y automática. Cómo resultado, se obtuvieron al final un total de 1062 comentarios extraídos de un total de diez archivos de audio.

C. Algoritmo de minería de opinión

El método se implementó en Python, incluyendo herramientas que permitieron llevar a cabo el propósito principal de esta investigación, es decir, la obtención de los indicadores de evaluación propuestos.

Este algoritmo sigue de forma general la siguiente estructura:

1. Definición de categorías a evaluar con su lista de atributos correspondientes, como se mencionó, para esta investigación se establecieron tres categorías, elementos visuales, diseño general y funcionalidad,

a su vez se declararon variables acumulativas de comentarios en sus tres polaridades posibles (Positivo, neutral y negativo), estas variables sirven para calcular los indicadores de evaluación al terminar todo el proceso.

2. El siguiente paso es crear un ciclo capaz de recorrer todos los comentarios identificados, estos son lematizados y entran al filtro de identificación de comentarios relevantes.
3. Una vez que un comentario pasa el filtro, se procede a identificar la polaridad y la categoría a la que pertenece para sumarlo a las variables acumulativas
4. Finalmente, al terminar el ciclo, se realiza el cálculo de los indicadores de evaluación

Los resultados obtenidos muestran una reducción significativa en los comentarios extraídos de las pruebas, casi el 90% al obtener solo 121 comentarios relevantes del total de 1062, sin embargo, todos aportan información relevante para el evaluador, ya que, se identificó que en las pruebas, el 82% de los comentarios eran neutrales, es decir, 832 del total, los cuales, en su mayoría no aportaban datos relevantes, esto se explica debido a que los participantes suelen narrar muchas de las acciones que realizan o dan lectura a lo que ven en pantalla, entre otros aspectos, en lugar de verbalizar comentarios con algún tipo de polaridad relacionados con criterios de usabilidad o de la experiencia del usuario.

Como resultado final, se obtuvieron los resultados mostradas en la figura 7, donde se observa que la satisfacción del usuario es medianamente aceptable en todas las categorías al tener cifras cercanas a 0.5 recordando que la escala es de 0 a 1, también se observa que la satisfacción más alta obtenida es respecto a la categoría de Diseño general con 0.517, seguido de resultados casi iguales relacionados con Elementos visuales y con la Funcionalidad, 0.487 y 0.484 respectivamente.

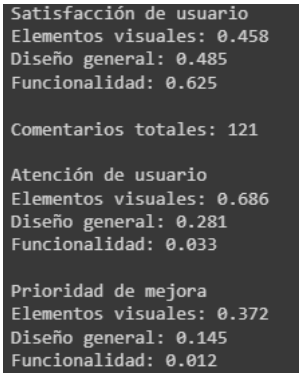

Fig. 7 Indicadores de evaluación

Por otra parte, con respecto a la métrica de atención del usuario, se observa que en definitiva hubo una categoría que destacó sobre el resto, los elementos visuales recibieron la mayoría de los comentarios de los usuarios, concretamente el 71%, con respecto al Diseño general y Funcionalidad solo el 14% y 15% respectivamente. Finalmente, la prioridad de mejora muestra que a pesar de que la categoría de Funcionalidad fue la categoría con menor nivel de satisfacción, sin embargo, no es la que se considera que deba mejorarse prioritariamente, en contraste a esto, los elementos visuales fueron atributos que más llamaron la atención de los usuarios al realizar las pruebas y al obtener una satisfacción de usuario regular, el cálculo de prioridad de mejora indica la categoría que esta categoría debería mejorarse antes que las demás.

D. CONCLUSIONES

Hoy en día el diseño de productos digitales y mejora de estos, sigue adoptando métodos tradicionales como el uso de encuestas auto informadas y entrevistas a los usuarios, herramientas que en su mayoría son tediosas y tardadas, en esta investigación, se desarrolló un método novedoso que utiliza tecnologías de procesamiento de lenguaje natural y minería de opinión realizar un análisis de usabilidad en un menor tiempo, comparado con los métodos tradicionales, mediante una técnica que permite a los usuarios expresarse de manera natural, basada en el Protocolo Think-Aloud, la cual, permite registrar comentarios espontáneos sobre los productos que evalúan, sin embargo, también se tienen algunas limitaciones, que pueden mejorarse en un futuro, por ejemplo, no se toma en cuenta el costo que conllevaría la mejora de alguna categoría específica, por lo que en un trabajo futuro sería agregar más factores al cálculo de los indicadores de evaluación, dependiendo de las necesidades y criterios de cada uno, de igual forma, podría considerarse una metodología más específica para la protocolo think-aloud, filtrando aún más los comentarios que emiten los usuarios al realizar la evaluación y así obtener desde un principio opiniones más útiles para los evaluadores de usabilidad y/o de la UX.

E. DISCUSIÓN

En esta investigación se logró combinar técnicas de procesamiento de lenguaje natural (PLN), transcripción automática de voz y análisis de sentimientos para desarrollar un método para extraer automáticamente opiniones de usuarios desde el enfoque de la evaluación de la experiencia del usuario (UX) mediante el protocolo think-aloud, para posteriormente ser procesadas y obtener un sistema de indicadores que refleje con mayor precisión la evaluación que los usuarios realizan. Además, se logró resolver un problema relacionado con el protocolo think-aloud, relacionado con la narrativa y exceso de comentarios sin carga emocional que emiten los usuarios al hablar, utilizando el léxico emocional y palabras relacionadas a categorías específicas para obtener solo comentarios de utilidad, a pesar de esto, podrían implementar ciertas pautas en las pruebas para mejorar la verbalización de los usuarios, manteniendo el enfoque principal de este tipo de pruebas, por ejemplo, realizar sesiones de práctica que ayude a los usuarios a familiarizarse con sus pensamientos y la prueba en general antes de realizar la formal.

REFERENCES

[1] Arce, J. I. B. (2019, julio 26). La matriz de confusión y sus métricas. Juan Barrios. https://www.juanbarrios.com/la-matriz-de-confusion-y-sus-metricas/

[2] Balsamiq. Rapid, effective and fun wireframing software. (s/f). Balsamiq.com. Recuperado el 17 de abril de 2023, de https://balsamiq.com/

[3] Budiharto, W., & Meiliana, M. (2018). Prediction and analysis of Indonesia Presidential election from Twitter using sentiment analysis. Journal of Big Data, 5(1). https://doi.org/10.1186/s40537-018-0164-1

[4] Castro-Sánchez, N. A., Baca-Gómez, Y. R., & Martínez, A. (2015). Development of affective lexicon for Spanish with Mexican slang expressions. Research in Computing Science, 100(1), 9–18. https://doi.org/10.13053/rcs-100-1-1

[5] Chen, H. (2021, enero 20). Does Word Error Rate matter? SmartAction. https://smartaction.ai/blog/does-word-error-rate-matter/

[6] Díaz Rangel, I., Instituto Politécnico Nacional, Sidorov, G., & Suárez Guerra, S. (2014). Creación y evaluación de un diccionario marcado con emociones y ponderado para el español. ONOMAZEIN, 29, 31–46. https://doi.org/10.7764/onomazein.29.5

[7] Gottipati, S., Shankararaman, V., & Lin, J. R. (2018). Text analytics approach to extract course improvement suggestions from students' feedback. Research and Practice in Technology Enhanced Learning, 13(1), 6. https://doi.org/10.1186/s41039-018-0073-0

[8] Le, N.-B.-V., & Huh, J.-H. (2021). Applying sentiment product reviews and visualization for BI systems in Vietnamese E-commerce website: Focusing on Vietnamese context. Electronics, 10(20), 2481. https://doi.org/10.3390/electronics10202481

[9] Namugera, F., Wesonga, R., & Jehopio, P. (2019). Text mining and determinants of sentiments: Twitter social media usage by traditional media houses in Uganda. Computational Social Networks, 6(1). https://doi.org/10.1186/s40649-019-0063-4

[10] Pérez, J. M., Giudici, J. C., & Luque, F. (2021). Pysentimiento: A Python Toolkit for Sentiment Analysis and SocialNLP tasks. En arXiv [cs.CL]. http://arxiv.org/abs/2106.09462

[11] Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust speech recognition via large-scale weak supervision. En arXiv [eess.AS]. http://arxiv.org/abs/2212.04356

[12] Sasikala, P., & Mary Immaculate Sheela, L. (2020). Sentiment analysis of online product reviews using DLMNN and future prediction of online product using IANFIS. Journal of Big Data, 7(1). https://doi.org/10.1186/s40537-020-00308-7

[13] Test Thinking Aloud. (s/f). Ryte.com. Recuperado el 17 de abril de 2023, de https://es.ryte.com/wiki/Test_Thinking_Aloud

[14] Wu, J., Wang, Y., Zhang, R., & Cai, J. (2018). An approach to discovering product/service improvement priorities: Using dynamic importance-performance analysis. Sustainability, 10(10), 3564. https://doi.org/10.3390/su10103564

[15] Yang, C., Wu, L., Tan, K., Yu, C., Zhou, Y., Tao, Y., & Song, Y. (2021). Online user review analysis for product evaluation and improvement. Journal of Theoretical and Applied Electronic Commerce Research, 16(5), 1598–1611. https://doi.org/10.3390/jtaer16050090