INTERACCION
Revista Digital de AIPO
Asociación Interacción Persona-Ordenador
Vol. 6, No 2 (2025)
ISSN 2695-6578

Evaluación efectiva de usabilidad mediante técnicas de análisis y extracción de conocimiento

Effective usability evaluation by means of analysis and knowledge extraction techniques

Shuoshuo Li
Escuela Politécnica Superior
Universidad Autónoma de Madrid
Madrid, España
shuoshuo.li@estudiante.uam.es

José Antonio Macías Iglesias
Escuela Politécnica Superior
Universidad Autónoma de Madrid
Madrid, España
j.macias@uam.es

Recibido: 15.11.2025 | Aceptado: 01.12.2025

Palabras Clave

Extracción de conocimiento
Análisis de sentimientos
Análisis de audio y video
Pensando en voz alta
Evaluación de usabilidad

Resumen

Este trabajo propone la aplicación de técnicas para automatizar evaluaciones de usabilidad basadas en el protocolo *Thinking Aloud* en con el objetivo de superar las limitaciones inherentes a los enfoques manuales tradicionales. Para ello, se realiza una revisión sistemática de la literatura que permitirá identificar avances recientes y vacíos existentes en la aplicación de técnicas automatizadas. El análisis considera tecnologías emergentes como el reconocimiento automático de voz, el procesamiento de lenguaje natural y el análisis multimodal de audio y video, evaluando su potencial para capturar y procesar datos de interacción de manera eficiente y objetiva. Asimismo, se examinan los retos asociados a la integración de estas tecnologías, incluyendo aspectos relacionados con la fiabilidad, la reducción de sesgos y la escalabilidad del proceso. A partir de los hallazgos, se propone el diseño de una herramienta de soporte orientada a la combinación de métodos de aprendizaje y análisis multimodal para optimizar la detección de emociones y la extracción de conocimiento en tiempo real. Esta aproximación busca mejorar la calidad y la eficiencia de las evaluaciones de usabilidad, ofreciendo un marco metodológico que contribuya a la evolución hacia procesos más automatizados y menos dependientes de la intervención humana para aumentar la objetividad.

Keywords

Knowledge extraction
Sentiment Analysis
Audio and video processing
Thinking aloud
Usability testing

Abstract

This study proposes the automation of the *Thinking Aloud* protocol in usability evaluations, aiming to overcome the limitations of traditional manual approaches. To achieve this, a systematic literature review is conducted to identify recent advances and existing gaps in the application of automated techniques. The analysis examines emerging technologies such as automatic speech recognition, natural language processing, and multimodal audio-video analysis, assessing their potential to capture and process interaction data efficiently and objectively. Furthermore, the review addresses challenges related to the integration of these tools into testing environments, including issues of reliability, bias reduction, and process scalability. Based on these findings, the design of a supporting tool is presented, focusing on the combination of learning methods and multimodal analysis to enhance real-time emotion detection and knowledge extraction. This approach seeks to improve the quality and efficiency of usability evaluations by providing a methodological framework that supports the transition toward more automated processes, reducing human intervention and increasing objectivity.

1. Introducción

Las pruebas de usabilidad son un medio fundamental para evaluar la experiencia del usuario y la facilidad de uso de los sistemas interactivos (Hernando & Macías, 2023). El protocolo Thinking Aloud (TA) se considera uno de los métodos más utilizado para realizar investigaciones y evaluaciones de usabilidad (Boren & Ramey, 2000). En particular, este método

se ha adoptado ampliamente debido a su capacidad para registrar de manera intuitiva los procesos de pensamiento y los problemas que enfrentan los usuarios al ejecutar tareas (Boren & Ramey, 2000). Durante una prueba con TA, los usuarios expresan en voz alta sus pensamientos mientras realizan una tarea específica. Este enfoque de "pensar en voz alta mientras se actúa" ayuda a identificar problemas de usabilidad ocultos en el sistema y a comprender los patrones cognitivos de los usuarios (Hertzum & Holmegaard, 2015).

Sin embargo, la mayoría de las pruebas tradicionales con TA se desarrollan manualmente, a través de anotaciones por parte de observadores (Macías & Castells, 2001, 2002; Macías & Culén, 2021). En algunos casos, se utiliza también un proceso de grabación, transcripción y análisis manual, lo que no solo es un proceso que consume mucho tiempo y esfuerzo, sino que también dificulta la eliminación de juicios subjetivos en el análisis (Hertzum & Holmegaard, 2015).

Actualmente, y con el avance de la tecnología, es posible explorar otros enfoques tecnológicos para la predicción de emociones y la extracción automática de características. En los últimos años, se han propuesto y aplicado diversos métodos basados en aprendizaje profundo para la extracción automática de características en la predicción del estado emocional a partir de señales de voz (Jahangir et al., 2021). Sin embargo, algunos estudios han señalado que estos enfoques aún presentan ciertas limitaciones, como una excesiva concentración en el método concurrente, sin considerar en profundidad las diferencias en la aplicación de otros enfoques (McDonald et al., 2012).
Con los avances en el reconocimiento automático de voz (ASR), procesamiento de lenguaje natural (NLP), aprendizaje profundo y análisis de video, la automatización del análisis de grabaciones y videos de usuarios se ha convertido en una dirección clave para poder mejorar la eficiencia y objetividad de las evaluaciones mediante TA. En ese sentido, la investigación actual se centra principalmente en cómo aprovechar las tecnologías existentes para convertir de manera efectiva los datos de voz, texto y video de los usuarios en indicadores cuantificables de emociones y comportamiento, un problema que está recibiendo una creciente atención (Pang & Lee, 2008).

Este artículo tiene como objetivo explorar cómo lograr una evaluación de usabilidad automatizada usando el protocolo TA. A través de la integración de herramientas avanzadas de detección de emociones y extracción de conocimiento, se busca desarrollar un sistema de evaluación eficiente, objetivo y basado en datos. Además, se analizará el estado actual de las tecnologías existentes en el procesamiento de datos de voz, texto y video, así como sus limitaciones (Pang & Lee, 2008). De esta forma, se propone también el diseño de una herramienta de suporte al evaluador, proporcionando un sólido respaldo técnico para la toma de decisiones.

Este artículo se estructura de la siguiente forma. En la Sección 1 introduce el contexto de la investigación. En la Sección 2 se presenta una Revisión Sistemática de la Literatura, formulando las preguntas de investigación y dando contestación a cada una de ellas. En la Sección 3 se discuten los resultados obtenidos, así como las oportunidades de investigación junto con un análisis de las posibles amenazas a la validez identificadas durante el desarrollo del estudio. En la Sección 4 se describe una propuesta de diseño de una herramienta de soporte al evaluador, en base a las indagaciones realizadas en secciones anteriores. Finalmente, en la Sección 5 se presentan las conclusiones del trabajo y las líneas de trabajo futuro que definirán la continuidad de la investigación.

2. Análisis de trabajo relacionado

Para investigar sobre las posibilidades de automatización del protocolo TA, se llevará a cabo una SLR (*Systematic Literature Review* o Revisión Sistemática de la Literatura) que permita recopilar y analizar datos utilizando un método repetible y analítico (Kitchenham et al., 2009; Fernández & Macías, 2021). Este enfoque es adecuado para situaciones que requieren un análisis preciso, y es una vía apropiada para identificar y examinar la evidencia existente. La idea es realizar una integración exhaustiva de la literatura relacionada con la automatización de las pruebas de usabilidad basadas en TA, la detección de emociones y la extracción de conocimiento. Esto proporcionará una base teórica sólida para la selección de tecnologías y el diseño de una herramienta de soporte.

2.1 Método

En términos generales, la metodología adoptada puede variar entre diferentes disciplinas. El método empleado en este estudio se basa en una versión simplificada de una SLR (Kitchenham & Charters, 2007), con una adaptación específica del marco PICOC (*Población, Intervención, Comparación, Resultados y Contexto*) a nuestro dominio de investigación (Padua, 2010). Esto nos permitirá comprender de manera integral los últimos avances en detección de emociones, extracción de conocimiento y análisis de video dentro de la evaluación automatizada de la usabilidad mediante TA. El método SLR no solo garantiza la repetibilidad y sistematicidad, sino que también permite una recopilación y análisis exhaustivo y riguroso de la literatura existente. A continuación, se presentarán las preguntas de investigación, la cadena de búsqueda, los criterios de inclusión y exclusión, y el cribado y la selección de artículos.

2.2 Preguntas de investigación

Se plantean las siguientes preguntas de investigación para guiar la SLR y analizar los resultados obtenidos, en base a los objetivos marcados en la sección anterior:

- RQ1: ¿Qué propuestas existen para el análisis de emociones, principalmente a partir de grabaciones con usuarios, analizando video y audio?

- RQ2: ¿Qué métodos específicos existen para implementar el protocolo Thinking Aloud en estudios de usabilidad, considerando aquellos que puedan ser implementados de manera automática a través de métodos de extracción de conocimiento y de emociones?

- RQ3: ¿Qué tecnologías y herramientas de apoyo existen para la detección de emociones a partir de videos grabados con usuarios?

2.3 Selección de palabras clave y fuentes bibliográficas

Para poder encontrar literatura que respalde y responda las preguntas anteriores, se han seleccionado una serie de palabras clave que compondrán la cadena de búsqueda bibliográfica:

Situación experimental y entorno: *Thinking Aloud*
Tratamientos: *AND (Design Thinking OR Usability OR User Experience OR UX OR User-Centered Design OR User-Centred Design)*
Variables de respuesta: *AND (Automatic OR Voice OR Video OR Speech Recognition OR Conversational OR Natural Language Processing OR Intelligent OR Sentiment Analysis OR Transcription)*

Por otro lado, las bases de datos utilizadas para la búsqueda fueron las siguientes:

(1) IEEE Xplore
(2) ACM Digital Library
(3) Scopus
(4) Springer
(5) Google Scholar

2.4 Criterios de selección de artículos

Se definieron los siguientes criterios para filtrar los artículos extraídos de las cinco bases de datos mediante la cadena de búsqueda anterior:

- Criterios de inclusión: Los artículos seleccionados deben estar relacionados con evaluaciones de usabilidad que utilicen TA, e incluir procesos de automatización a través de la detección de emociones o la extracción de conocimiento. También, los artículos seleccionados deben estar en inglés o español, y deben haber sido publicados después del 2010 para asegurar que son trabajos recientes.

- Criterios de exclusión: Se descartarán aquellos artículos que no estén relacionados directamente con la temática definida, aquellos que sean antiguos, con contenido en curso, demasiado cortos o poco desarrollados, y aquellos que se encuentren repetidos.

2.5 Cribado y selección de artículos

Inicialmente, se obtuvo una alta cantidad de artículos que posteriormente fueron filtrados mediante los criterios de inclusión y exclusión. En la Tabla 1 se muestra un resumen de estos procesamientos iniciales. Para aplicar los criterios de inclusión y exclusión se utilizaron los filtros automáticos que proporcionan las bases de datos. Además, se analizaron los resúmenes, palabras clave y contenido general de los artículos, descartando aquellos que no cumplían con los criterios establecidos. Como podemos ver en la Tabla 1, la cantidad de artículos se redujo significativamente, obteniendo al final un subconjunto más reducido y relevante de artículos.

Tabla 3: Resultado de la búsqueda de artículos

Base de datos	Número inicial de artículos encontrados	Número de artículos tras aplicar los criterios de inclusión y exclusión
(1) IEEE Xplore	382	25
(2) ACM Digital Library	1240	44
(3) Scopus	980	39
(4) Springer	1650	60
(5) Google Scholar	32000	95

Tras otra revisión más concienzuda de los artículos resultantes, finalmente fueron 15 los artículos considerados como primarios, es decir, aquellos relevantes y directamente relacionados con el tema de investigación propuesto, por lo que se utilizarán principalmente para la contestar a las preguntas de investigación planteadas. Además, para ampliar el alcance de la revisión de la literatura, se utilizó el método de la *bola de nieve*. Para ello, tras seleccionar los 15 artículos primarios, se revisaron sus referencias bibliográficas,

encontrando otros 6 artículos adicionales considerados como relevantes para la investigación, lo que aumentó el número de artículos a 21. Estos artículos adicionales enriquecieron la investigación, proporcionando más argumentaciones e información de interés para elaborar este artículo. Por último, se seleccionaron 14 documentos más, a través de búsquedas menos estructuradas en Google. Estos documentos, considerados como *literatura gris*, tratan sobre temas técnicos que son de interés para enriquecer la investigación. Esto permitió finalmente contar con un total de 35 artículos de interés que permitieron abordar cada una de las preguntas de investigación, aportando bibliografía básica para poner en contexto cada una de ellas. Por un lado, para abordar la RQ1 se encontraron artículos que discuten principalmente los métodos de reconocimiento de emociones en video y audio, así como el análisis multimodal de emociones. Para la RQ2 se encontraron artículos que exploran cómo utilizar tecnologías de reconocimiento automático de voz, procesamiento de lenguaje natural y aprendizaje profundo. Por otro lado, para la RQ3 se encontraron artículos que presentan principalmente tecnologías de reconocimiento facial, detección de emociones en tiempo real y herramientas de código abierto relacionadas.

3. Resultados y discusión

La revisión en profundidad de la literatura arrojó un resultado alentador para la investigación, y es que, hasta el momento, no se han encontrado soluciones ni propuestas concretas completamente automatizadas para la evaluación de la usabilidad utilizando TA a través de tecnologías de reconocimiento de audio y video para la captura y análisis del pensamiento del usuario. A pesar de los avances en la automatización de algunos aspectos, como el análisis de emociones a partir de audio y video, todavía no existe un sistema integrado que cubra las necesidades del TA de una manera completamente automatizada. Por lo tanto, este campo sigue siendo un área de investigación poco explorada, lo que subraya la importancia de este estudio y su contribución para avanzar hacia una automatización más completa de las pruebas de usabilidad.

A continuación, se analizan cada una de las preguntas de investigación, discutiendo los trabajos encontrados y las soluciones más relevantes aportadas por la literatura.

(RQ1) ¿Qué propuestas existen para el análisis de emociones, principalmente a partir de grabaciones con usuarios, analizando video y audio?

Durante mucho tiempo, en la investigación sobre usabilidad ha habido una gran variedad de métodos utilizados para el análisis de emociones basados en grabaciones de usuarios (audio y video). Actualmente, se emplean principalmente los siguientes métodos:

1. Reconocimiento de emociones unimodales en voz. Los investigadores analizan las señales de voz y video de los usuarios, extrayendo características como el volumen, ritmo, energía y tono para detectar emociones. Klaus R. Scherer fue uno de los primeros académicos en investigar la relación entre las características de las señales de voz y la expresión emocional en 1986 (Scherer, 1986). En 2010, Rafael A. Calvo y Sidney D'Mello discutieron las teorías psicológicas de las emociones y revisaron los métodos tradicionales de detección de emociones (como la fisiología, el análisis facial y el análisis de voz), así como los sistemas multimodales emergentes. Calvo y D'Mello (Calvo & D'Mello, 2010) propusieron métodos de análisis de emociones basados en aprendizaje automático y modelos estadísticos. En 2011, El Ayadi y otros concluyeron que las señales de voz son una forma rápida y efectiva de interacción hombre-máquina, enfocándose en el reconocimiento de emociones en la voz (SER) en aplicaciones como la traducción y asistentes inteligentes, y resumieron las técnicas más relevantes, como la extracción de características y los métodos de clasificación (El Ayadi et al., 2011). También se identificaron algunas observaciones sobre los patrones de comportamiento de los usuarios en diferentes condiciones. Por ejemplo, una investigación señala que los analistas emplearon diversos métodos para identificar problemas de usabilidad. En el modo de "trabajo en silencio", los usuarios tendieron a depender más de la búsqueda en el sistema, mientras que en el modo de "pensar en voz alta", este método mostró un carácter más activo y orientado a objetivos (McDonald, 2020). Este hallazgo ofrece una perspectiva valiosa para comprender las características cognitivas del método TA.
2. Métodos de reconocimiento de emociones multimodales. Comenzó en los años 90 y, con el desarrollo de la tecnología, la combinación de voz y video ha mejorado la precisión del reconocimiento de emociones. En 1995, la investigación de la profesora Rosalind Picard sentó las bases para este enfoque (Poria et al., 2017). En 2017, Scherer y Ellgring señalaron que la combinación de voz y video permite un reconocimiento de emociones más efectivo (Scherer & Ellgring, 2007). Por otro

lado, en 2011, Mohammad Soleymani y otros investigaron la base de datos multimodal MAHNOB-HCI, que se utiliza para la investigación de la detección de emociones y etiquetado (Soleymani et al., 2012). Realizaron experimentos utilizando datos de video, movimientos oculares, audio y señales fisiológicas, y demostraron que los métodos multimodales son más precisos que los unimodales. A lo largo del tiempo, los investigadores han comparado diferentes métodos de reconocimiento de emociones. En 2011, Schuller y otros resumieron los avances en el reconocimiento de emociones a partir de la voz (Schuller et al., 2011). En 2017, Scherer y Ellgring estudiaron la expresión multimodal de las emociones y descubrieron que las emociones están influenciadas por el contexto (Scherer & Ellgring, 2007). Estos métodos, combinados con técnicas de aprendizaje profundo y aprendizaje automático, han mejorado la precisión del reconocimiento.

(RQ2) ¿Qué métodos específicos existen para implementar el protocolo Thinking Aloud en estudios de usabilidad, considerando aquellos que puedan ser implementados de manera automática a través de métodos de extracción de conocimiento y de emociones?

Para automatizar el TA, se pueden utilizar cinco métodos, cada uno basado en diferentes tecnologías que convierten la voz en texto y luego usan tecnologías avanzadas para realizar el análisis de emociones y la extracción de conocimiento, permitiendo el análisis automático de los informes verbales de los usuarios:

3. Transcripción automática de voz. Utiliza la tecnología de reconocimiento automático de voz para convertir las descripciones verbales en texto. Un estudio preliminar realizado por Nguyen y otros (2017) respalda esta idea, al demostrar que la transcripción de protocolos de pensamiento en voz alta mediante ASR es viable y puede integrarse de manera efectiva en estudios de usabilidad (Kuhn et al., 2024). En 2012, Geoffrey Hinton y otros señalaron que las redes neuronales profundas (DNN) son más eficaces que los métodos tradicionales, mejorando la precisión del reconocimiento de voz, y mencionaron las experiencias de cuatro equipos de investigación exitosos (Hinton et al., 2012).
4. Comprensión del texto y análisis de emociones. Después de la transcripción automática de voz, el texto se introduce en un sistema de procesamiento de lenguaje natural, utilizando modelos de lenguaje pre-entrenados como BERT para el análisis. En 2019, Jacob Devlin y otros señalaron que BERT es un modelo de lenguaje basado en *Transformer*, que aprende representaciones del lenguaje mediante un entrenamiento bidireccional profundo, y es adecuado para diversas tareas NLP (Sun et al., 2020). En comparación con los modelos tradicionales, BERT mostró mejor rendimiento en varias tareas, pudiendo identificar eficazmente las inclinaciones emocionales de los usuarios, apoyando la detección de emociones. Además de estos modelos avanzados, Liu, Li y Wang (2016) propusieron un enfoque basado en procesamiento de lenguaje natural que puede ser usado para el análisis automatizado del protocolo TA, demostrando que es posible identificar patrones problemáticos en los informes verbales de los usuarios de manera automática (Zhang et al., 2024).
5. Extracción de palabras clave y descubrimiento de conocimiento. El uso de tecnologías de aprendizaje profundo permite extraer palabras clave de manera rápida y precisa. En Wang y otros (2019) propusieron un método automático de extracción basado en aprendizaje profundo, que soporta la extracción de palabras clave de las transcripciones, lo que facilita la posterior extracción de conocimiento y el análisis de emociones (Sun et al., 2020). Posteriormente se desarrolló un enfoque específico, que podría ser utilizado para automatizar el TA, orientado a reducir la carga manual y mejorar la precisión del proceso (Li et al., 2025).
6. Modelado de contexto y manejo de textos largos. En 2017, Vaswani y otros concluyeron que *Transformer* puede utilizarse para manejar textos largos y modelar el contexto (Vaswani et al., 2017), lo que permitiría mejorar la comprensión del proceso de pensamiento del usuario y la precisión en la detección de emociones.
7. Explicación del modelo y validación de resultados- En 2016, Ribeiro propuso el modelo LIME, que ayuda a automatizar el proceso de extracción de características emocionales y de conocimiento (Ribeiro et al., 2016). Este método utiliza redes neuronales profundas para la transcripción de voz, BERT para el análisis de emociones, aprendizaje profundo para la extracción de palabras clave, y *Transformer* para el procesamiento del contexto, mejorando así la precisión y la eficiencia de los datos.

Por otro lado, los métodos multimodales pueden ser también de utilidad para en análisis de gestos en el protocolo TA. Si bien la mayor parte del conocimiento se puede extraer a través de la voz, también pueden ser de interés las emociones expresadas por el usuario a través de gestos o expresiones faciales. Sin embargo, la combinación de voz y video ha mejorado la precisión del reconocimiento de emociones, y prueba de ellos son las investigaciones que afirman que la combinación de voz y video permite un reconocimiento de emociones más efectivo (Scherer & Ellgring, 2007), a través de trabajos donde se realizaron experimentos utilizando datos de video, movimientos oculares, audio y señales fisiológicas, demostrando que los métodos multimodales son más precisos que los unimodales (Soleymai et al., 2012). Oros trabajos descubrieron que las emociones están influenciadas por el contexto (Scherer & Ellgring, 2007), y la combinación de técnicas de aprendizaje profundo y aprendizaje automático pueden ser la solución para un reconocimiento automático de emociones más completo en la evaluación automática de la usabilidad mediante TA.

En los últimos años, el desarrollo de las tecnologías de reconocimiento de emociones ha dado un nuevo impulso a las investigaciones en este ámbito. Por ejemplo, algunos estudios han demostrado que los métodos de reconocimiento emocional basados en expresiones faciales pueden alcanzar una tasa de precisión superior al 90 % en condiciones en tiempo real, lo que evidencia su viabilidad y eficacia en aplicaciones prácticas (Abdat et al., 2011). Además, la comunidad académica ha llevado a cabo revisiones sistemáticas sobre la aplicación de la computación afectiva en el reconocimiento de emociones psicológicas, lo cual aporta un sólido respaldo teórico para la comprensión de la percepción emocional en la interacción persona-ordenador (Bakkialakshmi &. Sudalaimuthu, 2021). Otros trabajos han desarrollado programas de reconocimiento de emociones faciales basados en realidad virtual, dirigidos específicamente a la evaluación y tratamiento asistido de personas con esquizofrenia (Souto et al., 2019). Al mismo tiempo, las tecnologías multimodales de reconocimiento emocional sin contacto también están evolucionando rápidamente, y la literatura existente ha explorado sus múltiples aplicaciones, desafíos técnicos, soluciones propuestas y perspectivas futuras (Khan et al., 2024). De la misma forma, en los últimos años, los investigadores también han comenzado a explorar diferentes enfoques tecnológicos, como los métodos de reconocimiento de emociones en el habla basados en aprendizaje profundo (Bhavan et al., 2020), así como el uso de la realidad aumentada para la detección de emociones a partir de expresiones faciales (Bhardwaj, 2023). Estos avances enriquecen aún más el contexto tecnológico en el que se enmarca este artículo.

(RQ3) ¿Qué tecnologías y herramientas de apoyo existen para la detección de emociones a partir de videos grabados con usuarios?

Las tecnologías actuales se centran principalmente en el reconocimiento de expresiones faciales en los videos. Para ello, se lleva a cabo la extracción de características faciales de los fotogramas del video para analizar el estado emocional. Las tecnologías y herramientas principales se dividen en los siguientes cuatro enfoques:

8. Métodos de reconocimiento de emociones en video basados en aprendizaje profundo. En 2017, Li y otros propusieron la base de datos RAF-DB y optimizaron el aprendizaje de características mediante el modelo DLP-CNN, demostrando que este modelo supera a los métodos existentes (Li et al., 2022). Ese mismo año, Mollahosseini y otros crearon la base de datos AffectNet, que recopila más de un millón de imágenes faciales, descubriendo que los métodos de Redes Neuronales Profundas (DNN) son más efectivos que los métodos tradicionales. En una investigación previa, los mismos autores habían evidenciado que el uso de arquitecturas DNN más profundas podía mejorar significativamente la precisión en el reconocimiento de expresiones faciales, sentando así una base técnica sólida para sistemas de análisis emocional en video (Mollahosseini et al., 2016).
9. Tecnologías de detección de emociones mediante el reconocimiento en tiempo real de expresiones faciales. En Shuster, M., y otros (2017) investigaron la tecnología de reconocimiento de expresiones faciales en tiempo real. Esta tecnología permite capturar rápidamente los cambios emocionales de los usuarios durante su interacción, asegurando la precisión y la inmediatez de la detección de emociones (Bartlett et al., 2003). En 2019, C. Jiang, Y. Qiu, H. Gao y otros propusieron una plataforma de retroalimentación para usuarios. Utilizando tecnología de aprendizaje profundo y video de alta velocidad, esta plataforma captura y analiza en tiempo real las expresiones faciales de los usuarios para proporcionar retroalimentación emocional. Esto podría ayudar a los diseñadores a optimizar dinámicamente la interfaz de usuario y mejorar los resultados de las pruebas de usabilidad (Jiang et al., 2019). En general, estas tecnologías podrían ser

adecuadas para pruebas de usabilidad en línea y otros escenarios altamente interactivos.

10. Herramientas y plataformas de código abierto. El paquete OpenFace, desarrollado en 2018 por T. Baltrušaitis y otros, es una herramienta de análisis facial de código abierto que puede detectar puntos clave faciales, postura de la cabeza, unidades de acción y seguimiento ocular, y soporta procesamiento en tiempo real. OpenFace tiene una alta precisión y puede ejecutarse en cámaras comunes, siendo aplicable en campos como la interacción persona-ordenador, la computación emocional y el análisis médico (Baltrušaitis et al., 2016).
11. Explicación del modelo y mejoras. En 2017, Selvaraju y otros propusieron el método Grad-CAM, que genera imágenes visuales utilizando información de gradientes, ayudando a entender las áreas clave en el reconocimiento de expresiones faciales (Selvaraju et al., 2017). Las tecnologías actuales, basadas en visión por computadora y aprendizaje profundo, pueden detectar en tiempo real el estado emocional del usuario, proporcionando soporte para las pruebas de usabilidad.

En general, todas las tecnologías y herramientas analizadas parecen prometedoras para su utilización dentro de la interacción con el usuario, con el objetivo de poder detectar información y conocimiento con precisión. Esto permitiría automatizar en análisis de resultados en pruebas de usabilidad basadas en TA. A través de la integración de herramientas avanzadas de detección de emociones y extracción de conocimiento, se podría desarrollar un sistema de evaluación eficiente, objetivo y basado en datos. Esto proporcionará un sólido respaldo técnico para mejorar la experiencia del usuario y el diseño de la interacción (Rojas & Macías, 2015), aportando además una solución automatizada de calidad al proceso (Quintal & Macías, 2021).

3.1 Oportunidades de investigación

El análisis realizado a partir de las tres preguntas de investigación revela patrones interesantes de convergencia tecnológica y áreas donde la integración entre diferentes enfoques podría generar avances significativos. Mientras que cada modalidad de análisis (audio, video, texto) ha demostrado capacidades prometedoras de manera individual, la verdadera oportunidad para la automatización completa del protocolo TA reside en la integración inteligente de todos estos enfoques.

La convergencia más significativa se observa en el uso generalizado de arquitecturas de aprendizaje profundo que pueden ser utilizadas en todas las modalidades. Desde los modelos de transformación para procesamiento de texto hasta las redes convolutivas profundas para análisis de video y el procesamiento de audio, existe una base tecnológica común que facilita la integración. Esta convergencia sugiere que es factible desarrollar arquitecturas unificadas que puedan procesar múltiples modalidades de datos de manera coherente y sincronizada, si bien para el caso de TA tiene más sentido el análisis de audio que el análisis de gestos en video.

No obstante, persisten brechas significativas en áreas clave. Primero, la sincronización temporal precisa entre diferentes modalidades sigue siendo un desafío técnico complejo, especialmente cuando se requiere realizar un análisis en tiempo real. Segundo, la interpretación contextual de eventos multimodales en relación con tareas específicas de usabilidad requiere un desarrollo adicional. Tercero, el proceso de adaptación para tener en cuenta las diferencias individuales y culturales respecto a la parte de expresión emocional y de patrones de verbalización presenta desafíos que van más allá de las capacidades técnicas actuales.

3.1.1 Implicaciones para el diseño de herramientas integradas

Los hallazgos sugieren que el diseño de herramientas para automatizar el protocolo TA debe considerar una arquitectura modular que permita la integración flexible de diferentes componentes tecnológicos. Esta arquitectura debe considerar no solo la precisión técnica de cada componente individual, sino también la coherencia y complementariedad de las percepciones generadas por diferentes modalidades de análisis. Un aspecto particularmente importante es la necesidad de métodos que puedan manejar situaciones donde exista información contradictoria. Por ejemplo, cuando el análisis de audio sugiere frustración, pero el análisis de video indica concentración, la herramienta debe incluir mecanismos para resolver estas inconsistencias de manera inteligente, posiblemente considerando el contexto específico de la tarea y el historial de comportamiento del usuario.

3.1.2 Consideraciones éticas y de privacidad

La implementación de herramientas de análisis multimodal para la automatización del protocolo TA también plantea importantes consideraciones éticas y de privacidad que deben ser abordadas de manera proactiva. La capacidad de estas herramientas para detectar estados emocionales y cognitivos detallados de los usuarios requiere el desarrollo de marcos éticos claros que protejan la privacidad y

autonomía de los participantes en estudios de usabilidad. Es particularmente importante la consideración de cómo el conocimiento automatizado podría ser utilizado o mal utilizado, especialmente en contextos comerciales donde la información sobre respuestas emocionales de usuarios podría tener valor económico significativo. Las herramientas construidas deben incorporar salvaguardas técnicas y procedimentales que aseguren que los datos emocionales de los usuarios sean utilizados exclusivamente para mejorar la usabilidad y la experiencia del usuario, y no para propósitos de manipulación o explotación comercial.

3.1.3 Desarrollo de estándares y protocolos

Una necesidad crítica identificada a través de este análisis es el desarrollo de estándares y protocolos específicos para la implementación de herramientas para automatizar el TA. Estos estándares deben abordar aspectos técnicos como la sincronización de datos multimodales, la calibración para diferentes poblaciones de usuarios, y los métodos de validación de resultados automatizados frente a los análisis manuales tradicionales.

Los protocolos también deben especificar mejores prácticas para la recolección de datos, incluyendo configuraciones recomendadas de hardware, procedimientos de configuración experimental, e instrucciones para participantes que optimicen la calidad de los datos capturados sin comprometer la naturalidad de la experiencia del usuario durante las sesiones TA.

3.1.4 Validación y confiabilidad

Otra área crítica de investigación relacionada es el desarrollo de métodos robustos para validar la confiabilidad y validez de las herramientas de análisis automático en TA. Esto incluye el desarrollo de métricas específicas que puedan comparar de manera significativa los resultados automatizados con los análisis manuales expertos, considerando que diferentes analistas humanos también pueden tener interpretaciones ligeramente diferentes de los mismos datos.
La investigación en validación debe también abordar la cuestión de cuándo y en qué condiciones las herramientas automáticas proporcionan conocimiento que van más allá de lo que puede ser detectado mediante el análisis manual tradicional. Esto podría incluir la capacidad de detectar patrones sutiles en datos multimodales que serían difíciles de percibir por analistas humanos, o la identificación de correlaciones entre diferentes aspectos de la respuesta del usuario que emergen solo cuando se analiza la totalidad de los datos (del *dataset*) de manera sistemática.

3.1.5 Integración con paradigmas de investigación existentes

Finalmente, se debe abordar cómo las herramientas automáticas podrían ser integradas de manera efectiva en los métodos de investigación de usabilidad existentes. Esto incluye el desarrollo de *workflows* híbridos donde la automatización debe complementar, en lugar de reemplazar completamente, la experiencia del humano, permitiendo que los investigadores se enfoquen en aspectos de alto nivel como la interpretación de resultados y la generación de recomendaciones de diseño.

La integración efectiva también requiere el desarrollo de interfaces de usuario intuitivas que permitan a los investigadores en usabilidad, que pueden no tener experiencia técnica en aprendizaje automático o procesamiento de señales, utilizar e interpretar eficazmente los resultados de las herramientas de automatización. Estas interfaces deben proporcionar tanto resultados de alto nivel como acceso a detalles técnicos cuando sea necesario requerido para una validación o investigación más profunda.

3.2 Amenazas a la validez

Con respecto a las amenazas a la validez (Rojas & Macías, 2019) relacionadas con el estudio presentado, el sesgo de publicación, el de selección, junto con la arbitrariedad de la cadena de búsqueda, son las principales amenazas internas a tener en cuenta, mientras que la generalizabilidad puede entenderse como la amenaza externa a considerar en este contexto.

Por un lado, el sesgo de publicación se refiere a la tendencia de que los resultados positivos se publiquen con mayor frecuencia que los negativos, lo que afecta a los resultados de la búsqueda de la literatura. Para reducir este sesgo, se utilizaron múltiples bases de datos (como IEEE Xplore, ACM Digital Library, etc.) y motores de búsqueda (como Google Scholar), además de seleccionar algunos documentos considerados como *literatura gris*, asegurando que se pudiera realizar una búsqueda amplia de investigaciones relacionadas con la automatización del protocolo TA.

Por otro lado, el sesgo de selección se refiere a la inclusión de literatura no relevante debido a criterios de selección demasiado amplios o no estrictos. Para evitar este sesgo, se establecieron criterios claros de inclusión: los artículos deben abordar temas relacionados con la automatización del protocolo TA, la detección de emociones y la extracción de conocimiento, y deben estar escritos en español o inglés, con una fecha de publicación posterior a 2010. Al mismo tiempo, se excluyeron estrictamente aquellos artículos que no se

ajustaran al tema de investigación, que estuvieran repetidos o que tuvieran un contenido demasiado breve, asegurando así que los artículos seleccionados fueran altamente relevantes para el tema de la investigación.

En lo que se refiere a la arbitrariedad de la cadena de búsqueda, ciertamente el diseño de la misma tiene un impacto importante en los resultados obtenidos. Para reducir los problemas de omisión o redundancia causados por combinaciones inadecuadas de palabras clave, estas fueron seleccionadas por los dos autores del artículo, asegurando así que la expresión de búsqueda pudiera capturar con precisión todos los artículos clave relacionados con el tema de investigación planteado.

Finalmente está el problema de la generalizabilidad, es decir, la posibilidad de que los resultados de la investigación no sean aplicables a un campo más amplio. Para reducir este problema se utilizó el método *snowball* o de *la bola de nieve*, aumentando así el número de trabajos relacionados y asegurando que los artículos seleccionados representan mejor el tema de investigación tratado.

Aunque podrían seguir existiendo amenazas a la validez, las acciones llevadas a cabo permiten minimizarlas, proporcionaron una base sólida de literatura para el estudio y análisis de la automatización del protocolo TA a partir de métodos modernos de aprendizaje y extracción automatizada de conocimiento y emociones, que es el objeto de estudio.

4. Propuesta de herramienta automatizada

En base a las conclusiones presentadas en la sección anterior, se propone el diseño de una herramienta que permita el análisis automático de evaluaciones de usabilidad a través del protocolo TA, no como método único, sino para la toma de decisiones, de forma que los resultados puedan ser comparados con la percepción de otros observadores humanos. Esto permitiría cubrir las carencias existentes en este ámbito y desarrollar un sistema automatizado que combine datos multimodales y proporcione retroalimentación rápida y eficaz en evaluaciones de usabilidad basadas en TA.

4.1 Arquitectura software

La herramienta estará basada principalmente en el análisis de audio; interpretará tanto ficheros de audio como de video que contengan las opiniones de los usuarios grabadas a partir de sesiones TA. La herramienta se implementará mediante una arquitectura modular fundamentada en el patrón Modelo-Vista Controlador. La selección de este patrón arquitectónico responde a la necesidad de integrar flexiblemente diferentes componentes de análisis especializados, mantener la escalabilidad de la herramienta para manejar volúmenes crecientes de datos multimedia, y asegurar la mantenibilidad del código para facilitar futuras extensiones. El patrón MVC se adapta particularmente bien a las características de la herramienta a diseñar debido a la clara separación de responsabilidades que requiere el procesamiento de datos multimedia. El modelo maneja la persistencia de datos complejos incluyendo archivos multimedia, metadatos de análisis, y resultados procesados. La vista gestiona múltiples interfaces especializadas para diferentes tipos de usuarios, desde evaluadores/investigadores individuales hasta administradores de equipos de UX. El controlador coordinará flujos de trabajo complejos que involucran procesamiento asíncrono, integración con APIs externas, y generación de reportes dinámicos.

En la Figura 1 se muestran los principales módulos y etapas funcionales de la herramienta a construir. En concreto, se proponen ocho funcionalidades secuenciales que abarcan desde la entrada que proporciona el evaluador, compuesta por grabaciones de sesiones TA, hasta la salida que proporciona la herramienta, y que estará compuesta por un informe detallado sobre los problemas concretos de usabilidad identificados durante las sesiones TA suministradas como entrada.

Figura 3: Las ocho etapas funcionales de la herramienta.

A continuación, se describen estas etapas:

- Entrada y validación: En esta etapa, la herramienta permitirá cargar archivos multimedia, usando formatos estándar de vídeo y audio. También validará dicho formato, la duración y la calidad, haciendo las adaptaciones necesarias para mejorar en lo posible la calidad de la entrada.

- Preprocesamiento: En esta etapa se extraen las pistas de audio, incluyendo las que se encuentren en archivos de video, se normalizará el volumen, y se aplican filtros básicos de ruido.
- Transcripción ASR (*Automatic Speech Recognition*): En esta etapa, el audio procesado en la etapa anterior se convertirá en texto utilizando tecnologías de reconocimiento automático de voz.
- Análisis NLP (*Natural Language Processing*): En esta etapa, el texto obtenido en la etapa anterior se procesa mediante técnicas de procesamiento de lenguaje natural para su *tokenización* y análisis sintáctico.
- Análisis (emocional) de sentimientos: En esta etapa se aplican técnicas de análisis de sentimientos para cuantificar la polaridad emocional de cada segmento obtenido en la etapa anterior.
- Detección de problemas: En esta etapa se aplican algoritmos especializados para detectar patrones específicos que permitan identificar problemas concretos de usabilidad.
- Generación de sugerencias: En esta etapa, los problemas encontrados en la etapa anterior se clasifican por nivel de severidad y se generan las recomendaciones específicas para cada uno de dichos problemas de usabilidad.
- Salida de reporte: En esta etapa, los resultados obtenidos se almacenan en una base de datos, y además se generan visualizaciones y reportes específicos para el usuario final (evaluador), con información sobre el análisis realizado.

Respecto a los detalles técnicos, por un lado, la extracción de audio incluirá un análisis inteligente de los archivos de video para identificar y extraer únicamente las pistas de audio relevantes para el análisis TA. Esto incluye la detección automática de múltiples pistas de audio en archivos complejos, identificación de la pista que contiene verbalizaciones del usuario en contraposición al audio de la propia herramienta o el ruido ambiental, y la preservación de sincronización temporal precisa que permita correlacionar posteriormente transcripciones específicas con momentos exactos en la grabación original.

La normalización de audio implica la utilización de algoritmos sofisticados que ajustan no solamente el volumen general, sino también la dinámica de la señal para optimizar el reconocimiento automático. Esto incluye el uso de compresión dinámica para reducir variaciones extremas de volumen sin eliminar información emocional importante, la ecualización para enfatizar frecuencias de voz humana mientras se suprimen ruidos típicos de grabaciones de campo, y la supresión del ruido que elimina segmentos de silencio o ruido de fondo que pueden interferir con el reconocimiento. La segmentación divide estratégicamente el audio en *chunks* de duración optimizada para su procesamiento a través de APIs de reconocimiento. La duración debe representar un balance cuidadoso entre varios factores, relacionados con la maximización de la precisión de reconocimiento (por ejemplo, los segmentos demasiado cortos pierden contexto, y los que son demasiado largos aumentan errores). Se debe también optimizar el uso de APIs comerciales (que frecuentemente tienen límites de duración por petición), y facilitar el procesamiento paralelo para reducir el tiempo total de análisis.

Por otro lado, el módulo de reconocimiento automático de voz es responsable de la transformación crítica de verbalizaciones auditivas en el texto estructurado que posteriormente puedan ser procesadas por algoritmos especializados de análisis de lenguaje natural para, posteriormente, poder detectar problemas de usabilidad. Esta transformación representa uno de los desafíos técnicos más complejos de la herramienta a construir, ya que debe manejar las características específicas del habla conversacional típica en sesiones del protocolo TA, incluyendo frases falsas, pausas cognitivas, lenguaje informal, y expresiones emocionales espontáneas.

La arquitectura del módulo de reconocimiento automático debe ser diseñada considerando las limitaciones y variabilidades inherentes en las grabaciones de campo típicas de la investigación en UX, donde las condiciones de audio pueden ser subóptimas debido a factores como el ruido ambiental, la calidad variable de los equipos de grabación, los diferentes acentos y patrones de habla de los participantes, y las fluctuaciones en el volumen de voz durante la verbalización de los pensamientos. Esta realidad práctica requiere la implementación de múltiples capas de procesamiento y optimización que van más allá de la simple aplicación de APIs de reconocimiento de voz comerciales. La implementación técnica de este módulo debe integrar múltiples tecnologías complementarias que trabajen en conjunto para optimizar la precisión de la transcripción. Esto es particularmente importante, ya que los diferentes motores ASR tienen fortalezas específicas: algunos son adecuados para reconocimiento de lenguaje conversacional informal, otros para audio con ruido de fondo, y algunos para acentos específicos o terminología técnica. La capacidad de cambiar dinámicamente entre motores según las características detectadas del audio permite optimizar la precisión de transcripción caso por caso.

Otra parte importante es lo relacionado con el análisis de sentimientos y el mapeo a problemas concretos de

usabilidad a detectar. Para ello, se utilizarán herramientas para analizar textos informales y conversacionales, estudiando la polaridad e intensidad tanto positiva como negativa de los sentimientos identificados. Adicionalmente, la herramienta implementará un algoritmo de detección de patrones basado en expresiones regulares optimizadas, utilizado para identificar expresiones lingüísticas específicas relacionadas con problemas conocidos de usabilidad, como se ha indicado en la literatura analizada. Para ello, se realizarán clasificaciones en base a la severidad del problema encontrado y a su factor emocional asociado. La herramienta incluirá también un motor de reglas que permita mapear automáticamente los tipos de problemas detectados a sugerencias específicas de mejora, proporcionando valor inmediato a los usuarios. Durante todo el proceso, se tendrán en cuenta métricas y valores necesarios para la estimación y cálculo de la bondad de las predicciones realizadas.

4.2 Prototipo

En la Figura 2 se presenta un prototipo de la herramienta web construida, atendido a las consideraciones narradas en la sección anterior.

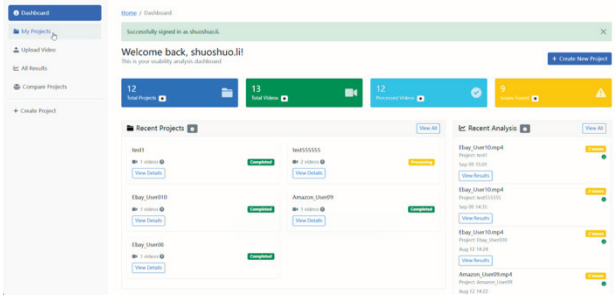

Figura 2: Prototipo de la herramienta.

Por un lado, la herramienta permite la gestión de usuarios y proyectos. De esta forma, los evaluadores pueden crear sus propios proyectos de evaluación y subir los videos y audios de las sesiones TA que deseen analizar. Una vez subidos, la herramienta procede a realizar las etapas descritas en la Figura 1, obteniendo como resultado una serie de indicadores y salidas, en forma de *dashboard* o cuadro de mandos, que le servirán al evaluador para la toma de decisiones. Todos los resultados se guardan de manera permanente en una base de datos, de forma que puedan ser accedidos en cualquier momento por parte de los evaluadores. Además, la herramienta permitirá la comparación de distintos tipos de proyectos para identificar problemas de usabilidad comunes a todos ellos.

5. Conclusión y trabajo futuro

El protocolo TA es uno de los más utilizados en las pruebas de evaluación de la usabilidad con usuarios, ya que permite, de una forma expresiva, conocer las sensaciones de este mientras interactúa con la aplicación a evaluar, lo que facilita obtener información de primera mano a los evaluadores. La mayoría de las evaluaciones tradicionales mediante TA se desarrollan manualmente, a través de anotaciones que toman los observadores. En algunos casos, se utiliza también un proceso de grabación para su posterior procesamiento, o de transcripción y análisis manual, lo que consume tiempo y esfuerzo.

Con los avances en el reconocimiento automático de voz, procesamiento de lenguaje natural, aprendizaje profundo y análisis de video, la automatización del análisis de grabaciones y videos de usuarios es una solución interesante para mejorar la eficiencia y objetividad de las evaluaciones mediante TA, lo que permitiría un proceso más sistemático de análisis y obtención automática de resultados, logrando pautas para integrar de manera satisfactoria la inteligencia artificial en la interacción con el usuario (Macías 2008; Macías 2012).

Según la literatura existente, no existen todavía propuestas concretas en ese sentido, pero sí hay tecnologías y herramientas que ayudarían a implementar soluciones que permitirían la automatización del protocolo TA y su posterior análisis, especialmente a través del análisis de emociones y de la extracción automática de conocimiento. Esto incluye el reconocimiento automático de voz, que ha permitido mejorar la precisión de la transcripción de voz, haciendo el proceso más eficiente. Por otro lado, el análisis de emociones y procesamiento de lenguaje natural tiene cabida a través de herramientas que permiten analizar el texto transcrito e identificar automáticamente las emociones del usuario. Además, el reconocimiento de emociones multimodales, que permite combinar datos de voz y video, permite capturar con mayor precisión el estado emocional del usuario. Todo ello permitiría extraer y analizar en tiempo real información para asistir en el análisis emocional.

En base a estas indagaciones, se proponen pautas para la construcción de una herramienta de soporte que automatice la detección de problemas de usabilidad en evaluaciones mediante el protocolo TA. La herramienta sigue un flujo lógico compuesto por ocho pasos que permiten cargar información inicial (audio y video) sobre las sesiones TA con usuarios y, como último paso, la emisión de informes de resultados sobre los problemas de usabilidad encontrados, siendo la ejecución del resto de pasos totalmente transparente para el usuario evaluador.

Como líneas futuras para continuar esta investigación, se propone la finalización y testeo de la herramienta de

automatización propuesta, lo que permitiría observar hasta qué punto se puede aumentar la eficiencia, reduciendo al mismo tiempo la intervención humana y mejorando la objetividad. También se pueden plantear otros escenarios que presentan de por sí retos de investigación en esta línea, como la posible falta de precisión en entornos ruidosos y las dificultades técnicas en el análisis emocional en tiempo real cuando se procesan grandes volúmenes de datos. También, la inclusión del análisis facial proveniente de vídeo podría ser explorado como otro canal de entrada de información multimodal, si bien este tipo de información es menos relevante en evaluaciones TA, ya que el audio es la principal fuente de entrada y de conocimiento para evaluaciones basadas en este protocolo.

Agradecimientos

Esta investigación ha sido subvencionada a través de los proyectos de investigación TED2021-129381B-C21, PID2021-122270OB-I00 y PID2024-155231OB-I00, de la Agencia Estatal de Investigación.

Referencias

Abdat, F., Maaoui, C., & Pruski, A. (2011). Human–computer interaction using emotion recognition from facial expression. *UKSim 5th European Symposium on Computer Modeling and Simulation* (pp. 196–201). https://doi.org/10.1109/EMS.2011.20

Bakkialakshmi, V. S., & Sudalaimuthu, T. (2021). A survey on affective computing for psychological emotion recognition. *5th International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques (ICEECCOT)* (pp. 480–486). https://doi.org/10.1109/ICEECCOT52851.2021.9707947

Baltrušaitis, T., Robinson, P., & Morency, L. P. (2016). OpenFace: An open source facial behavior analysis toolkit. *IEEE Winter Conference on Applications of Computer Vision (WACV)* (pp. 1–10). https://doi.org/10.1109/WACV.2016.7477553

Bartlett, M. S., Littlewort, G., Fasel, I., & Movellan, J. R. (2003). Real-time face detection and facial expression recognition: Development and applications to human–computer interaction. *Proceedings of the 2003 Conference on Computer Vision and Pattern Recognition Workshop* (p. 53). https://doi.org/10.1109/CVPRW.2003.10057

Bhavan, A., Sharma, M., Piplani, M., Chauhan, P., Hitkul, S., & Shah, R. R. (2020). Deep learning approaches for speech emotion recognition. In B. Agarwal, R. Nayak, N. Mittal, & S. Patnaik (Eds.), *Deep learning-based approaches for sentiment analysis* (Algorithms for Intelligent Systems). Springer. https://doi.org/10.1007/978-981-15-1216-2_10

Bhardwaj, V., Joshi, A., Bajaj, G., Sharma, V., Rushiya, A., & Bharghavi, S. S. (2023). Emotion detection from facial expressions using augmented reality. *2023 5th International Conference on Inventive Research in Computing Applications (ICIRCA)* (pp. 1–5). https://doi.org/10.1109/ICIRCA57980.2023.10220824

Boren, T., & Ramey, J. (2000). Thinking aloud: Reconciling theory and practice. *IEEE Transactions on Professional Communication, 43*(3), 261–278. https://doi.org/10.1109/47.867942

Calvo, R. A., & D'Mello, S. (2010). Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing, 1*(1), 18–37. https://doi.org/10.1109/T-AFFC.2010.7

El Ayadi, M., Kamel, M., & Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition, 44*(3), 572–587. https://doi.org/10.1016/j.patcog.2010.09.020

Fernández, J., & Macías, J. A. (2021, September). Heuristic-based usability evaluation support: A systematic literature review and comparative study. *In Proceedings of the XXI International Conference on Human-Computer Interaction* (pp. 1-9). https://doi.org/10.1145/3471391.3471395

Hernando, R., & Macías, J. A. (2023). Development of usable applications featuring QR codes for enhancing interaction and acceptance: a case study. *Behaviour & Information Technology*, *42*(4), 360-378. https://doi.org/10.1080/0144929X.2021.2022209

Hertzum, M., & Holmegaard, K. D. (2015). Thinking aloud influences perceived time. *Human Factors, 57*(1), 101–109. https://doi.org/10.1177/0018720814549709

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-R., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., & Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine, 29*(6), 82–97. https://doi.org/10.1109/MSP.2012.2205597

Jahangir, R., Teh, Y., Wah, H., Faiqa, F., & Mujtaba, G. (2021). Deep learning approaches for speech emotion recognition: State of the art and research challenges. *Multimedia Tools and Applications, 80*(16), 23745–23812. https://doi.org/10.1007/s11042-020-09874-7

Jiang, C., Qiu, Y., Gao, H., Fan, T., Li, K., & Wan, J. (2019). An edge computing platform for intelligent operational monitoring in Internet data centers. *IEEE Access, 7*, 133375–133387. https://doi.org/10.1109/ACCESS.2019.2939614

Kitchenham, B., & Charters, S. (2007). *Guidelines for performing systematic literature reviews in software engineering*(EBSE 2007 Technical Report). EBSE. https://www.durham.ac.uk/media/durham-university/departments-/computer-science/research/technical-reports/Guidelines-for-Performing-Systematic-Literature-Reviews-in-Software-Engineering-2007.pdf

Kitchenham, B., Brereton, O. P., Budgen, D., Turner, M., Bailey, J., & Linkman, S. (2009). Systematic literature reviews in software engineering – A systematic literature review. *Information and Software Technology, 51*(1), 7–15. https://doi.org/10.1016/j.infsof.2008.09.009

Khan, U. A., Xu, Q., Liu, Y., Lagstedt, A., Alamäki, A., & Kauttonen, J. (2024). Exploring contactless techniques in multimodal emotion recognition: Insights into diverse applications, challenges, solutions, and prospects. *Multimedia Systems, 30*(115). https://doi.org/10.1007/s00530-024-01302-2

Kuhn, K., Kersken, V., Reuter, B., Egger, N., & Zimmermann, G. (2024). Measuring the accuracy of automatic speech recognition solutions. *ACM Transactions on Accessible Computing, 16*(3), Article 25, 23 pages. https://doi.org/10.1145/3636513

Li, S., & Deng, W. (2022). Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing, 13*(3), 1195–1215. https://doi.org/10.1109/TAFFC.2020.2981446

Li, S., Huang, X., Wang, T., Zheng, J. & Lajoie, S. (2025). Using text mining and machine learning to predict reasoning activities from think-aloud transcripts in computer-assisted learning. *Journal of Computing in Higher Education, 37*, 477–496. https://doi.org/10.1007/s12528-024-09404-6

Macías, J. A. (2008). Intelligent assistance in authoring dynamically generated web interfaces. *World Wide Web, 11*(2), 253–286. https://doi.org/10.1007/s11280-008-0043-3

Macías, J. A. (2012). Enhancing interaction design on the semantic web: A case study. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 42*(6), 1365–1373. https://doi.org/10.1109/TSMCC.2012.2187052

Macías, J. A., & Castells, P. (2001). A generic presentation modeling system for adaptive web-based instructional applications. In *CHI'01 Extended Abstracts on Human Factors in Computing Systems* (pp. 349-350). https://doi.org/10.1145/634067.63427

Macías, J. A., & Castells, P. (2002). Tailoring dynamic ontology-driven web documents by demonstration. In *Proceedings Sixth International Conference on Information Visualisation* (pp. 535-540). IEEE. https://doi.org/10.1109/IV.2002.1028826

Macías, J. A., & Culén, A. L. (2021). Enhancing decision-making in user-centered web development: a methodology for card-sorting analysis. *World Wide Web, 24*(6), 2099-2137. https://doi.org/10.1007/s11280-021-00950-y

McDonald, S., Cockton, G., & Irons, A. (2020). The impact of thinking-aloud on usability inspection. *Proceedings of the ACM on Human-Computer Interaction, 4*(CSCW1), 1–22. https://doi.org/10.1145/3397876

McDonald, S., Edwards, H. M., & Zhao, T. (2012). Exploring think-alouds in usability testing: An international survey. *IEEE Transactions on Professional Communication, 55*(2), 2–19. https://doi.org/10.1109/TPC.2011.2182569

Mollahosseini, A., Hasani, B., & Mahoor, M. H. (2016). Going deeper in facial expression recognition using deep neural networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW 2016)* (pp. 1–6). https://doi.org/10.1109/CVPRW.2016.7477553

Padua, A. G. (2010). Propuesta de un proceso de revisión sistemática de experimentos en ingeniería del software. *Proceedings of the 13th Ibero-American Conference on Software Engineering (CIbSE 2010)* (pp. 313–318). Universidad Politécnica de Madrid.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval, 2*(1–2), 1–135. https://doi.org/10.1561/1500000011

Poria, S., Cambria, E., Bajpai, R., & Hussain, A. (2017). A review of affect analysis: From unimodal analysis to multimodal fusion. *Information Fusion, 37*, 98–125. https://doi.org/10.1016/j.inffus.2017.02.003

Quintal, C., & Macías, J. A. (2021). Measuring and improving the quality of development processes based on usability and accessibility. *Universal Access in the Information Society, 20*(2), 203–221. https://doi.org/10.1007/s10209-020-00726-7

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*(pp. 1135–1144). https://doi.org/10.1145/2939672.2939778

Rojas, L. A., & Macías, J. A. (2015). An agile information-architecture-driven approach for the development of user-centered interactive software. *Proceedings of the XVI International Conference on Human–Computer Interaction* (Article No. 50, pp. 1–8). https://doi.org/10.1145/2829875.2829919

Rojas, L. A., & Macías, J. A. (2019). Toward collisions produced in requirements rankings: A qualitative approach and experimental study. *Journal of Systems and Software, 158*, 110417. https://doi.org/10.1016/j.jss.2019.110417

Scherer, K. R. (1986). Vocal communication of emotion: A review of research paradigms. *Speech Communication, 5*(1–2), 1–49. https://doi.org/10.1016/0167-6393(86)90070-X

Scherer, K. R., & Ellgring, H. (2007). Multimodal expression of emotion: Affect programs or componential appraisal patterns? *Emotion, 7*(1), 158–171. https://doi.org/10.1037/1528-3542.7.1.158

Schuller, B., Batliner, A., Steidl, S., & Seppi, D. (2011). Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Communication, 53*(9–10), 1062–1087. https://doi.org/10.1016/j.specom.2011.01.011

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. *2017 IEEE International Conference on Computer Vision (ICCV)*(pp. 618–626). https://doi.org/10.1109/ICCV.2017.74

Soleymani, M., Lichtenauer, J., Pun, T., & Pantic, M. (2012). A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing, 3*(1), 42–55. https://doi.org/10.1109/T-AFFC.2011.25

Souto, T., Silva, H., Leite, A., Baptista, A., Queirós, C., & Marques, A. (2019). Facial emotion recognition: Virtual reality program for facial emotion recognition—a trial program targeted at individuals with schizophrenia. *Rehabilitation Counseling Bulletin, 63*(2), 79–90. https://doi.org/10.1177/0034355219847284

Sun, Y., Qiu, H., Zheng, Y., Wang, Z., & Zhang, C. (2020). SIFRank: A new baseline for unsupervised keyphrase extraction based on pre-trained language model. *IEEE Access, 8*, 10896–10906. https://doi.org/10.1109/ACCESS.2020.2965087

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*, 5998–6008.

Zhang, J., Borchers, C., Aleven, V., & Baker, R. S. (2024). Using large language models to detect self-regulated learning in think-aloud protocols. *Proceedings of the 17th International Conference on Educational Data Mining* (pp. 157–168). International Educational Data Mining Society. https://doi.org/10.5281/zenodo.12729790