

Documentación de Bitácora para investigación

Escucha inteligente: transcripción, análisis de sentimientos y clasificación de emociones en español

Moyano, María Emilia

Estudiante de Ingeniería informática UNNOBA
memoyano411@comunidad.unnoba.edu.ar

Koroluk, Mijael Andrés

Estudiante de Ingeniería informática UNNOBA
makoroluk@comunidad.unnoba.edu.ar

Badano, Valentino

Estudiante de Ingeniería informática UNNOBA
vbadano@comunidad.unnoba.edu.ar

Introducción

Este trabajo fue realizado en el marco de la asignatura Sistemas Inteligentes, con el objetivo de aplicar técnicas y herramientas propias de la inteligencia artificial al análisis automático de sentimientos (AS) y emociones (CE) en interacciones humanas. La propuesta integra diferentes componentes de procesamiento de lenguaje natural (PLN), incluyendo reconocimiento automático del habla (RAH), análisis de sentimientos y emociones y síntesis de resultados, para construir una herramienta funcional y aplicable a diversos contextos reales.

1. Analizador Automático de Sentimientos y Emociones

1.1 Descripción

En contextos donde se analiza contenido multimedia, como grabaciones de reuniones, entrevistas, presentaciones o debates resulta de gran utilidad contar con herramientas automáticas capaces de identificar el tono emocional predominante y ofrecer un resumen interpretativo del contenido.

El AS y CE busca detectar polaridades y emociones básicas expresadas a través del lenguaje, y constituye un área clave dentro de la inteligencia artificial aplicada al PLN [1].

Este tipo de análisis puede brindar apoyo en múltiples ámbitos, como el monitoreo de experiencias del cliente, la evaluación del clima laboral, la detección de estados anímicos en contextos terapéuticos, o incluso la evaluación de la dinámica emocional en discusiones grupales.

1.2 Resultado esperado

Se propone desarrollar un sistema que analice automáticamente los sentimientos y las emociones expresadas en interacciones habladas, a partir de la transcripción textual obtenida mediante RAH. Una vez segmentadas las intervenciones por participante, se aplicarán modelos de PLN como PySentimiento [2, 3], especializado en el AS y CE en español, y RoBERTa-base-GoEmotions [4], entrenado sobre un conjunto amplio de emociones en inglés, para identificar emociones básicas (alegría, tristeza, enojo, etc.) y polaridad/sentimiento (positiva, negativa o neutra).

Como resultado, se generará un informe en formato PDF que incluirá gráficos y tablas con un resumen de los hallazgos, destacando la emoción y el sentimiento predominante por participante y en el contenido general.

2. Comprensión de la problemática

En el contexto actual, donde los datos textuales y multimedia crecen exponencialmente, la capacidad de identificar emociones, opiniones y polaridades en tiempo real ofrece ventajas significativas para la toma de decisiones informadas. El AS y CE permiten evaluar aspectos clave como el estado de ánimo colectivo, la satisfacción de los participantes, y la dinámica interpersonal, lo que resulta particularmente útil en sectores como recursos humanos, educación, marketing o servicios al cliente. Sin embargo, el análisis manual de estos datos es costoso, subjetivo y no escalable, especialmente cuando el contenido involucra múltiples participantes o modalidades (texto, audio y video).

Tradicionalmente, el enfoque del AS y CE se ha centrado en la evaluación de textos escritos [1,5]. Sin embargo, en los últimos años, ha habido un creciente interés en expandir este análisis a modalidades de audio y visuales, así como a combinaciones de estas [6,7]. Esto permite una comprensión más rica y matizada de las emociones humanas, ya que la comunicación no verbal y las expresiones auditivas pueden aportar información valiosa que complementa el texto escrito.

Adicionalmente, aunque se han logrado avances notables en el AS y CE, la mayoría de las investigaciones existentes se han centrado exclusivamente en el idioma inglés, dejando un vacío significativo en el desarrollo de herramientas para idiomas con menor cantidad de datos.

2.1 Objetivo general

El objetivo general de este trabajo es desarrollar un sistema automatizado que permita el AS y CE expresadas en interacciones habladas en idioma español, a partir de la transcripción generada de archivos de audio o vídeo. El sistema estará orientado a ofrecer una herramienta útil para interpretar el contenido emocional de conversaciones humanas, especialmente en contextos como reuniones, entrevistas o presentaciones.

2.2 Objetivos específicos

Incorporar una herramienta de transcripción automática de audio y video a texto: Incorporar un componente que permita convertir archivos de audio o video en idioma español a texto, utilizando tecnologías de RAH.

Identificar emociones y polaridades expresadas por múltiples participantes: Identificar y etiquetar emociones y polaridades asociadas a cada participante de la interacción, tanto a nivel individual como global, considerando su predominancia a lo largo de la conversación.

Validar el sistema mediante métricas de desempeño: Validar el funcionamiento del sistema utilizando métricas cuantitativas que incluyan la precisión en la transcripción, la exactitud en la clasificación emocional y la consistencia en la diarización de participantes. Para evaluar el grado de confiabilidad en la clasificación múltiple, se utilizará el coeficiente Kappa de Fleiss.

2.3 Estado del Arte

El AS y CE han evolucionado significativamente a partir del desarrollo de modelos avanzados de PLN. Inicialmente, se emplearon técnicas supervisadas basadas en representaciones simples del texto, como bag-of-words (BoW) y TF-IDF, combinadas con clasificadores tradicionales como máquinas de soporte vectorial (SVM) y regresión logística [5].

La introducción de representaciones vectoriales semánticas (word embeddings), tales como Word2Vec [49, 50] y GloVe [51], permitió capturar relaciones más profundas entre palabras mediante vectores densos, mejorando la capacidad de los modelos para identificar matices afectivos. Posteriormente, las redes neuronales profundas, en particular las redes convolucionales (CNN) y recurrentes (RNN) [52], posibilitaron modelar con mayor precisión la estructura lingüística y las dependencias contextuales del texto.

Un avance disruptivo fue la aparición de los modelos basados en transformers [8], que mediante el mecanismo de autoatención (self-attention) procesan

secuencias completas simultáneamente, captando relaciones complejas entre palabras sin depender del procesamiento secuencial.

Uno de los modelos más influyentes es BERT [9], preentrenado con grandes volúmenes de texto y posteriormente ajustado para tareas específicas como la clasificación de sentimientos. Su arquitectura ha dado lugar a múltiples variantes y ha establecido un estándar de referencia en el campo. En este contexto, Demszky et al. (2020) desarrollaron GoEmotions [10], un corpus en inglés con 27 etiquetas emocionales derivadas de comentarios en Reddit. Utilizando modelos basados en RoBERTa [53], una variante optimizada de BERT, lograron los siguientes resultados: macro F1 de 0.69 para sentimientos, 0.64 para emociones básicas de Ekman y 0.46 para el conjunto completo de 27 emociones. Estos resultados reflejan la complejidad creciente en el reconocimiento afectivo a medida que se incrementa la granularidad emocional. Sin embargo, estos modelos, al igual que herramientas comerciales consolidadas como Google Cloud Natural Language [23], Azure Text Analytics [24] y Amazon Comprehend [25], presentan limitaciones para el análisis de interacciones orales en español; además, estos últimos no están disponibles gratuitamente y carecen de acceso abierto y opciones de personalización detallada de reportes.

Para el español, Perez et al. (2022) crearon PySentimiento [2], una biblioteca que adapta modelos preentrenados como BETO [11] y variantes de RoBERTa, para clasificar polaridad y emociones básicas en español. En evaluaciones reportadas por sus autores, BETO alcanza un macro F1 de 67.2 tanto para sentimientos como para emociones básicas, mientras que RoBERTuito [54], una variante de RoBERTa optimizada para español, logra 70.2 en sentimientos y 55.3 en emociones, evaluados en conjuntos de datos como TASS [55] y EmoEvent [56]. Estos resultados evidencian que, si bien los modelos actuales obtienen buenos desempeños en clasificación de polaridad, el reconocimiento emocional más fino sigue siendo un desafío en el procesamiento del lenguaje afectivo en español.

Si bien los enfoques clásicos y de aprendizaje automático han permitido avances en el análisis de polaridad en español [42.43.44], el reconocimiento emocional más fino continúa siendo un desafío debido a factores como la subjetividad, ambigüedad semántica y sensibilidad cultural. En los últimos años, los modelos generativos de lenguaje (LLMs) han abierto nuevas posibilidades en clasificación emocional, especialmente mediante técnicas como EmotionPrompt [57], el aprendizaje multilingüe en pocos ejemplos y el uso de prompts estructurados. Estas estrategias han demostrado ser efectivas en tareas cero-shot (sin ajuste fino) y en lenguas con escasos recursos anotados como el español. No obstante, persisten limitaciones en la

consistencia de las salidas y la cobertura de categorías emocionales más complejas, lo que sugiere que, pese a su potencial, los LLMs aún requieren adaptaciones específicas para un análisis afectivo robusto y culturalmente sensible [58, 59].

En consecuencia, aunque existen avances relevantes en análisis afectivo para español, la mayoría de las soluciones se enfocan en polaridad y no en emociones específicas, y pocas integran análisis sobre interacciones orales [6].

2.4 Arquitectura general del sistema

El sistema propuesto para el AS y CE sigue un flujo de procesamiento secuencial y modular, comenzado desde la recepción del archivo de audio o video, y culminando con la generación de un informe consolidado. Esta arquitectura, ilustrada en la Figura 1, promueve la flexibilidad y escalabilidad al permitir la futura integración de nuevas herramientas o mejoras en componentes específicos sin afectar a la totalidad del proceso.

La solución está diseñada para procesar diálogos que involucran múltiples participantes con locuciones distinguibles en español, aceptando un archivo de audio o video en formatos .mp3, .mp4 y .wav y devolviendo como resultado un archivo PDF con los resultados. Sus módulos principales incluyen la transcripción automática de voz a texto (Speech-to-Text), la diarización¹, la fusión y segmentación, el preprocesamiento del texto, el AS y CE, el análisis individual por participante, el análisis global y la generación de reportes.

¹ Diarización: proceso automático que segmenta una grabación de audio para identificar y distinguir cuándo y quién habla en una conversación con múltiples participantes, sin necesidad de conocer previamente sus identidades.

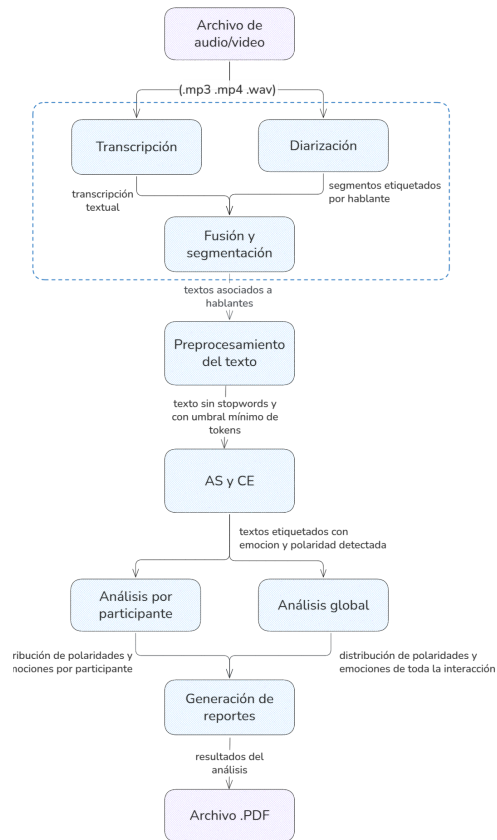


Figura 1. Diagrama de arquitectura modular

A continuación, se resumen las etapas principales del sistema:

- **Entradas aceptadas:** Archivo de audio y video.
- **Salida:** Documento en formato PDF con los resultados consolidados del análisis.
- **Módulos principales:**
 - Transcripción automática de voz a texto (Speech-to-Text).
 - Diarización para segmentar y etiquetar hablantes.
 - Fusión y segmentación, para asociar los fragmentos de texto con los hablantes identificados.
 - Preprocesamiento del texto, para optimizar la calidad del texto y mejorar la precisión del análisis
 - AS y CE mediante modelos de PLN.

- Análisis por participante.
- Análisis global de la interacción.
- Generación de reportes con resultados detallados.

2.5 Alcance de la solución propuesta

El sistema acepta un archivo en los formatos MP3, MP4 y WAV, sin límite de duración técnica. Los archivos deben contener diálogos entre dos o más participantes con locuciones claramente diferenciables, y la lengua predominante debe ser el español.

El sistema produce un informe automático en formato PDF que incluye:

- Gráficos de sentimientos y emociones, tanto por participante como para toda la interacción.
- Nube de palabras y listado de palabras frecuentes.
- Distribución temporal de los sentimientos, tanto a nivel global como por participante.
- Tabla con las intervenciones de cada participante junto con su AS y CE.

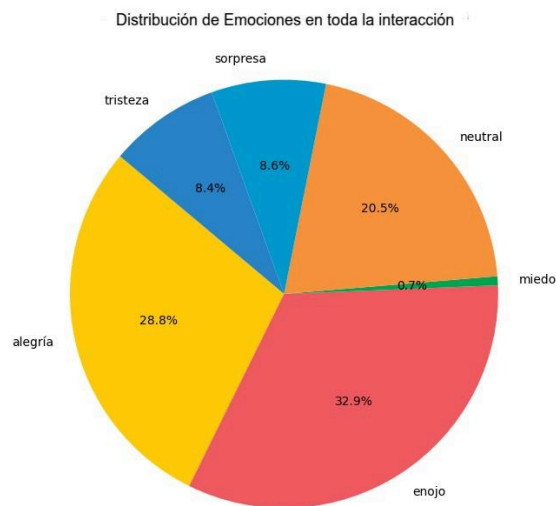


Figura 2. Distribución de emociones en toda la interacción

Participante	Duración (seg)	Texto	Sentimiento	Emocion
Scarlet	3.481	Amor, nunca dijimos eso. Tal vez fue tu suposición, pero nunca lo expresamos.	NEU	alegría
Adam	3.741	No sé cuándo lo dijimos, pero lo dijimos. Yo creí... ¿Por aquella vez por teléfono?	NEU	sorpresa

Tabla 3. Ejemplo de tabla de interacción por participante.

2.6 Desafíos

Variabilidad del Habla: La precisión del sistema de RAH puede verse afectada por la alta variabilidad en el habla, especialmente en idioma español, caracterizado por una gran diversidad dialectal y regional. Entre los factores que dificultan la transcripción precisa se incluyen:

- **Acentos regionales:** Un mismo término puede pronunciarse de manera distinta en distintas regiones, generando ambigüedades fonéticas que reducen la exactitud del reconocimiento.
- **Velocidad del habla:** Ritmos de habla inusualmente rápidos o lentos dificultan la segmentación y alineación adecuada del texto.
- **Estilo de habla informal o entrecortado:** Intervenciones espontáneas, titubeos o frases incompletas pueden generar inconsistencias en la transcripción.

Ruido de Fondo: La presencia de ruido ambiental o interferencias sonoras durante la grabación del audio puede afectar considerablemente la calidad de la transcripción. Esto es especialmente crítico en entornos no controlados como entrevistas, clases o reuniones en espacios abiertos.

Identificación de Participantes (diarización): Diferenciar correctamente a los hablantes dentro de una grabación es una tarea compleja, sobre todo cuando existen solapamientos de voz, cambios de tono o poca diferencia entre

timbres vocales. La diarización incorrecta impacta directamente en la asignación de emociones a cada participante.

Precisión en la Detección de Emociones: La detección automática de emociones presenta limitaciones, principalmente debido a la ambigüedad de las expresiones en el texto transcrito y a la escasa adaptación de los modelos de PLN al idioma español, lo que afecta su rendimiento en contextos informales o multiculturales.

Resumen Emocional Global: Uno de los desafíos más significativos es la generación de un resumen emocional global que capture el tono predominante de toda la interacción. En este contexto, es importante aclarar que el término "tono" se refiere al contexto emocional de la interacción y no al tono de voz. Los principales retos incluyen:

- **Integración de emociones individuales:** Combinar las emociones detectadas en cada intervención para obtener una visión global coherente.
- **Determinación de la emoción predominante:** Diseñar una métrica que permita identificar la emoción y polaridad predominantes en toda la interacción, teniendo en cuenta factores como la duración de las intervenciones y la frecuencia de las emociones detectadas.

3. Investigación de tecnologías

3.1 Reconocimiento automático de voz

El RAH ha experimentado avances significativos en los últimos años gracias a modelos basados en redes neuronales profundas, especialmente aquellos entrenados con grandes volúmenes de datos. Herramientas como Whisper, Google Speech-to-Text, Azure Speech y Amazon Transcribe destacan por su capacidad de transcribir audio a texto con alta precisión, especialmente en inglés.

Google Speech-to-Text. Desarrollada por Google Cloud, esta herramienta ofrece soporte para más de 125 idiomas y variantes. Su infraestructura permite transcripciones en tiempo real y sobre grabaciones preexistentes, con

funcionalidades avanzadas como diarización de locutores, reconocimiento en ambientes ruidosos y personalización de vocabulario mediante modelos adaptativos [12]. Su facilidad de integración vía API y su rendimiento en contextos profesionales la han convertido en una de las soluciones más utilizadas. No obstante, el costo está asociado a la cantidad de minutos de audio transcritos, lo que puede representar un impacto relevante en los presupuestos de proyectos a gran escala.

Amazon Transcribe. Esta herramienta [13] también permite transcripciones tanto en tiempo real como en archivos pregrabados. A pesar de ofrecer soporte para más de 50 idiomas, su cobertura lingüística es algo más limitada que la de Google. Incluye características como marcas de tiempo por palabra, diarización y una modalidad médica que reconoce terminología especializada. Adicionalmente, permite glosarios personalizados para adaptar los resultados al dominio del usuario. Sin embargo, su precisión puede disminuir en idiomas poco frecuentes o con acentos no estándar.

Azure Speech Service. Parte de los servicios cognitivos de Microsoft [14], proporciona soporte para más de 100 idiomas y variantes, con modelos personalizables para contextos específicos. A las capacidades estándar de transcripción y diarización, se suman funciones como detección de intenciones y traducción automática, lo que permite su uso en sistemas conversacionales complejos. Su integración nativa con otros servicios de Azure lo vuelve especialmente versátil, aunque los costos dependen del tipo de uso y de la región geográfica.

Whisper. Desarrollado por OpenAI, Whisper (v2) [15] está entrenado en 680000 horas de datos supervisados multitarea y multilingüaje recolectados en la web. Gracias a esta base de entrenamiento diversa y masiva, presenta una notable robustez frente a acentos, ruido de fondo y lenguaje técnico. Su arquitectura de tipo transformer encoder-decoder procesa el audio dividiéndolo en segmentos de 30 segundos, que se convierten en espectrogramas log-Mel antes de ser decodificados a texto. Whisper permite transcripción multilingüe y traducción al inglés, pero carece de funciones nativas de diarización y requiere recursos computacionales significativos para su ejecución eficiente.

WhisperX. Es una extensión del modelo Whisper, diseñada específicamente para mejorar la precisión y eficiencia en la transcripción de audio mediante capacidades avanzadas de alineación temporal y manejo de segmentos complejos. Esta herramienta [16, 17] combina la potencia del reconocimiento de voz de Whisper con técnicas de forced alignment, que permiten obtener marcas de tiempo altamente precisas para cada palabra en el audio. Por esta razón, resulta ideal para aplicaciones que requieren un análisis detallado del contenido vocal, tales como el AS y CE en interacciones multimedia o la generación de subtítulos sincronizados.

Además, WhisperX soporta múltiples idiomas y puede procesar archivos de audio de calidad variable, integrando capacidades de diarización para separar voces de diferentes participantes en una conversación. Su arquitectura modular permite una fácil integración con sistemas de análisis de texto, lo que la hace especialmente útil para proyectos que combinan procesamiento de voz y texto en flujos de trabajo más amplios. En comparación con herramientas tradicionales, WhisperX destaca por su precisión en ambientes ruidosos y su adaptabilidad a diferentes dominios lingüísticos.

A continuación, se presenta una tabla comparativa que resume las características principales de las herramientas descritas:

Herramienta	Idiomas Soportados	Diarización	Adaptabilidad	Uso Principal	Ventajas Destacadas	Limitaciones
Google Speech-to-Text	Más de 125 idiomas y variantes	Sí	Alta. Permite adaptar el reconocimiento mediante vocabularios personalizados y modelos de contexto.	Transcripciones en tiempo real y pregrabadas	Alta precisión en múltiples entornos. Integración sencilla vía API. Amplio soporte multilingüe.	Costos asociados a la duración del audio. Impacto presupuestario en proyectos de gran escala.
Amazon Transcribe	Más de 50 idiomas	Sí	Media. Soporta glosarios personalizados para adaptar resultados al dominio del usuario.	Transcripción de audios pregrabados y en tiempo real	Marcas de tiempo precisas. Soporte para terminología médica especializada [18].	Cobertura limitada para idiomas menos comunes o acentos no estándar.

Azure Speech Service	Más de 100 idiomas	Sí	Alta. Permite entrenamiento de modelos personalizados con datos específicos del dominio.	Transcripción, análisis de intenciones y traducción	Integración con otros servicios de Azure. Versátil para sistemas conversacionales.	Costos variables según uso y región. Requiere infraestructura Azure para integración completa.
Whisper	Más de 100 idiomas	No (nativo)	Baja. No cuenta con mecanismos de personalización directa, aunque su entrenamiento diverso le otorga cierta robustez.	Transcripciónes offline	Código abierto. Alta precisión en entornos ruidosos.	Sin diarización nativa. Altos requerimientos computacionales (especialmente sin GPU).
WhisperX	Más de 100 idiomas	Sí	Media. No permite personalización explícita, pero su alineación precisa y robustez multilingüe mejoran su desempeño en distintos dominios.	Transcripción con alineación precisa	Marcas de tiempo detalladas. Soporte multilingüe.	Requiere instalación manual de dependencias, configuración específica de modelos de alineación y uso preferente de GPU para un rendimiento óptimo.

Tabla 4. Comparación de diferentes herramientas de transcripción de voz.

3.2 AS y CE

El AS y CE han evolucionado significativamente gracias al desarrollo de diversas herramientas y modelos diseñados para adaptarse a diferentes idiomas, contextos y niveles de complejidad. Estas técnicas resultan clave para interpretar matices afectivos en el lenguaje humano y se han convertido en componentes esenciales en aplicaciones como monitoreo de redes sociales, atención al cliente o evaluación de dinámicas grupales.

Entre las bibliotecas de código abierto, PySentimiento [2,3] se destaca por su capacidad de identificar tanto polaridad como emociones básicas. Esta herramienta se basa en modelos de arquitectura Transformer, como BETO (una versión de BERT entrenada en español) y RoBERTa, y ofrece soporte

adicional para inglés, italiano y portugués. Su diseño accesible permite aplicar AS y CE de forma eficiente sobre texto plano y datos transcritos automáticamente.

En paralelo, herramientas como TextBlob [19] y VADER [20], proporcionan enfoques simplificados para el análisis de polaridad en inglés, siendo especialmente útiles en contextos de texto corto o en redes sociales.

En cuanto a los modelos de lenguaje profundo, tecnologías basadas en arquitecturas transformer como BERT [9] y LLaMA [21] permiten una comprensión contextualizada y detallada del lenguaje natural. Estos modelos, preentrenados con grandes volúmenes de datos y adaptables mediante fine-tuning, son capaces de capturar matices emocionales complejos en diferentes idiomas. Un caso particular es el modelo roberta-base-go_emotions [22], entrenado específicamente para clasificar un amplio espectro de 27 emociones en inglés, lo que lo convierte en una herramienta potente para análisis emocionales detallados.

En el ámbito comercial, servicios como Google Cloud Natural Language API [23], Azure Text Analytics [24] y Amazon Comprehend [25] ofrecen soluciones escalables y multilingües, facilitando la integración en sistemas empresariales. Sin embargo, su adopción puede verse limitada por los costos asociados.

Adicionalmente, existen soluciones como Gemini (Google DeepMind) [26] y DeepSeek [27], que, aunque no están diseñadas exclusivamente para el AS y CE, incorporan capacidades avanzadas de interpretación contextual, generación conversacional y búsqueda semántica, ampliando el espectro de aplicaciones para la interpretación emocional y semántica en textos.

A continuación, se presenta una tabla comparativa con las características principales de estas herramientas y modelos:

Modelo/ Herramienta	Idiomas soportados	Descripción	Ventajas	Desventajas
PySentimiento	Español, Inglés, Italiano, Portugués	Biblioteca especializada en AS y CE, optimizada para español.	Analiza tanto la polaridad como la emoción.	Soporte limitado y menor precisión en idiomas distintos al español.

			Código abierto y fácil integración.	
BERT	Multilingüe	Modelo de lenguaje profundo para NLP, adaptable mediante fine-tuning para análisis emocional.	Alta precisión en benchmark GLUE y tareas emocionales. Fine-tuning mejora resultados	Requiere fine-tuning específico para cada tarea. Alto uso computacional (CPU/GPU).
LLaMA	Multilingüe	Modelo LLM abierto y flexible para tareas NLP diversas.	Buen desempeño en benchmarks de comprensión de lenguaje.	Necesita fine-tuning para tareas específicas. Requiere recursos computacionales importantes.
TextBlob	Inglés	Biblioteca diseñada para facilitar el procesamiento de datos textuales y tareas de procesamiento del lenguaje natural.	Fácil de usar, útil para tareas básicas de PLN. [19]	Solo analiza polaridad y subjetividad en inglés.
VADER	Inglés	Lexicón y herramienta basada en reglas para análisis de sentimiento, optimizada para redes sociales.	Preciso en textos cortos y redes sociales [20].	No detecta emociones específicas, solo polaridad.
roberta-base-go_emotions	Inglés	Modelo RoBERTa entrenado con GoEmotions para clasificación de 27 emociones en inglés.	Amplia detección de emociones específicas en inglés.	Limitado a inglés. No analiza la polaridad.
Google Cloud NLP	Multilingüe	Servicio en la nube para análisis de texto con capacidades de sentimiento y entidad.	Escalable, fácil integración vía API, soporte multilingüe.	Costos y dependencia de la nube.
Azure Text Analytics	Multilingüe	Plataforma en la nube para análisis de texto con análisis de sentimientos y entidades.	Integración con Microsoft, escalable.	Costos y dependencia de la plataforma.
Amazon Comprehend	Multilingüe	Servicio en la nube para análisis de texto con enfoque en sentimiento y entidades.	Escalable, integración sencilla en el ecosistema AWS.	Costos y dependencia del servicio.
DeepSeek	Multilingüe	Plataforma especializada que combina modelos de lenguaje	Alta velocidad y eficiencia para	No está diseñado específicamente

		con tecnologías para búsqueda semántica.	análisis en tiempo real.	para clasificación emocional. Plan gratuito limitado.
Gemini	Multilingüe	Modelo generativo avanzado con capacidades conversacionales y razonamiento contextual.	Excelente comprensión contextual.	No está diseñado específicamente para clasificación emocional. Plan gratuito limitado.

Tabla 5. Comparación de diferentes herramientas/modelos de AS y CE.

3.3 Datasets

Para nuestro trabajo de AS y CE consideramos dos opciones principales al seleccionar datasets:

- Datasets con las emociones clasificadas:
 - Audios de una sola persona hablando [28]
 - Extractos de películas [29]
- Datasets sin clasificar: En esta opción, utilizamos múltiples herramientas de clasificación de sentimientos y emociones para analizar los datos. Aunque no se pueden validar objetivamente, esta estrategia ofrece resultados más fiables al combinar varias aproximaciones. Entre los datasets destacados se incluyen:
 - Conversaciones entre dos o más personas [30]
 - Grabaciones de una persona en Español argentina [31]
 - Extractos de TED talks en Español [32]

Finalmente, optamos por la segunda opción debido a la flexibilidad que ofrece al trabajar con datos de diversos contextos y escenarios. Este enfoque permite explorar cómo las herramientas de análisis emocional responden en situaciones reales y no estructuradas. Para enriquecer el alcance y aplicabilidad del sistema, seleccionamos audios de los datasets mencionados y los complementamos con fragmentos adicionales de series, películas y otras fuentes relevantes en español.

3.4 Entorno de ejecución

Para llevar a cabo este proyecto, seleccionamos Google Colaboratory [33] como entorno principal debido a su capacidad para proporcionar recursos computacionales avanzados de forma gratuita. Este servicio alojado de Jupyter Notebook permite realizar tareas intensivas como la transcripción y el AS y CE de manera eficiente.

Recursos asignados:

- **GPU NVIDIA Tesla T4:** 16 GB de memoria GDDR6, ideal para la inferencia de modelos y la aceleración de cargas de trabajo intensivas.
- **RAM del sistema:** Hasta 12 GB disponibles en Google Colab.
- **Frameworks compatibles:** PyTorch y TensorFlow, garantizando la integración óptima con herramientas como WhisperX y modelos de AS y CE.

Este entorno no requiere configuración previa y está diseñado para maximizar la eficiencia en proyectos que incluyen PLN, generación de transcripciones y evaluación emocional. La elección de Google Colab asegura un equilibrio entre accesibilidad y rendimiento, permitiendo manejar volúmenes de datos significativos con rapidez y precisión.

A pesar de sus numerosas ventajas, Google Colaboratory presenta ciertas limitaciones relevantes para su uso. La plataforma opera bajo un modelo de recursos compartidos, lo que implica que la disponibilidad de hardware especializado, como unidades de procesamiento gráfico (GPU) y memoria RAM, pueden variar en función de la demanda global del sistema [33]. Las sesiones de ejecución tienen una duración máxima estimada de 12 horas, aunque este límite puede fluctuar según el uso del sistema y otras condiciones dinámicas [33]. Además, las sesiones pueden finalizar de manera anticipada si se detecta inactividad o si el entorno es utilizado fuera de los lineamientos esperados, como cuando se privilegia la interacción con interfaces web por sobre la ejecución directa de código [33]. A esto se suma que el funcionamiento en las máquinas virtuales es transitorio, es decir, al finalizar la sesión todos los datos no guardados en un almacenamiento externo son eliminados. Estas características obligan a una planificación cuidadosa

para preservar la integridad del trabajo, especialmente en proyectos que requieren persistencia de datos y ejecución prolongada [34].

3.5. Herramientas elegidas

Transcripción de texto

WhisperX (v3.3.1) fue seleccionado como la herramienta principal para la transcripción automática de audio/video en este proyecto debido a su conjunto único de características que se alinean perfectamente con las necesidades de nuestra investigación:

1. **Capacidades de Diarización:** WhisperX incluye funcionalidades avanzadas para la diarización, es decir, la identificación y separación de los distintos participantes en un archivo de audio. Esto resulta crucial para nuestro objetivo de asociar emociones y polaridades específicas a cada orador, permitiendo un análisis detallado a nivel individual [16].
2. **Código Abierto y Gratuito:** Al ser una herramienta de código abierto y de acceso gratuito, WhisperX ofrece la flexibilidad de ser adaptado y personalizado según los requerimientos específicos de nuestra herramienta. Esto también elimina barreras económicas que otras soluciones comerciales podrían presentar, facilitando su implementación en un contexto de investigación [17].
3. **Soporte Multilingüe:** Una característica clave de WhisperX es su capacidad de trabajar eficazmente con múltiples idiomas, incluido el español. Esto es especialmente relevante, dado que una gran parte de las investigaciones existentes se centran en inglés, dejando a las lenguas de pocos recursos menos exploradas. WhisperX llena este vacío, permitiendo que nuestro sistema procese contenido en español con alta precisión [17].
4. **Precisión en Reconocimiento de Voz:** WhisperX se basa en el modelo Whisper de OpenAI, conocido por su alta precisión en la transcripción de audio, incluso en condiciones de ruido o con acentos variados. Esta robustez es esencial para garantizar transcripciones de calidad que sirvan como base confiable para el AS y CE [16].

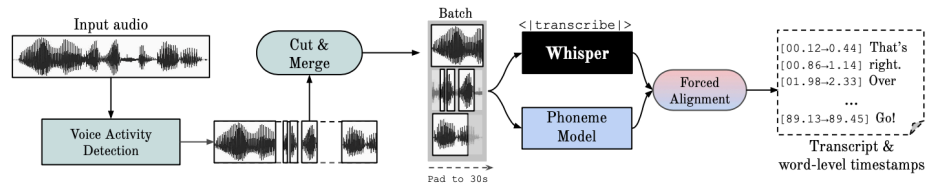


Figura 3. Flujo de procesamiento de audio con WhisperX [16, 17].

Modelos de Lenguaje Natural (LLMs)

Para el AS y CE sobre los textos transcritos, el sistema se apoya en dos herramientas principales:

- **PySentimiento** (v0.7.3): biblioteca especializada en AS y CE, con soporte específico para el idioma español [3].
- **roberta-base-go_emotions**: modelo basado en RoBERTa entrenado con el dataset GoEmotions. Aunque está orientado al inglés, se utilizó como complemento para ampliar la cobertura emocional en casos específicos [4].

Ambos modelos fueron seleccionados por su accesibilidad gratuita, compatibilidad con el idioma español (directa o mediante traducción), y facilidad de implementación en entornos como Google Colab.

Modelos utilizados como referencia

Además de los modelos utilizados en el sistema, se recurrió a otros modelos de lenguaje de gran escala con el fin de validar y comparar los resultados obtenidos. Estos modelos no forman parte del sistema de análisis, sino que se usaron como punto de referencia externo: **Gemini** [26], **LLaMA** [21] y **DeepSeek** [27].

Estos modelos fueron empleados para clasificar polaridades y emociones en los mismos fragmentos analizados por el sistema, con el objetivo de evaluar el grado de concordancia entre herramientas. Se eligieron por su disponibilidad gratuita para consultas limitadas, su alto rendimiento en tareas de

comprensión contextual y su capacidad multilingüe. La descripción completa del procedimiento comparativo puede consultarse en la sección 6.

4. Diseño de la solución

El flujo de trabajo propuesto consta de etapas secuenciales y modulares que permiten realizar un análisis detallado del contenido emocional en archivos de audio o video. Tal como se muestra en el esquema general presentado en la Sección 2.4 “*Arquitectura general del sistema*”, el sistema se compone de distintos módulos que operan de forma complementaria para facilitar la extracción, segmentación, clasificación y resumen de la información emocional.

Este diseño modular aporta claridad y flexibilidad, facilitando un procesamiento escalonado y la integración eficiente de cada etapa. A continuación, se describen brevemente las funciones principales de cada módulo:

- **Módulo de transcripción de voz:** Convierte el archivo de audio o video en texto, generando una transcripción escrita que sirve de base para el análisis posterior.
- **Módulo de diarización:** Identifica y distingue los diferentes hablantes en el audio o video, asignando etiquetas únicas a cada uno.
- **Módulo de fusión y segmentación:** Combina los resultados de la transcripción y la diarización, organizando los textos en segmentos claramente asociados a cada hablante.
- **Módulo de preprocesamiento de texto:** Limpia y filtra los segmentos de texto eliminando stopwords y descartando aquellos con menos de cuatro tokens. Esto reduce el ruido y mejora la calidad del análisis, preservando sólo fragmentos con contenido semántico relevante.
- **Módulo de AS y CE:** Procesa el texto segmentado para detectar emociones y polaridades, generando datos etiquetados que reflejan el contenido emocional de cada intervención.
- **Módulo de análisis por participante:** Analiza individualmente las emociones y polaridades de cada participante, identificando las predominantes durante su participación.

- **Módulo de análisis global:** Integra la información emocional de todos los participantes para ofrecer una visión global de la interacción, destacando las emociones y polaridades predominantes en toda la conversación.
- **Módulo generador de reportes:** Consolida los resultados de los análisis individuales y globales en un documento estructurado, que incluye gráficos y tablas, facilitando la interpretación y presentación de los resultados al usuario final.

4.0 Evolución de la herramienta y justificación del rediseño

La versión inicial del sistema se apoyaba en la transcripción y diarización automática mediante **WhisperX**, y en el AS y CE en español utilizando **PySentimiento**. Si bien WhisperX demostró ser una herramienta robusta para extraer y segmentar el texto hablado —y por ello se mantuvo en el sistema rediseñado—, se identificaron limitaciones significativas en el desempeño del análisis emocional, especialmente en cuanto a precisión y concordancia con modelos de referencia.

Para cuantificar estas limitaciones, se calculó el coeficiente Kappa de Fleiss [35] con el fin de medir el acuerdo entre los resultados de la herramienta y dos sistemas de referencia avanzados (LLaMA y Gemini). Los valores obtenidos fueron los siguientes:

Métrica	Valor de Kappa de Fleiss
Polaridad Predominante por Participante	0.5205
Polaridad Predominante en la Conversación	0.5878
Emoción Predominante por Participante	-0.0207
Emoción Predominante Durante Toda la Interacción	-0.1368

Tabla 6. Coeficiente de Kappa de Fleiss para la comparación de modelos en la detección de polaridad y emoción.

Los resultados muestran un nivel de acuerdo bajo a moderado en la clasificación de polaridad, y muy bajo (incluso negativo) en la identificación de emociones predominantes. Esto evidencia una falta de precisión en la

versión original del sistema, especialmente en la detección de emociones más sutiles o en contextos de ambigüedad semántica.

Ante estas limitaciones, se propuso un rediseño del sistema, incorporando mejoras destinadas a optimizar tanto la calidad del texto analizado como la robustez del análisis emocional. Las principales modificaciones incluyen:

- **Preprocesamiento textual avanzado:** eliminación de stopwords, normalización de expresiones y filtrado de intervenciones con baja densidad léxica para reducir el ruido en los datos de entrada.
- **Análisis complementario en inglés:** traducción y reevaluación de segmentos ambiguos, clasificados como “neutros” u “otros”, mediante modelos entrenados en inglés, con el fin de aprovechar la mayor disponibilidad de recursos en ese idioma.
- **Ponderación temporal:** se implementó un mecanismo de cálculo de emociones y polaridades predominantes basado en la duración del tiempo hablado, en lugar de considerar únicamente la frecuencia de aparición. Este enfoque asigna mayor peso a las emociones expresadas durante periodos más extensos, permitiendo una representación más precisa de la carga emocional.
Por ejemplo, si un participante manifiesta una emoción durante 60 segundos y otra diferente durante solo 10 segundos, la primera será considerada predominante, independientemente del número de intervenciones en que se haya expresado.

Estas modificaciones permitieron mejorar la detección emocional, reducir inconsistencias en las clasificaciones y aumentar la concordancia con modelos de referencia, generando resultados más coherentes con el contenido emocional real de las interacciones.

4.1 Transcripción, Diarización y Fusión/Segmentación del habla

La primera etapa del proceso consiste en la transcripción automática del audio a texto mediante WhisperX, que además realiza la diarización automática para identificar y distinguir a los diferentes hablantes, asignándoles etiquetas genéricas como SPEAKER_00, SPEAKER_01, entre otras. Esta funcionalidad es fundamental para estructurar la información y asignar

correctamente las intervenciones a cada participante, facilitando un AS y CE segmentado y personalizado.

WhisperX detecta los segmentos activos de habla mediante un modelo especializado que identifica momentos de actividad vocal significativa. Estos segmentos, cuya duración puede variar, se limitan mediante técnicas como la operación de min-cut para no superar aproximadamente los 30 segundos, asegurando un procesamiento eficiente. Luego, el sistema agrupa segmentos que corresponden al mismo hablante analizando características acústicas como tono, timbre y otras propiedades vocales, garantizando una diarización precisa.

Para mejorar la coherencia y continuidad en el AS y CE se agrupan todas las frases consecutivas emitidas por un mismo hablante en unidades semánticas completas, consolidando texto, tiempos de inicio y fin, y palabras asociadas para generar bloques significativos.

Finalmente, se selecciona un fragmento representativo de audio para cada participante, utilizando la librería Pydub [47], facilitando que el usuario pueda editar manualmente las etiquetas asignadas automáticamente y reemplazar las identificaciones genéricas por nombres personalizados.

4.2 Preprocesamiento del texto

Antes de realizar el AS y CE, es fundamental llevar a cabo un preprocesamiento del texto para mejorar la calidad de los datos y la precisión de los modelos de clasificación. Este preprocesamiento incluye:

- **Eliminación de stopwords:** Las stopwords son palabras vacías o funcionales que aparecen con mucha frecuencia en el lenguaje (como preposiciones, artículos y conjunciones: “el”, “de”, “y”, “que”, etc.) y que por sí solas no aportan información semántica relevante para el análisis emocional. Para este fin, se utiliza la librería de Python NLTK (Natural Language Toolkit) [36], que ofrece las listas estándares de stopwords en varios idiomas, incluyendo el español.
- **Filtrado por número mínimo de tokens:** Luego de eliminar las stopwords, se descartaron las frases que contienen menos de 4 tokens (palabras). Este umbral se estableció tras analizar tres conjuntos

distintos de transcripciones, evaluando cómo variaban las distribuciones de polaridad y emociones en función de la longitud mínima del texto.

En todos los casos, se observó que las frases con menos de cuatro palabras producían resultados menos confiables: aumentaba la proporción de etiquetas neutras o ruidosas (como “otros” en emoción), y se debilitaba la diferenciación entre participantes. Por el contrario, al aplicar un umbral de 4 tokens, las distribuciones se estabilizaban, permitiendo identificar emociones más definidas y polaridades claramente diferenciadas entre los hablantes.

Estas observaciones pueden verse en las Figuras 4 y 5, que corresponden al análisis de uno de los tres archivos evaluados. En la Figura 4 se muestran las distribuciones de polaridad y emociones antes del filtrado, mientras que en la Figura 5 se presentan los resultados después de aplicar el umbral mínimo de cuatro tokens. Se evidencia una reducción significativa del ruido y una mejora en la discriminación emocional entre los participantes.

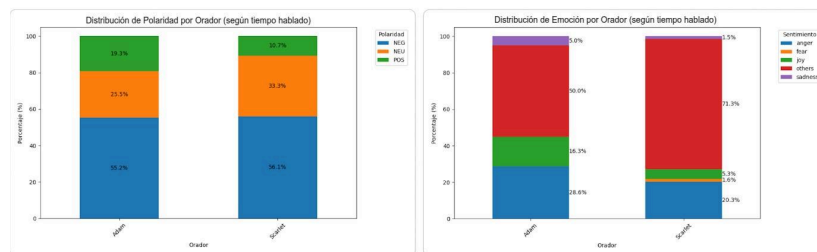


Figura 4. Distribución de polaridad y emociones sin la eliminación de stopwords.

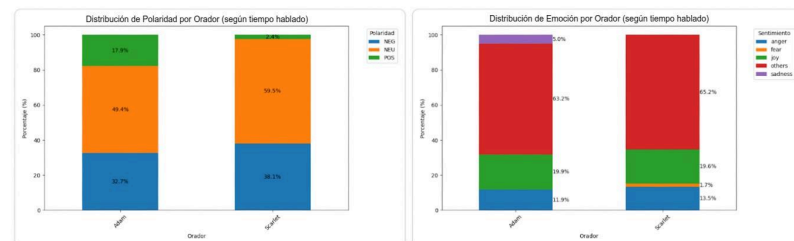


Figura 5. Distribución de polaridad y emociones con la eliminación de stopwords y filtrado por umbral mínimo de 4 tokens.

Además, con este umbral se logra conservar entre el 68% y el 74% de los textos totales en los conjuntos analizados, manteniendo una adecuada representatividad del corpus. Umbrales más bajos conservaban más datos pero introducían ruido, mientras que umbrales más altos comenzaban a excluir información emocionalmente relevante.

Este proceso de limpieza y filtrado reduce el ruido en los datos textuales y optimiza el rendimiento de los modelos de análisis de sentimientos y emociones.

4.3 AS y CE

4.3.1 Análisis en español

Sobre el texto procesado, se lleva a cabo el análisis de polaridades y emociones con PySentimiento. Este modelo clasifica cada frase en dos dimensiones principales:

- **Polaridad:** Determina si la carga emocional de la frase es positiva (POS), neutral (NEU) o negativa (NEG).
- **Emoción:** Identifica una de las emociones básicas propuestas por el psicólogo Paul Ekman [46], quien, a partir de estudios transculturales, estableció que ciertas emociones son universales y reconocibles en distintos contextos culturales y lingüísticos. En los casos en que la emoción no coincide con ninguna de las categorías propuestas, la herramienta la clasifica como “others” (neutral).

El sistema adoptado reconoce las siguientes emociones, extraídas del modelo de Ekman:

Emoción	Descripción
Alegría	Expresión de felicidad o satisfacción.

Tristeza	Sensación de pena o melancolía.
Enojo	Respuesta a una injusticia o frustración.
Miedo	Reacción ante un peligro o amenaza.
Sorpresa	Emoción ante un evento inesperado.
Disgusto	Rechazo o aversión a una situación o estímulo.

Tabla 7. Emociones básicas universales según Ekman.

4.3.2 Análisis Complementario en Inglés para Frases Ambiguas

En los casos en los que la polaridad resulta NEU o la emoción corresponda a la categoría “others”, se aplica una segunda fase de análisis en inglés para mejorar la precisión del reconocimiento emocional. Este proceso consta de los siguientes pasos:

1. Se traduce la transcripción original a inglés utilizando la librería **Deep-Translator** con el motor de Google [48].
2. Eliminación de stopwords en inglés, empleando la librería NLTK, seguida de un nuevo filtrado para descartar frases con menos de cuatro tokens.
3. Re-análisis de polaridad utilizando **PySentimiento** en su versión inglesa.
4. Análisis de emociones mediante el modelo **roberta-base-go_emotions**, entrenado con el dataset GoEmotions de Google [10], que permite identificar una amplia gama de emociones.

Los resultados obtenidos con este último modelo son posteriormente mapeados a las seis emociones básicas utilizadas por PySentimiento para mantener la coherencia analítica en el sistema. Para ello, se aplica el esquema de agrupación propuesto en el paper de GoEmotions, garantizando así la comparabilidad y consistencia entre ambos modelos.

4.4 Cálculo Ponderado de Emociones y Polaridades

Para determinar la emoción predominante de cada participante, se emplea una métrica que pondera la duración del tiempo hablado en cada categoría emocional, en lugar de basarse únicamente en la frecuencia de ocurrencia. Esto significa que las emociones expresadas durante períodos más largos tienen un peso mayor en la clasificación, reflejando con mayor precisión la carga emocional real de cada intervención.

El procedimiento consiste en agrupar las intervenciones por participante y emoción, sumando el tiempo total en el que cada emoción está presente. Luego, se calcula el porcentaje que representa cada emoción respecto al total del tiempo hablado por el participante. La emoción predominante será aquella con el mayor porcentaje acumulado.

De forma similar, para determinar la polaridad predominante en toda la interacción, se suman las duraciones de cada polaridad considerando a todos los participantes y se identifica la polaridad que representa la mayor proporción del tiempo total.


Este enfoque ponderado por duración permite obtener un análisis más fiel y representativo de las dinámicas emocionales en la conversación.

4.5 Prototipo de Interfaz de Usuario


Como parte del diseño conceptual del sistema, se desarrolló un mockup de la interfaz de usuario con el objetivo de anticipar cómo será la interacción con la herramienta una vez implementada. Este prototipo visual simula el flujo de trabajo esperado por el usuario, desde la carga del archivo hasta la visualización de los resultados.

La interfaz está organizada en etapas claramente definidas: primero, permite cargar un archivo de audio o video en formatos como MP3, MP4 o WAV, ya sea mediante un botón de búsqueda o arrastrando el archivo a la pantalla. Además, ofrece la opción de ingresar manualmente el número de participantes en la conversación, con el fin de optimizar el rendimiento del sistema de diarización automática (Figura 6). Luego, se brinda la posibilidad de editar las

etiquetas de los hablantes detectados, reemplazando los nombres genéricos por identificaciones personalizadas, acompañadas de un fragmento de audio representativo que facilita la identificación de cada participante (Figura 7). Finalmente, se muestran los resultados del análisis emocional a través de distintos gráficos, que detallan la polaridad y las emociones detectadas por orador y a lo largo del tiempo, así como un resumen global de la interacción. La interfaz también incluye la opción de descargar un reporte completo en formato PDF (Figura 8). Este diseño preliminar no ha sido implementado aún, pero constituye una guía clara para futuras etapas de desarrollo enfocadas en la experiencia del usuario y la usabilidad del sistema.

 **Analizador de Sentimientos y Emociones**



Cargar Audio o Video




Arrastra y suelta tu archivo
o

[Buscar Archivo](#)

Formatos aceptados: MP3, MP4, WAV

 ejemplo-audio.mp3 

Número de participantes (opcional)

 **Advertencia**

Este sistema utiliza WhisperX (v3.3.4) para la diarización automática de participantes.
Si no se indica correctamente el número de participantes, los resultados pueden ser inexactos o erróneos.


 **Generar Reporte**

Figura 6. Pantalla principal del sistema de análisis de sentimientos y emociones.

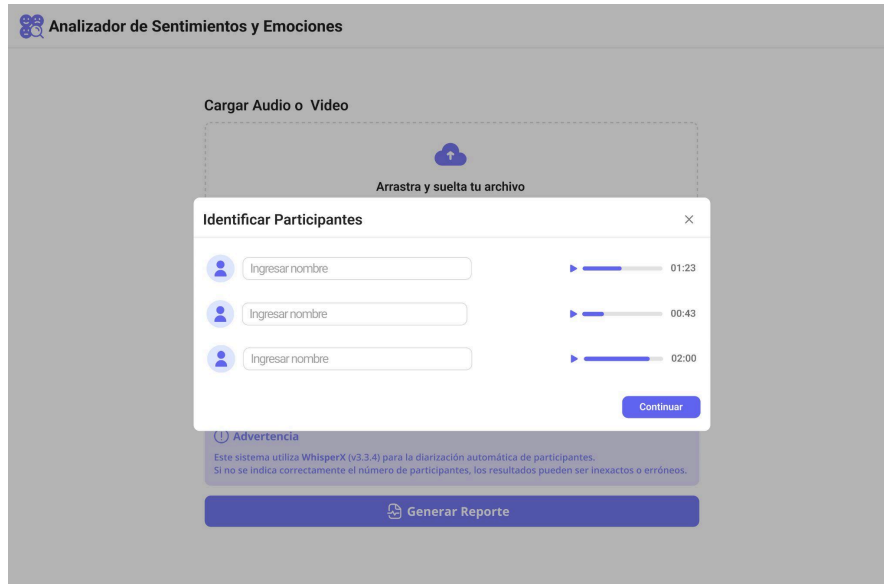


Figura 7. Identificación de participantes luego del procesamiento del audio.

Analizador de Sentimientos y Emociones - Informe





Figura 8. Reporte con los resultados del AS y CE.

5. Implementación de la solución

La implementación de la solución se encuentra disponible en el siguiente repositorio de GitHub: [Analizador de Sentimientos y Emociones](#) [60]

6. Verificación y validación

Archivo	Descripción
 spreadsheet	Archivo de pruebas
 KappaDeFleiss - O...	Archivo de pruebas - Optimización de la herramienta

Transcripción y Diarización

Para evaluar la precisión de las transcripciones y diarizaciones generadas por WhisperX, se realizó una comparación con transcripciones de referencia obtenidas de los audios y videos suministrados a la herramienta. La mayoría de los fragmentos utilizados corresponden a películas y series, seleccionados por la accesibilidad a sus guiones o subtítulos, lo que facilitó la validación de resultados.

En el caso de los guiones, la disponibilidad fue limitada, particularmente en español, mientras que en inglés se encontraron en mayor cantidad. Por otro lado, el uso de subtítulos presentó la ventaja de proporcionar transcripciones preexistentes. Sin embargo, se requirió realizar manualmente la diarización, asignando cada frase al hablante correspondiente. Con el objetivo de optimizar el proceso de validación y reducir la carga manual, se decidió utilizar fragmentos de audio de corta duración.

Se realizó la prueba de obtener fragmentos de larga duración mediante la descarga de videos desde YouTube, junto con los subtítulos autogenerados proporcionados por la plataforma. Sin embargo, se observó que la calidad de estos subtítulos no era adecuada para realizar una evaluación confiable de las transcripciones, debido a errores frecuentes y falta de precisión en la transcripción del contenido.

Para cuantificar el nivel de coincidencia entre la transcripción generada por WhisperX y la referencia textual (ya sea guión o subtítulo), se desarrolló una función denominada `generar_diferencias`. Esta herramienta compara ambos textos utilizando `HtmlDiff` de la librería `difflib` [37], generando un archivo visual en formato HTML con las diferencias resaltadas. Además calcula el

porcentaje de similitud mediante el algoritmo SequenceMatcher de la biblioteca difflib. Esto permitió obtener un indicador objetivo de exactitud y facilitar el análisis de errores de transcripción de forma automatizada.

AS y CE

Una vez obtenidas y procesadas las transcripciones, se aplicaron tres modelos para el AS y CE: **DeepSeek**, **Gemini** y la **herramienta propuesta**. Cada modelo fue configurado para clasificar las emociones en seis categorías: alegría, sorpresa, disgusto, tristeza, miedo y enojo, además de determinar la polaridad del texto como positiva (POS), neutral (NEU) o negativa (NEG).

Inicialmente se había considerado el modelo LLaMA (17B) como parte del conjunto de referencia. Sin embargo, tras realizar pruebas comparativas preliminares, se observó que DeepSeek (70B) ofrecía un rendimiento superior. Ambos modelos, LLaMA y DeepSeek, fueron accedidos a través de la APIs de Groq [38], un servicio de inferencia de modelos de lenguaje optimizado para acelerar consultas sobre arquitecturas de gran tamaño. Esto permite atribuir las diferencias de desempeño exclusivamente a los modelos evaluados y no a la infraestructura de ejecución.

La decisión de reemplazar LLaMA por DeepSeek se basó en dos observaciones fundamentales. En primer lugar, DeepSeek mostró mayor coherencia y especificidad en la clasificación emocional, con respuestas más alineadas tanto con los contextos analizados como con los generados por Gemini y nuestro modelo propio, lo que sugiere una mejor sintonía conceptual entre estos sistemas. En segundo lugar, LLaMA presentó dificultades para seguir las instrucciones del prompt, devolviendo con frecuencia respuestas vacías o etiquetas ambiguas como “mixta”, en lugar de una emoción claramente definida. Dado que se utilizó el mismo prompt estandarizado para todos los modelos, estas inconsistencias justificaron su exclusión del análisis final.

Debido a las limitaciones de la versión gratuita del servicio, se incorporó un retardo de 10 segundos entre cada consulta para evitar errores por tasa de uso o restricciones de acceso.

El prompt utilizado para los modelos de lenguaje de gran escala (LLMs) fue formulado en inglés, ya que este idioma mostró un rendimiento superior respecto de su equivalente en español en pruebas preliminares. El texto exacto del prompt fue el siguiente:

“Perform emotion and polarity analysis on the following sentence: {sentence}.

Classify polarity as one of (not other): 'POS' (positive), 'NEG' (negative), or 'NEU' (neutral).

For emotion, assign exactly one of: 'neutral', 'joy', 'sadness', 'anger', 'surprise', 'disgust', or 'fear'.

The text must be the original.

Return the result in JSON format like:

{"text": "...", "polarity": "...", "emotion": "..."}".

Como parte del proceso de validación, se aplicó el coeficiente kappa de Fleiss con el fin de medir el grado de acuerdo entre los modelos utilizados. En una primera instancia, se analizó una sección de la película “Historia de un Matrimonio” de Noah Baumbach, obtenido de YouTube, comparando las salidas entre los pares de modelos: nuestro modelo vs. Gemini, Gemini vs. DeepSeek, y nuestro modelo vs. DeepSeek. Este análisis permitió identificar discrepancias relevantes y guiar una serie de ajustes al sistema desarrollado, orientados a mejorar su coherencia con los modelos de referencia.

Los principales ajustes realizados fueron los siguientes:

- Para el análisis de polaridad con PySentimiento en inglés, se eliminaron las stopwords tras la traducción, lo que mejoró la precisión.
- En cambio, para el análisis de emociones con roberta-base-go_emotions en inglés, se observó un mejor desempeño al mantener las stopwords.
- Para los LLMs, el análisis arrojó mejores resultados al eliminar las stopwords antes de enviar el texto (transcripción original en español).

Posteriormente, este procedimiento fue replicado en una serie de audios y videos adicionales. Para cada uno, se calcularon tanto la polaridad como la emoción predominante a nivel global, así como por hablante. Esta estrategia permitió validar la estabilidad del modelo propuesto en distintos contextos discursivos y fuentes de audio, destacando su capacidad para adaptarse a diferentes estructuras conversacionales sin pérdida significativa de precisión.

6.1 Métrica

Para evaluar el acuerdo entre los modelos de análisis emocional seleccionados, utilizamos Fleiss' Kappa como métrica estándar. Esta métrica nos permite medir el nivel de concordancia entre los clasificadores en tareas de categorización, asegurando que los resultados sean consistentes y confiables.

El coeficiente Kappa de Fleiss [35, 39] es una medida estadística ampliamente utilizada para evaluar la fiabilidad entre evaluadores en estudios donde se requiere categorizar observaciones en múltiples clases por parte de varios participantes.

A diferencia del coeficiente Kappa de Cohen (1960), que se limita a la comparación entre dos codificadores, el Kappa de Fleiss (1981) permite el análisis del acuerdo cuando participan 2 o más observadores. El coeficiente Kappa de Fleiss añade el cálculo del sesgo del codificador (precisión-error) y el cálculo de la concordancia (calibración).

Coeficiente Kappa de Cohen (para dos observadores):

$$k = \frac{p_0 - p_c}{1 - p_c}$$

P₀ se define como la proporción de concordancia observada realmente y se calcula sumando las marcas que representan la concordancia y dividiendo por el número total de ellas;

P_c es la proporción esperada por azar y se calcula sumando las probabilidades de acuerdo por azar para cada categoría

Coefficiente Kappa de Fleiss (para más de dos observadores):

$$\overline{K} = 1 - \frac{n m^2 - \sum_{i=1}^n \sum_{j=1}^r x_{ij}^2}{n m (m-1) \sum_{j=1}^r \overline{p_j} \overline{q_j}}$$

Los símbolos de la fórmula vienen identificados por las siguientes correspondencias:

- **n**: se corresponde con el número total de conductas o códigos a registrar;
- **m**: identifica el número de codificaciones;
- **x_{ij}**: define el número de registros de la conducta i en la categoría j;
- **r**: indica el número de categorías de que se compone el sistema nominal;
- **p**: es la proporción de acuerdos positivos entre codificadores;
- **q**: es la proporción de acuerdos negativos (no acuerdos) en codificadores (1 - p)

Diversos autores han propuesto escalas para la interpretación de los valores de Kappa. En particular, Altman (1991) establece la siguiente clasificación:

Valor de Kappa	Nivel de Concordancia
< 0.20	Pobre
0.21 - 0.40	Débil
0.41 - 0.60	Moderada
0.61 - 0.80	Buena
0.81 - 1.00	Muy Buena

Tabla 8. Interpretación del coeficiente Kappa de Fleiss

Por otro lado, Fleiss ofrece una interpretación más restringida, considerando los valores entre 0.40 y 0.60 como "regulares", entre 0.61 y 0.75 como "buenos" y superiores a 0.75 como "excelentes".

En el marco del análisis automático de sentimientos y emociones, la fiabilidad entre evaluadores es fundamental para evaluar la consistencia entre diferentes clasificadores automáticos. Por ello, en nuestra investigación se utiliza el coeficiente Kappa de Fleiss para medir el grado de acuerdo entre tres modelos — DeepSeek, Gemini y nuestra solución — en la identificación de polaridades y emociones en segmentos de audio transcritos. Esta comparación permite validar la consistencia y fiabilidad de los clasificadores en el contexto de nuestro sistema.

6.2 Resultados

Transcripción de texto

Para evaluar la precisión del sistema de transcripción y diarización utilizado, se realizaron pruebas con WhisperX sobre fragmentos de audio breves extraídos de películas y series, seleccionados por su accesibilidad a subtítulos o guiones oficiales. Estos recursos permitieron contar con una referencia confiable para verificar el contenido transcrito.

Se compararon dos configuraciones distintas del modelo:

- Una en la que se especificaba el número de participantes.
- Otra en la que no se proporcionaba esta información.

Los resultados mostraron que indicar el número de participantes mejora la calidad de la diarización, logrando asignar correctamente las intervenciones al hablante correspondiente en el 68.42% de los casos. En los casos restantes (31.58%), donde no se indicó la cantidad de hablantes, la transcripción resultó levemente menos precisa, con errores principalmente en la segmentación por orador, aunque sin afectar significativamente la fidelidad del contenido textual.

Por lo tanto, se concluye que el desempeño de WhisperX es robusto, aunque su rendimiento mejora al configurarse adecuadamente con el número de participantes.

Análisis de sentimientos y emociones

A partir de las transcripciones textuales de las interacciones analizadas, se aplicaron modelos de clasificación para determinar la polaridad (positiva, negativa o neutra) y la emoción (alegría, tristeza, miedo, etc.) asociada a cada frase. Un ejemplo del formato resultante se presenta a continuación:

	speaker	text	start (s)	end (s)	duration (seg)	polaridad	emocion
10	Paloma	La vaca es el animal más maltratado de todo el campo. Las separan de sus crías apenas nacen. Les sacan los cuernos con un descornador eléctrico que es sumamente doloroso. Las inseminan artificialmente cada dos meses. Basta de hablar de vacas.	74.07 7	86.22 1	12.14	NEG	anger
11	Manuel	¿Y a las vacas felices cómo les sacan la leche? Viene la vaca solita y le dice al tambero, tomá, Jorge, probá esta leche cortesía de la casa.	86.28 1	95.50 5	9.22	NEU	neutral
12	Grace	Manuel, no jodas a la chica. Las vacas felices pastan libremente, comen comida natural, cuidan a sus crías,	95.64 5	102.1 87	6.54	POS	joy

Tabla 9. Fragmento representativo del análisis de polaridad y emoción por segmento de habla.

Evaluación inicial de la Herramienta (versión previa al rediseño)

Tal como se detalla en la sección 4.0 (*“Evolución de la herramienta y justificación del rediseño”*), la primera versión del sistema se basó únicamente en análisis emocional en español utilizando PySentimiento. Esta configuración fue evaluada frente a modelos de referencia (Gemini y LLaMA). Los resultados evidenciaron limitaciones importantes en la detección de polaridad y un desempeño deficiente en la identificación de sentimientos predominantes, según lo reflejado en el coeficiente Kappa de Fleiss. Para detalles cuantitativos, se remite a la Tabla 6.

Estos hallazgos motivaron una revisión profunda del modelo, que incluyó la implementación de etapas de preprocesamiento más rigurosas, análisis multilingüe y métodos de ponderación más sensibles, con el fin de optimizar la precisión y coherencia del sistema.

Evaluación de la Herramienta Rediseñada

El rediseño del sistemas incluyó nuevas etapas de procesamiento (explicadas en la sección 4), lo cual permitió una mejora significativa en la precisión del análisis emocional. A continuación se presentan los resultados obtenidos al comparar nuestro modelo actualizado con dos modelos de referencia (Gemini y DeepSeek), utilizando el coeficiente Kappa de Fleiss.

Análisis de frases individuales por modelo

Se evaluó cada frase de la transcripción asignándole una polaridad y una emoción mediante tres herramientas: **nuestra herramienta**, **Gemini** y **Groq** (DeepSeek). Para medir el nivel de acuerdo entre modelos, se calculó el coeficiente kappa de Fleiss entre pares. Los modelos se agruparon por archivo, y se reportaron los promedios de k por cada par de modelos.

Comparación	Polaridad	Emoción
Nuestra herramienta	0.6709 (Bueno)	0.3050 (Débil)

vs. Gemini		
Nuestra herramienta vs. DeepSeek	0.6065 (Moderado)	0.2874 (Débil)
Gemini vs. DeepSeek	0.7718 (Bueno)	0.6781 (Bueno)

Tabla 10. Resultados de Kappa de Fleiss para el análisis de frases individuales

El análisis de consistencia entre modelos en la asignación de polaridad y emoción a nivel de frase revela diferencias importantes en el desempeño de los clasificadores. En general, se observa un mayor acuerdo en la clasificación de polaridad que en la de emociones específicas, lo cual es consistente con la literatura previa sobre análisis afectivo automatizado.

Polaridad: Los valores de Kappa de Fleiss muestran que hay un acuerdo bueno entre los modelos para la identificación de la polaridad (positiva, negativa o neutral). En particular, el mayor nivel de concordancia se da entre Gemini y DeepSeek ($k = 0.7718$), seguido por nuestro modelo en comparación con Gemini ($k = 0.6709$) y DeepSeek ($k = 0.6065$). Esto sugiere que la tarea de determinar la orientación afectiva general de cada frase es relativamente estable entre herramientas, posiblemente porque se basa en señales lingüísticas más evidentes y menos ambiguas que las que intervienen en la clasificación emocional.

Emociones: En contraste, los valores de kappa para la clasificación de emociones específicas (como alegría, tristeza, enojo, etc.) son notablemente más bajos, especialmente en las comparaciones que involucran nuestro modelo ($k \approx 0.29\text{--}0.30$). Estos resultados indican un nivel de acuerdo débil, lo que sugiere que los modelos no coinciden frecuentemente en la emoción asignada a una misma frase. Sin embargo, la comparación entre Gemini y DeepSeek muestra una concordancia buena ($k = 0.6781$), lo que podría deberse a una mayor similitud entre sus arquitecturas o conjuntos de entrenamiento.

Este patrón refuerza la idea de que la detección de polaridad es una tarea más robusta y consistente entre modelos que la clasificación detallada de

emociones [39]. Los valores de Kappa muestran una mejora considerable respecto al modelo original, pero persisten niveles débiles de acuerdo en emociones específicas. Esto indica que, si bien se ha logrado un avance significativo en términos de estabilidad y coherencia del modelo, aún existen limitaciones importantes en la capacidad de identificar emociones con precisión a nivel de frase.

Aunque se observa una mayor coincidencia entre Gemini y DeepSeek en la asignación de emociones, esta concordancia no garantiza que ambos modelos clasifiquen correctamente, sino que podrían estar cometiendo errores similares [40]. Por esta razón, resulta fundamental contar con un conjunto de referencia basado en etiquetado humano confiable, que sirva como punto de comparación para evaluar objetivamente el desempeño de los sistemas automáticos [41].

Para futuras mejoras metodológicas, sería conveniente ampliar la base de entrenamiento en español, con datos etiquetados manualmente que reflejen la diversidad lingüística y cultural del habla real. Asimismo, podría explorarse la incorporación de enfoques híbridos, como la combinación de emociones o el modelado de matices contextuales, con el fin de abordar con mayor eficacia la ambigüedad semántica y la variabilidad expresiva propias de las interacciones habladas.

Análisis de conductas predominantes (por hablante y a nivel global)

Además del análisis frase por frase, se evaluó la polaridad y la emoción predominante por hablante y por interacción completa. Aquí también se calculó el coeficiente de Kappa entre modelos:

	Polaridad predominante (nivel global)	Emoción predominante (nivel global)	Polaridad predominante (por hablante)	Emoción predominante (por hablante)
3 modelos (conjuntos)	0.5610 (Moderada)	0.7612 (Buena)	0.6756 (Buena)	0.4368 (Moderada)
Nuestra herramienta	0.3600 (Débil)	0.6322 (Buena)	0.6616 (Buena)	0.3512 (Débil)

vs. Gemini				
Nuestra herramienta vs. DeepSeek	0.5733 (Moderada)	0.6322 (Buena)	0.5938 (Moderada)	0.3322 (Débil)
Gemini vs. DeepSeek	0.7538 (Buena)	1 (Muy Buena)	0.7676 (Buena)	0.6161 (Buena)

Tabla 11. Resultados de Kappa de Fleiss para el análisis por participante y global.

El análisis de conductas predominantes, que considera la polaridad y la emoción dominante tanto a nivel de hablante como de la interacción completa, arroja resultados más consistentes que el análisis frase por frase. Esta segunda aproximación permite evaluar no solo eventos puntuales, sino tendencias emocionales y actitudinales a lo largo de toda la conversación, lo que facilita la identificación de patrones estables.

Polaridad predominante: En cuanto a la polaridad predominante, los resultados muestran un nivel de acuerdo moderado a bueno entre los tres modelos cuando se evalúa el total de hablantes ($k = 0.6756$) y la conversación completa ($k = 0.5610$). Esto sugiere que los sistemas logran identificar de forma relativamente estable si la actitud general de un hablante o de toda la interacción es “positiva”, “negativa” o “neutral”. Las comparaciones por pares revelan una mejor concordancia entre Gemini y DeepSeek ($k = 0.7538$ y 0.7676), mientras que el acuerdo entre nuestro modelo y los comerciales es más bajo ($k \approx 0.36$ y 0.57).

Emoción predominante: En la clasificación de emociones predominantes, se alcanzaron los mejores niveles de acuerdo de todo el análisis. El valor de kappa para la emoción global entre los tres modelos fue alto ($k = 0.7612$), y la concordancia entre Gemini y DeepSeek fue perfecta en ese aspecto ($k = 1$). Estos resultados indican que, cuando se consideran las emociones dominantes en contextos más amplios, los modelos tienden a coincidir con mayor frecuencia. A nivel de hablante, el acuerdo también es aceptable ($k = 0.4368$ entre los tres modelos), aunque menor que a nivel

global, probablemente debido a las variaciones individuales y estilos comunicativos.

Comparaciones: Las comparaciones por pares muestran nuevamente que nuestro modelo presenta mayor discrepancia respecto a los comerciales, especialmente en la clasificación de emociones predominantes por hablante ($k \approx 0.33\text{--}0.35$), lo que indica que sigue habiendo desafíos en el reconocimiento de matices emocionales en el habla individual.

Estos resultados evidencian que el análisis a nivel macro (por hablante o conversación completa) ofrece mayor robustez y coherencia entre modelos que el análisis frase por frase. Además, reflejan una mejoría notable del sistema propio tras el rediseño, en especial en la capacidad de captar la polaridad general y las emociones dominantes en interacciones prolongadas, acercándose al desempeño de herramientas comerciales más avanzadas.

6.3 Conclusión general y posibles mejoras

Los resultados obtenidos evidencian una evolución significativa en el desempeño de la herramienta tras su rediseño, especialmente en la detección de polaridad predominante y emociones globales por hablante e interacción completa. En comparación con herramientas comerciales, nuestro sistema alcanzó un acuerdo un acuerdo razonable en polaridad, pero presentó limitaciones en la identificación de emociones específicas a nivel frase, donde el acuerdo fue bajo.

En relación con estudios previos en español centrados en análisis de polaridad [42, 43, 44], nuestros resultados reflejan una mejora en la precisión del análisis general de emociones. Estas investigaciones, que utilizaron métodos tradicionales y aprendizaje automático, señalaron dificultades para capturar matices culturales y contextuales propios del español, limitaciones que nuestro sistema busca mitigar mediante la incorporación de modelos preentrenados más robustos, como PySentimiento.

Comparado con proyectos más avanzados en inglés, como GoEmotions, que clasifica 27 emociones utilizando modelos como RoBERTA, nuestro sistema, enfocado en español y en las seis emociones básicas de Ekman, obtiene resultados comparables en el análisis global por participante, aunque con

menor granularidad y precisión a nivel frase. PySentimiento, por su parte, reporta una F1 de 0.72 para emociones en español, desempeño similar al observado en nuestras pruebas, pero sobre datasets más estandarizados.

Estas comparaciones muestran que, si bien el sistema desarrollado ha logrado avances importantes y un acercamiento a herramientas de referencia, persisten desafíos propios del análisis emocional en español, particularmente en la detección de emociones específicas en unidades breves y en la consideración de variabilidad cultural y lingüística.

En el análisis frase por frase, la identificación de polaridad alcanzó niveles de acuerdo que van de moderados a buenos, mientras que la detección de emociones específicas continúa siendo una tarea compleja, con bajos niveles de concordancia entre modelos. Esto evidencia la dificultad inherente del análisis emocional en unidades lingüísticas breves, donde los matices contextuales son difíciles de captar, incluso para modelos avanzados.

Sin embargo, al pasar a un análisis de conductas predominantes por hablante y por interacción completa, los niveles de concordancia mejoran significativamente. La clasificación de emociones globales y por participante muestra una mayor estabilidad entre herramientas, sugiriendo que el procesamiento de tendencias emocionales en contextos más amplios es una estrategia más efectiva para lograr resultados confiables.

En conclusión, el rediseño del sistema cumplió su objetivo principal: mejorar la sensibilidad para detectar polaridad y emociones dominantes, acercándose a la precisión de modelos comerciales. No obstante, el sistema aún enfrenta limitaciones vinculadas a los retos que plantea el lenguaje emocional, especialmente en frases cortas, ambigüedad semántica y variabilidad interindividual en la expresión afectiva.

A partir de estos resultados obtenidos y las limitaciones identificadas en el presente trabajo, se plantean diversas líneas de investigación orientadas a fortalecer y ampliar las capacidades del sistema desarrollado. Estas futuras direcciones buscan mejorar la precisión y robustez del análisis emocional, específicamente en el idioma español. A continuación, se describen las principales áreas de desarrollo que pueden contribuir significativamente a estos objetivos.

- **Construcción y anotación de dataset en español:** Ampliar la base de datos mediante etiquetado humano de emociones es fundamental para superar la baja concordancia en la detección de emociones específicas. Esto permitirá entrenar modelos más sensibles a matices afectivos y evaluar rigurosamente el desempeño en español, superando las limitaciones señaladas en estudios previos [42, 43, 44].
- **Integración de análisis multimodal (texto y audio):** La incorporación de señales prosódicas a partir de modelos como wav2vec2-base-finetuned-sentiment-classification-MESD [45], con alta precisión en español, podría mejorar el reconocimiento emocional, sobre todo en casos ambiguos o neutros, complementando la información textual.
- **Optimización del procesamiento multilingüe:** Ajustar parámetros para la detección de emociones en inglés, como el umbral mínimo de tokens, podría perfeccionar la identificación emocional en los textos traducidos.
- **Incorporación de emociones compuestas o dimensionales:** Complementar el enfoque categórico actual (centrado en las seis emociones básicas de Ekman) con modelos dimensionales, como el modelo PAD (Valencia, Activación, Dominancia), permitiría capturar matices emocionales más sutiles y resolver ambigüedades frecuentes en emociones como la sorpresa, el miedo o el disgusto.
- **Análisis de error sistemático:** Establecer un procedimiento de análisis de errores frecuentes, por tipo de emoción, longitud de la frase, hablante o estructura lingüística, permitiría identificar patrones de fallo específicos del modelo. Esta información puede orientar ajustes precisos en el preprocesamiento, el modelo o el umbral de decisión.

7. Referencias

- [1] Murthy, G. S. N., Allu, S. R., Andhavarapu, B., Bagadi, M., & Belusonti, M.: Text based sentiment analysis using LSTM. Int. J. Eng. Res. Technol. 9(5), 290–295 (2020).
<https://www.ijert.org/text-based-sentiment-analysis-using-lstm>

- [2] Pérez, J.M., Rajngewerc, M., Giudici, J.C., Furman, D.A., Luque, F., Alonso Alemany, L., Martínez, M.V.: pysentimiento: A Python toolkit for opinion mining and social NLP tasks. arXiv preprint (2021). <https://arxiv.org/abs/2106.09462>
- [3] PySentimiento. Repositorio de PySentimiento. <https://github.com/pysentimiento/pysentimiento>
- [4] SamLowe. RoBERTa-base-GoEmotions. HuggingFace model hub. https://huggingface.co/SamLowe/roberta-base-go_emotions
- [5] Mullen, T., & Collier, N.: Sentiment analysis using support vector machines with diverse information sources. In: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 412–418 (2004). <https://aclanthology.org/W04-3253/>
- [6] Poria, S., Cambria, E., Howard, N., Huang, G.-B., & Hussain, A.: Fusing audio, visual and textual clues for sentiment analysis from multimodal content. *Neurocomputing* 174(A), 50–59 (2016). <https://www.sciencedirect.com/science/article/abs/pii/S0925231215011297>
- [7] Gómez-Zaragozá, L., del Amor, R., Castro-Bleda, M. J., & Vale, V. (2024). EMOVOME: A dataset for emotion recognition in spontaneous real-life speech. arXiv preprint. <https://arxiv.org/pdf/2403.02167>
- [8] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I.: Attention is all you need. arXiv preprint (2023). <https://arxiv.org/abs/1706.03762>
- [9] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint (2019). <https://arxiv.org/abs/1810.04805>
- [10] Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., & Ravi, S.: GoEmotions: A Dataset of Fine-Grained Emotions. arXiv preprint (2020). <https://arxiv.org/pdf/2005.00547>
- [11] Cañete, J., Chaperon, G., Fuentes, R., Ho, J.-H., Kang, H., & Pérez, J.: Spanish Pre-trained BERT Model and Evaluation Data. arXiv preprint (2023). <https://arxiv.org/abs/2308.02976>
- [12] Google Cloud: Cloud Natural Language <https://cloud.google.com/natural-language/docs?hl=es-419>

- [13] Amazon Web Services: Amazon Comprehend.
<https://aws.amazon.com/es/comprehend/>
- [14] Microsoft: Azure Language Service documentation.
<https://learn.microsoft.com/en-us/azure/ai-services/language-service/>
- [15] OpenAI: Whisper: Open-source automatic speech recognition (2022). <https://github.com/openai/whisper>
- [16] Bain, M., Huh, J., Han, T., & Zisserman, A.: WhisperX: Time-Accurate Speech Transcription of Long-Form Audio. arXiv preprint (2023). <https://arxiv.org/abs/2303.00747>
- [17] WhisperX [Repositorio GitHub]
<https://github.com/m-bain/whisperX>
- [18] Amazon Web Services. Amazon Transcribe Medical. Amazon.
<https://aws.amazon.com/es/transcribe/medical/>
- [19] TextBlob. TextBlob: Simplified Text Processing.
<https://textblob.readthedocs.io/en/dev/>
- [20] VADER Sentiment Analysis. Valence Aware Dictionary and sEntiment Reasoner. <https://github.com/cjhutto/vaderSentiment>
- [21] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, É., & Lample, G. LLaMA: Open and efficient foundation language models. (2023). arXiv. <https://arxiv.org/pdf/2302.13971>
- [22] Modelo RoBERTa-base-GoEmotions en HuggingFace:
https://huggingface.co/SamLowe/roberta-base-go_emotions
- [23] Google Cloud: Cloud Natural Language
<https://cloud.google.com/natural-language/docs?hl=es-419>
- [24] Microsoft: Azure Language Service documentation.
<https://learn.microsoft.com/en-us/azure/ai-services/language-service/>
- [25] Amazon Web Services: Amazon Comprehend.
<https://aws.amazon.com/es/comprehend/>
- [26] Google: Documentación de la API de Gemini.
<https://ai.google.dev/gemini-api/docs?hl=es-419>
- [27] DeepSeek: Documentación de la API de DeepSeek.
<https://api-docs.deepseek.com/>
- [28] Zenodo (2024). Dataset de emociones en audios:
<https://zenodo.org/records/10694370>

- [29] Zenodo (2018). Extractos de películas: <https://zenodo.org/records/1326428>
- [30] CIEMPIESS (2017). Dataset de conversaciones: <https://ciempiess.org/downloads>
- [31] OpenSLR (2019). Audios en español argentino: <https://openslr.org/61/>
- [32] OpenSLR (2020). TED talks en español: <https://openslr.org/67/>
- [33] Google: Colaboratory FAQ (2024). <https://research.google.com/colaboratory/faq.html>
- [34] Bisong, E. (2019). Google Colaboratory. En Building machine learning and deep learning models on Google Cloud Platform. Apress.
- [35] Torres Gordillo, J. J., & Perera Rodríguez, V. H.: Cálculo de la fiabilidad y concordancia entre codificadores de un sistema de categorías para el estudio del foro online en e-learning. Revista de Investigación Educativa 27(1), 89–103 (2010). <https://revistas.um.es/rie/article/view/94291>
- [36] NLTK. Natural Language Toolkit [Repositorio de software]. GitHub. <https://github.com/nltk/nltk>
- [37] Python Software Foundation. difflib — Helpers for computing deltas. In Python 3 Standard Library documentation. <https://docs.python.org/3/library/difflib.html>
- [38] Groq. Groq API. <https://groq.com/>
- [39] Mohammad, S. M. (2020). Sentiment analysis: Detecting valence, emotions, and other affectual states from text. In J. A. Calvo, S. D’Mello, J. Gratch, & A. Kappas (Eds.), The Oxford Handbook of Affective Computing (2nd ed.). Elsevier. <https://www.saifmohammad.com/WebDocs/emotion-survey.pdf>
- [40] Fort, K., Adda, G., & Cohen, K. B. (2011). Amazon Mechanical Turk: Gold mine or coal mine? Computational Linguistics, 37(2), 413–420. <https://aclanthology.org/J11-2010.pdf>
- [41] Buechel, S., & Hahn, U. (2017). EmoBank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, Volume 2: Short Papers, 578–585. <https://aclanthology.org/E17-2092.pdf>

- [42] Dubiau, L., & Ale, J. M.: Análisis de Sentimientos sobre un Corpus en Español: Experimentación con un Caso de Estudio. (2013). <http://sedici.unlp.edu.ar/handle/10915/76148>
- [43] Pauli, P. A.: Análisis de sentimiento: Comparación de algoritmos predictivos y métodos utilizando un lexicon español. (2019). Trabajo Final, Instituto Tecnológico de Buenos Aires (ITBA), Escuela de Ingeniería Informática, Buenos Aires, Argentina. <https://ri.itba.edu.ar/server/api/core/bitstreams/db2a9097-b8f4-4205-8048-8f9fdc76cd66/content>
- [44] Cedeño-Moreno, D., & Vargas, M.: Aprendizaje automático aplicado al análisis de sentimientos. *I+D Tecnológico* 16(2), 59–66 (2020). <https://revistas.utp.ac.pa/index.php/id-tecnologico/article/view/2833>
- [45] *somosnlp*. (2022). *wav2vec2-base-finetuned-sentiment-mesd* [Modelo de aprendizaje automático]. Hugging Face. <https://huggingface.co/somosnlp-hackathon-2022/wav2vec2-base-finetuned-sentiment-mesd>
- [46] Ekman, P. (1992). An argument for basic emotions. *Cognition & Emotion*, 6(3–4), 169–20
- [47] *pydub* [Repositorio de software]. GitHub. <https://github.com/jjaaro/pydub>
- [48] *deep-translator* (versión 1.11.4) [Repositorio de software]. GitHub. <https://github.com/nidhaloff/deep-translator>
- [49] Church, K. W.: Word2Vec. *Natural Language Engineering* 23(1), 155–162 (2016). <https://www.cambridge.org/core/services/aop-cambridge-core/content/view/B84AE4446BD47F48847B4904F0B36E0B/S1351324916000334a.pdf/word2vec.pdf>
- [50] Mikolov, T., Chen, K., Corrado, G., & Dean, J.: Efficient Estimation of Word Representations in Vector Space. *arXiv preprint* (2013). <https://arxiv.org/abs/1301.3781>
- [51] Pennington, J., Socher, R., & Manning, C. D.: GloVe: Global Vectors for Word Representation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543 (2014). <https://nlp.stanford.edu/pubs/glove.pdf>
- [52] Yin, W., Kann, K., Yu, M., & Schütze, H.: Comparative Study of CNN and RNN for Natural Language Processing. *arXiv preprint* (2017). <https://arxiv.org/pdf/1702.01923>

- [53] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V.: RoBERTa: A Robustly Optimized BER
- [54] Pérez, J. M., Furman, D. A., Alonso Alemany, L., & Luque, F.: RoBERTuito: a pre-trained language model for social media text in Spanish. arXiv preprint (2021). <https://arxiv.org/pdf/2111.09453>
- [55] TASS: Taller de Análisis de Sentimientos en Español. <http://tass.sepln.org/>
- [56] Plaza-del-Arco, F. M., Strapparava, C., Ureña-López, L. A., Martín-Valdivia, M. T.: EmoEvent: A Multilingual Emotion Corpus based on different Events. In: Proceedings of the Twelfth Language Resources and Evaluation Conference, pp. 1492–1498 (2020). <https://aclanthology.org/2020.lrec-1.186.pdf>
- [57] EmotionPrompt: EmotionPrompt: Elevating AI with Emotional Intelligence. Medium (2024). <https://medium.com/aimonks/emotionprompt-elevating-ai-with-emotional-intelligence-baee341f521b>
- [58] Mahmood, M. A., Maab, I., Sibtain, M., Sarwar, A., Arsalan, M., & Hussain, M.: Advancements in Sentiment Analysis: A Methodological Examination of News using multiple LLMs. (2025). https://www.anlp.jp/proceedings/annual_meeting/2025/pdf_dir/P9-3.pdf
- [59] Cheng, Z., Cheng, Z.-Q., He, J.-Y., Sun, J., Wang, K., Lin, Y., Lian, Z., Peng, X., & Hauptmann, A.: Emotion-LLaMA: Multimodal Emotion Recognition and Reasoning with Instruction Tuning. arXiv preprint (2024). <https://arxiv.org/pdf/2406.11161>
- [60] Analizador de sentimientos y emociones. Repositorio de Github. <https://github.com/memimoyano/Analizador-de-Sentimientos-y-Emociones>