Decires, Vol. 26, No. 35, diciembre de 2025, pp. 7-28
DOI: 10.22201/cepe.14059134e.2025.26.35.479

Recibido: 29 de julio de 2025
Aceptado: 22 de agosto de 2025

Un corpus de oraciones para el análisis de emociones en estudiantes de inglés mediante algoritmos de inteligencia artificial

A Corpus of Sentences for Emotion Analysis in Spanish-Speaker Students of English Using Artificial Intelligence Algorithms

Roberto Ángel Meléndez Armenta
ramelendeza@itsm.edu.mx
ORCID: 0000-0001-8994-0944
Instituto Tecnológico Superior de Misantla, Veracruz

Kevin Bernardo Herrera Carmona
192t0441@itsm.edu.mx
Tecnológico Nacional de México - Instituto Tecnológico Superior de Misantla

Francisco Fernandez-Dominguez
fjfernandezd@itsm.edu.mx
Tecnológico Nacional de México - Instituto Tecnológico Superior de Misantla
ORCID: 0009–0007–9602–1963

Edgar Degante-Aguilar
edgar.da@teziutlan.tecnm.mx
Tecnológico Nacional de México - Instituto Tecnológico Superior de Teziutlán
ORCID: 0009-0001-2382-944X

Resumen: El reconocimiento de las emociones desempeña un papel crucial dentro de las aulas tradicionales, ya que facilita una mejor comprensión de los comportamientos del estudiantado y permite implementar estrategias pedagógicas más efectivas. Aunque en los últimos años el reconocimiento de emociones mediante inteligencia artificial (IA) ha mostrado avances significativos (Kalateh et al., 2024; Hashem et al., 2023), en México este campo aún se consi-

dera emergente, principalmente por la escasa atención a las particularidades lingüísticas y culturales de los hispanohablantes. Mares et al. (2025) evidencian esta limitación al identificar únicamente siete conjuntos de datos en español para el análisis del campo afectivo, lo que restringe tanto el desarrollo como la precisión de modelos robustos en este ámbito. Un problema que destaca es la ausencia de conjuntos de datos que reflejen cómo la comunidad estudiantil mexicana expresa emociones en inglés, lo que revela una brecha significativa en la representación cultural y lingüística en las investigaciones de IA. Ante esta situación, se desarrolló un software destinado a la creación de un corpus de oraciones con distintas emociones. El proceso de construcción del corpus se dividió en dos fases principales: 1) la selección y definición del conjunto de oraciones en inglés y 2) la grabación de oraciones pronunciadas en inglés con seis emociones diferentes. En este experimento participaron 53 estudiantes universitarios del sureste de México, quienes aportaron grabaciones que sirvieron como base para entrenar modelos de IA. Los resultados del estudio demuestran que es posible desarrollar modelos capaces de clasificar emociones en inglés a partir de archivos de voz generados por hablantes de español. Además, el análisis de las oraciones recopiladas permitió identificar características acústicas clave que diferencian las seis categorías emocionales, mostrando la importancia de esta metodología para mejorar la comprensión de emociones en contextos multilingües.

Palabras clave: corpus lingüístico; emociones en la didáctica; IA ; enseñanza de segundas lenguas.

Abstract: Emotion recognition plays a crucial role in traditional classroom settings, as it enables a better understanding of student behavior and supports the implementation of more effective pedagogical strategies. Although recent years have seen significant advances in emotion recognition through artificial intelligence (Kalateh et al., 2024; Hashem et al., 2023), this field is still considered emerging in Mexico, mainly due to the limited attention given to the linguistic and cultural specificities of Spanish-speaking populations. Mares et al. (2025) highlight this limitation by identifying only seven Spanish-language datasets for affective analysis, which constrains both the development and accuracy of robust models in this domain. A particularly pressing issue is the lack of datasets that capture how Mexican students express emotions in English, revealing a substantial gap in cultural and linguistic representation within artificial intelligence research. In response to this situation, a software tool was developed to construct a corpus of emotion-labeled sentences. The corpus creation process was divided into two

Recibido: 29 de julio de 2025
Aceptado: 22 de agosto de 2025

main phases: (1) the selection and definition of English-language sentences, and (2) the recording of those sentences spoken with six distinct emotional expressions. A total of 53 university students from southeastern Mexico participated in the experiment, contributing voice recordings that were used as the foundation for training artificial intelligence models. The study's results demonstrate the feasibility of developing models capable of classifying emotions in English from voice recordings produced by Spanish-speaking individuals. Furthermore, the analysis of the collected sentences revealed key acoustic features that distinguish the six emotional categories, underscoring the value of this methodology for enhancing emotion recognition in multilingual contexts.

Keywords: linguistic corpus; emotions in didactics; artificial intelligence; second language teaching.

INTRODUCCIÓN

Actualmente el reconocimiento de emociones mediante algoritmos de inteligencia artificial (IA) es un campo emergente con un amplio potencial para mejorar la interacción humano-computadora. Esta tecnología ha ganado notoriedad debido a su creciente aplicación en diversos contextos, como el análisis de sentimientos en redes sociales, la evaluación del nivel de satisfacción en compras en línea y la detección de estados emocionales en estudiantes ante situaciones académicas específicas.

En el ámbito del reconocimiento de emociones, contar con múltiples conjuntos de datos es fundamental para el desarrollo y la validación de modelos robustos. El desempeño de un algoritmo de IA depende en gran medida de la cantidad y calidad de los conjuntos de datos utilizados en las fases de entrenamiento y validación (Zha et al., 2025). La diversidad de datos permite comparar enfoques, evaluar el rendimiento de distintas técnicas y garantizar una mayor generalización de los sistemas propuestos (Kalateh et al., 2024). Sin embargo, como señalan Mares et al. (2025), existe una notable escasez de conjuntos de datos en español, especialmente en el ámbito del análisis emocional, ya que en su estudio identifican únicamente siete corpus disponibles en este idioma. Esta limitación se vuelve aún más significativa en contextos locales, como lo es en estudiantes de México.

Con el fin de contribuir al cierre de esta brecha, la presente investigación propone la creación de un corpus de oraciones en audios que capture de manera auténtica las expresiones emocionales del alumnado mexicano al hablar inglés.

Este corpus no solo enriquecerá los recursos disponibles para el procesamiento de lenguaje natural (PLN) en español, sino que también facilitará el estudio de patrones culturales y lingüísticos únicos en otros idiomas.

De manera general, un corpus de oraciones consiste en una colección estructurada de oraciones en un idioma particular, obtenida tanto de fuentes primarias escritas, tales como libros, publicaciones científicas, redes sociales y sitios web (Laserna y Torres, 2022) como orales (transcripciones) (Sultana et al., 2021). El corpus desarrollado en este trabajo está compuesto por un conjunto de audios de voz etiquetados de acuerdo con la emoción que expresan, los cuales fueron analizados mediante algoritmos de inteligencia artificial. Este proceso de análisis busca responder a las siguientes preguntas en torno al corpus elaborado:

- ¿Cómo evaluar la calidad del etiquetado emocional del corpus de audios de voz mediante algoritmos de IA?

- ¿Cómo puede evaluarse la expresión vocal de emociones en los idiomas español e inglés de estudiantes hispanohablantes realizada mediante la implementación de algoritmos de IA (SVM y Random Forest) aplicados a espectrogramas de frecuencia extraídos de un corpus de oraciones en audio?

La voz es una señal sonora que se produce por el paso del aire a través de las cuerdas vocales y se crean variaciones de sonido en la pronunciación de diferentes fonemas. Esta señal contiene una amplia gama de información, según Egger et al. (2019) y Nassif et al. (2019), y entre las principales características que pueden extraerse de la señal de voz se encuentran:

- Contenido del habla. Información semántica o lingüística (palabras y frases pronunciadas).

- Identidad del hablante. Rasgos únicos que permiten reconocer a la persona que habla.

- Idioma hablado. Identificación del idioma utilizado por el hablante.

- Acento. Información sobre la variante regional o nacional del idioma hablado.

- Entonación y ritmo. Los elementos prosódicos que reflejan el flujo natural del habla.

Recibido: 29 de julio de 2025
Aceptado: 22 de agosto de 2025

- Emociones. El estado emocional del hablante (felicidad, enojo, tristeza, etc.) que influye en tono, energía y ritmo.

- Estado de salud. Indicadores de afecciones físicas o mentales que afectan la claridad, fluidez o tono.

- Edad. Cambios naturales en el tono, estabilidad y fuerza vocal a lo largo del tiempo.

- Género. Información relacionada con características vocales típicas de hombres o mujeres.

- Condición socioeconómica. Posibles indicios en la pronunciación, énfasis o expresión vinculados a la educación o entorno social.

- Aplicaciones psicológicas y clínicas. Posibilidad de detectar emociones en personas con dificultades comunicativas (autismo, síndrome de enclaustramiento).

- Adaptabilidad e interacción humano-máquina. Capacidad de los sistemas para ajustar sus respuestas con base en las emociones detectadas.

- Relación con postura corporal. La voz puede variar dependiendo de la postura del hablante.

En particular, los sistemas de reconocimiento de emociones por voz (SER, por sus siglas en inglés) se basan en la extracción de características como cruce por cero, *pitch* (intensidad), energía, tono, espectrogramas, entre otras. Los avances en la capacidad de procesamiento computacional permiten aplicar técnicas de aprendizaje robustas basadas en datos 2D, como las imágenes de espectrogramas obtenidas de los corpus de audios etiquetados por emociones. Actualmente se ha mostrado que las redes neuronales recurrentes (RNN, por sus siglas en inglés), los algoritmos máquinas de vectores de soporte (SVM, por sus siglas en inglés) y los bosques aleatorios (RF, por sus siglas en inglés) son algunas de las herramientas ampliamente utilizadas en el tratamiento de señales de voz y forman parte fundamental de los transformadores (*transformers*), los cuales son una arquitectura de redes convolucionales y recurrentes que muestran un alto desempeño en el procesamiento del lenguaje natural (NLP, por sus siglas en inglés) (Vaswani et al., 2017).

Para este trabajo se seleccionaron los algoritmos SVM y RF. Ambos han demostrado un desempeño competitivo en tareas de clasificación supervisada, especial-

mente en contextos con conjuntos de datos limitados y características acústicas extraídas de señales de voz. SVM es eficaz en espacios de alta dimensionalidad y ofrece buenos resultados al trazar fronteras óptimas entre clases emocionales. Por su parte, RF destaca por su robustez ante datos ruidosos y su capacidad de manejar variables no lineales. Por otro lado, se trabajó con cinco emociones: alegría, enojo, tristeza, desagrado y miedo, seleccionadas con base en el modelo dimensional de emociones de Russell (1980). A estas se añadió una expresión neutral, que funciona como punto de referencia central dentro del plano bidimensional de valencia y activación propuesto por dicho modelo. Tres de las emociones consideradas —alegría, enojo y tristeza— se distribuyen en el primer, segundo y tercer cuadrante, respectivamente, lo cual permite una cobertura representativa y diferenciable del espacio emocional. Con ello, se espera facilitar el proceso de construcción del conjunto de datos propuesto y, finalmente, mejorar los resultados del algoritmo de clasificación emocional. En cambio, las emociones de desagrado y miedo se ubican en el segundo cuadrante, junto con el enojo, lo que implica una mayor concentración de emociones negativas en dicha región del modelo y una menor dispersión global en el plano emocional.

SITUACIÓN ACTUAL

La revisión de la literatura demuestra que la creación de un corpus de oraciones para el reconocimiento de emociones en estudiantes es fundamental para el desarrollo de sistemas empleados en el ámbito académico; Ashraf et al. (2023) así lo demuestran en su estudio sobre el análisis de sentimientos en urdu, en el cual se revela la capacidad de los algoritmos para aprender patrones y realizar predicciones precisas sobre las emociones del alumnado.

La identificación de emociones ha ganado popularidad por su amplia aplicación práctica en diversos contextos, por ejemplo, análisis de sentimientos a partir de *datasets* conformados por comentarios en redes sociales, satisfacción del usuario por las compras de productos en línea, identificación de emociones en estudiantes ante situaciones concretas, entre otros ejemplos de gran alcance e importancia. En otras palabras, para alcanzar los objetivos deseados, se requiere construir conjuntos de información a partir de escritos, ya sean completos o fragmentos, con el propósito de formar un cuerpo de oraciones que actuará como base para su análisis posterior. Este conjunto de datos de oraciones se entiende como una base de información (*dataset*) compuesta por textos que han sido elegidos y marcados previamente. Su utilidad radica en el análisis de las emociones expresadas en ellos a través de procesos de reconocimiento y cate-

Decires, Vol. 26, No. 35, diciembre de 2025, pp. 7-28
DOI: 10.22201/cepe.14059134e.2025.26.35.479

Recibido: 29 de julio de 2025
Aceptado: 22 de agosto de 2025

gorización empleados por IA. Esta *dataset* se construye para un propósito específico y proporciona los datos necesarios para entrenar modelos que aprenden a identificar patrones y características asociados a diferentes emociones a partir de las oraciones etiquetadas.

El avance en el campo de la definición de corpus se centra en diversos propósitos. Destacan numerosas investigaciones con aplicación de corpus de oraciones basadas en texto para su uso con técnicas de NLP; esta rama es la más popular por la alta disponibilidad de textos en redes sociales, que abarcan desde el análisis de texto básico hasta la construcción de taxonomías y la predicción de resultados textuales. Por ejemplo, Kim et al. (2022) investigan el impacto de las palabras altisonantes en el análisis de sentimientos, mientras que Kalateh et al. (2024) proponen un método holístico para capturar emociones humanas a través de múltiples modalidades. Gonzalez-Gomez et al. (2024) analizan la evolución del NPL y su adaptación a diversas disciplinas. Por otro lado, Aljuhani et al. (2021) mejoran el reconocimiento de emociones en árabe y Sen et al. (2022) presentan un análisis exhaustivo del NPL aplicado al bengalí. Además, Ashraf et al. (2023) desarrollan un enfoque para el análisis de sentimientos en urdu, mientras que Nasution y Onan (2024) comparan la calidad de las anotaciones generadas por humanos y por modelos de lenguaje.

Althari y Alsulmi (2022) analizaron las posibilidades del aprendizaje basado en transformadores para la identificación de negaciones en textos biomédicos. Shim et al. (2019) desarrollaron un planteamiento para disminuir la cantidad de etiquetado manual requerido en el análisis de sentimientos. Asimismo, Mifrah y Benlahmar (2022) efectuaron una comparación de modelos de aprendizaje profundo en el contexto de la clasificación de sentimientos.

Kalateh et al. (2024) señalan que la calidad y fiabilidad de estas anotaciones son la clave del rendimiento y la utilidad de las aplicaciones de NPL e indican que un corpus bien construido, con oraciones etiquetadas con precisión, permite a los algoritmos aprender patrones y realizar predicciones precisas sobre el estado emocional expresado en la *dataset*. Un corpus con esas características genera asimismo un punto de partida para nuevas investigaciones en otros campos de aplicación, como el de Althari y Alsulmi (2022).

Entre las diversas implementaciones en proyectos de investigación y artículos científicos es posible observar una amplia gama de conjuntos de datos utilizados para clasificar las emociones mediante el empleo de IA. Cada corpus presenta características particulares que determinan la pertinencia de su aplicación. CREMA-D (Crowd Sourced Emotional Multimodal Actors Dataset) es un corpus de datos audiovisuales con la capacidad de representar las emociones básicas de neutro, felicidad, ira, disgusto, miedo y tristeza; los datos están crea-

dos en inglés, es multimodal y etiquetado mediante *crowdsourcing*; sus características y bondades son apropiadas para este proyecto de investigación, dado que su principal fortaleza radica en que reúne a participantes provenientes de distintas regiones geográficas y con un rango amplio de edades (Cao et al., 2014), por lo que se le considera como uno de los corpus más fiables en diversos experimentos científicos. Sin embargo, existen otros corpus con características similares. RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) es un corpus audiovisual cuya característica principal consiste en incorporar la calma como una emoción básica; además, incluye habla y canto en sus registros (Rzayeva y Alasgarov, 2019). SAVEE (Surrey Audio-Visual Expressed Emotion) es un corpus con datos de grabaciones audiovisuales con voces de actores y se encuentra en inglés (Zehra et al., 2021). TESS (Toronto Emotional Speech Set), por el contrario, es un corpus construido con grabaciones de voces femeninas (Parry et al., 2019); IEMOCAP (Interactive Emotional Dyadic Motion Captura) es multimodal, en inglés e incluye una captura de movimiento facial y transcripciones que ayudan a conseguir resultados prometedores en áreas alternas, como el análisis multimodal de emociones y el estudio de la comunicación expresiva humana (Antoniou et al., 2023). Los corpus KSUEmotions (Meftah et al., 2021), Saudi Dialect Corpus (Aljuhani et al., 2021) y Urdu Tweet (Amjad et al., 2021) no clasifican las emociones básicas; emplean únicamente *datasets* de texto, y las emociones se decantan por resultados positivos o negativos.

Método utilizado en la investigación

a) Demografía del grupo participante

Quienes conformaron el corpus fueron estudiantes de licenciatura pertenecientes a diversas carreras ofrecidas en el Instituto Tecnológico Superior de Misantla durante el semestre comprendido entre febrero y julio de 2024. La selección de participantes se realizó con base en un muestreo aleatorio simple, del cual se obtuvo una muestra de n=53 participantes, 33 hombres y 20 mujeres, en un rango de edad de 18 a 25 años. Es importante mencionar que el Instituto Tecnológico Superior de Misantla otorgó la aprobación ética establecida por el Comité de Ética de la institución para realizar el estudio dentro de sus instalaciones. Además, se obtuvo el consentimiento informado por escrito de quienes formaron parte del estudio. Cabe señalar que todas las personas participantes son nativas de la región y poseen como lengua materna el español; ninguna cuenta con un dominio

Decires, Vol. 26, No. 35, diciembre de 2025, pp. 7-28
DOI: 10.22201/cepe.14059134e.2025.26.35.479

Recibido: 29 de julio de 2025
Aceptado: 22 de agosto de 2025

bilingüe del idioma inglés. En algunos casos hubo participantes que presentaron dificultades de pronunciación, atribuibles a factores como el nerviosismo o la incomodidad durante el proceso de grabación. Para mitigar estas situaciones se realizaron ajustes como el cambio de horario o modificación del lugar de grabación, con el objetivo de generar un ambiente más cómodo para el estudiantado. No obstante, cuando estas medidas no resultaron efectivas, de común acuerdo entre quien participaba y la persona encargada del registro, se decidió no incluir sus grabaciones en el corpus final, por lo que no formaron parte del grupo final de 53 participantes.

b) Oraciones seleccionadas

El proyecto CREMA-D es una *dataset* en inglés que recopila las voces de 91 actores (48 hombres y 43 mujeres) de entre 20 y 74 años, y contiene las grabaciones de cada participante expresando 12 frases con distintas emociones (Cao et al., 2014).

A partir de las oraciones definidas en la *dataset* CREMA-D, se eligieron las oraciones y se tradujeron al español, con la finalidad de crear una versión de las oraciones en este idioma, las cuales fueron utilizadas en esta investigación (ver Tabla 1).

Tabla 1. Oraciones, en español e inglés, utilizadas por quienes participaron para generar el corpus

Oraciones en inglés	Oraciones en español
• It's eleven o'clock. • I wonder what this is about. • The airplane is almost full. • Don't forget a jacket. • The sun is very bright.	• Son las once en punto. • Me pregunto de qué trata esto. • El avión está casi lleno. • No olvides la chamarra. • El sol está muy brillante.

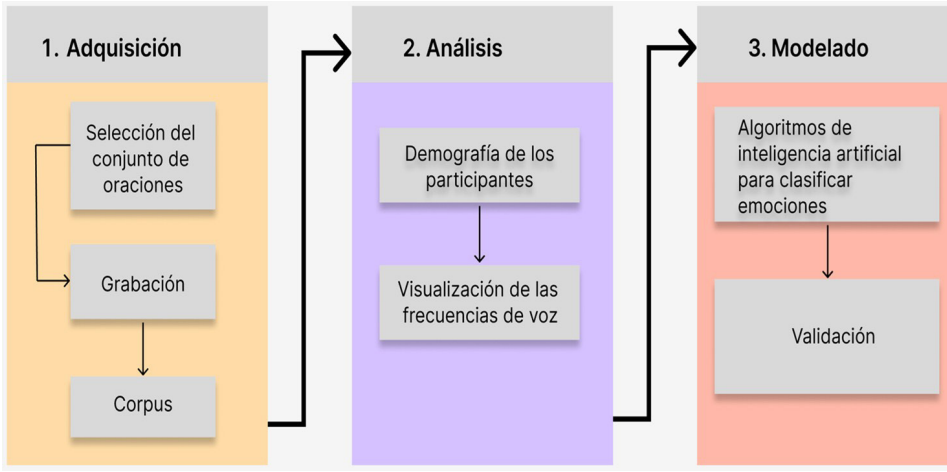
Fuente: elaboración propia.

Las oraciones fueron grabadas con seis emociones diferentes: enojo, tristeza, neutralidad, alegría, desagrado y miedo.

c) Procedimiento

La Figura 1 muestra los pasos de la metodología propuesta para la generación de un corpus destinado a realizar el análisis de emociones en estudiantes mediante algoritmos de IA.

Figura 1. Metodología propuesta para el desarrollo de esta investigación



Fuente: elaboración propia.

RESULTADOS

a) Corpus de oraciones

El proceso de grabar a 53 participantes que expresaron seis emociones diferentes para cada una de las cinco oraciones planteadas generó un corpus de archivos de audio que están asociados a estados emocionales en estudiantes de universidad. El corpus obtenido en esta investigación se dividió en dos conjuntos de datos que fueron etiquetados de la siguiente manera:

- Student Emotion Sentences in English Dataset (SES-ED), el cual consta de 1586 archivos de audio (4 grabaciones fueron descartadas por exceso de ruido).

- Student Emotion Sentences in Spanish Dataset (SES-SD), el cual consta de 1588 archivos de audio (2 grabaciones fueron descartadas por exceso de ruido).

Decires, Vol. 26, No. 35, diciembre de 2025, pp. 7-28
DOI: 10.22201/cepe.14059134e.2025.26.35.479

Recibido: 29 de julio de 2025
Aceptado: 22 de agosto de 2025

Para asegurar la calidad acústica de las grabaciones, la recolección de los audios se llevó a cabo en una sala de clases acondicionada para minimizar el ruido ambiental. En cada sesión participaban únicamente una persona voluntaria del cuerpo estudiantil y quien se encargaba de operar el software de grabación, que también guiaba el proceso y validaba la calidad del material obtenido.

Previo a la grabación, se proporcionó al grupo de participantes una explicación clara del objetivo general de la investigación y de la importancia de su participación. Como parte de la preparación, se solicitó que cada estudiante identificara y comprendiera las emociones a representar (alegría, tristeza, enojo, miedo, desagrado y neutralidad) y que pusiera en práctica su capacidad para expresarlas de forma vocal. Para facilitar este proceso, se entregó con al menos un día de anticipación una hoja con las seis oraciones a grabar en ambos idiomas (español e inglés), lo que les permitió practicar y minimizar posibles errores de pronunciación que pudieran afectar los resultados del análisis.

Durante la grabación, cada participante pronunció las seis oraciones, primero en español y posteriormente en inglés, representando una emoción distinta en cada una. Para gestionar este proceso, se diseñó un software especializado que indicaba en pantalla la oración a grabar, la emoción correspondiente y los botones de inicio y finalización. Al concluir cada grabación, tanto quien participaba como la persona facilitadora evaluaban en conjunto si la expresión emocional y la pronunciación eran adecuadas. En caso de detectar inconsistencias o errores, se procedía a repetir la grabación de la oración correspondiente.

Cabe destacar que uno de los objetivos principales de este trabajo es comprobar, mediante algoritmos de inteligencia artificial (IA), la veracidad y coherencia del corpus emocional generado y evaluar si las emociones expresadas por el grupo de participantes pueden ser reconocidas automáticamente a partir de las características acústicas de sus voces.

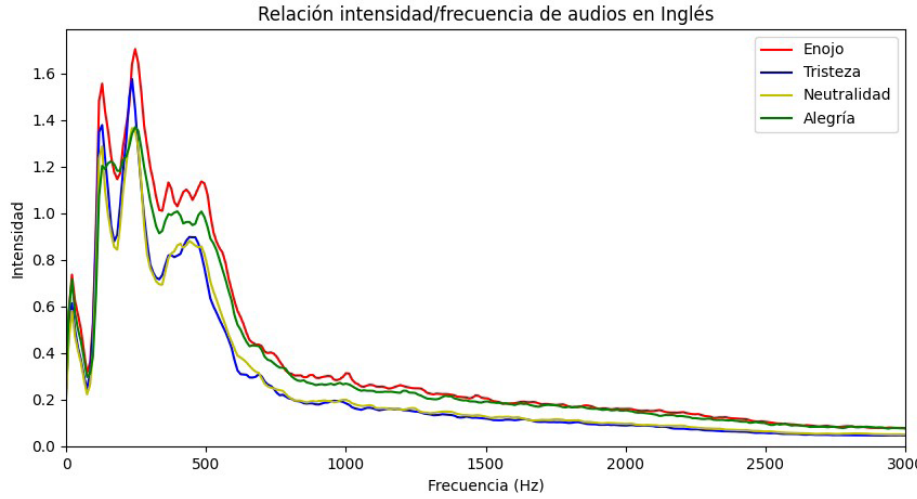
b) Frecuencias de voz

Hashem et al. (2023) representan en un espacio circular de dos dimensiones un conjunto amplio de emociones. Las que tienen una representación general en el plano dimensional son enojo, tristeza, neutralidad y alegría, razón por la cual el análisis de voz se realizó sobre estas cuatro emociones principales.

Las características espectrales, como los coeficientes espectrales de frecuencia Mel (MFCC por sus siglas en inglés), son unas de las principales utilizadas en los sistemas SER, como lo muestran Abdul y Al-Talabani (2022), quienes indican que el análisis acústico en general está basado en la característica MFCC para

aplicaciones con diversos clasificadores. En las Figuras 2 y 3 se realiza un análisis simple en la frecuencia de los audios y se grafica la relación intensidad/frecuencia de las cuatro emociones principales: alegría, tristeza, enojo y neutral en inglés y español, respectivamente. En este análisis se utilizó la Transformada de Fourier de Tiempo Reducido (STFT, por sus siglas en inglés) en ventanas de 93 milisegundos (ms).

Figura 2. Señales de voz en inglés de las cuatro emociones principales en función de la frecuencia



Fuente: elaboración propia.

La Figura 2 muestra el promedio de intensidad para cada componente en frecuencia de los audios en las distintas emociones. Las cuatro curvas son características de una señal de voz en la que se aprecian dos componentes principales de frecuencia (máximos); sin embargo, muestran un ligero desfase y distintos niveles de intensidad. Las curvas de enojo y alegría son de mayor intensidad que las otras (tristeza y neutralidad); la curva de alegría muestra más variaciones en frecuencias bajas respecto al resto. Por su parte, la curva de tristeza se diferencia de la de neutralidad porque cuenta con componentes de intensidad mayores en frecuencias bajas. La descripción anterior coincide con lo reportado por Shen et al. (2011), quienes analizaron siete emociones. En su estudio, el enojo se caracteriza por niveles elevados de energía y una alta concentración de componentes en frecuencias agudas; la felicidad también presenta una energía elevada, pero con una distribución espectral más equilibrada. En contraste, la tristeza se asocia con

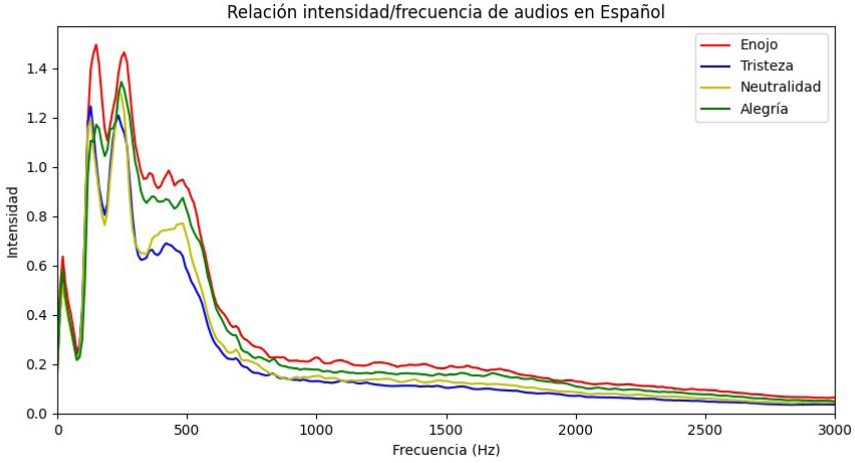
Decires, Vol. 26, No. 35, diciembre de 2025, pp. 7-28
DOI: 10.22201/cepe.14059134e.2025.26.35.479

Recibido: 29 de julio de 2025
Aceptado: 22 de agosto de 2025

bajos niveles de energía y una menor presencia de componentes en las frecuencias altas.

La Figura 3 pertenece a los audios en español; muestra un comportamiento similar a los de la Figura 2, pero presenta dos diferencias, principalmente: 1) la intensidad es menor para todas las curvas y 2) los componentes de frecuencia alta son menores en todos los casos.

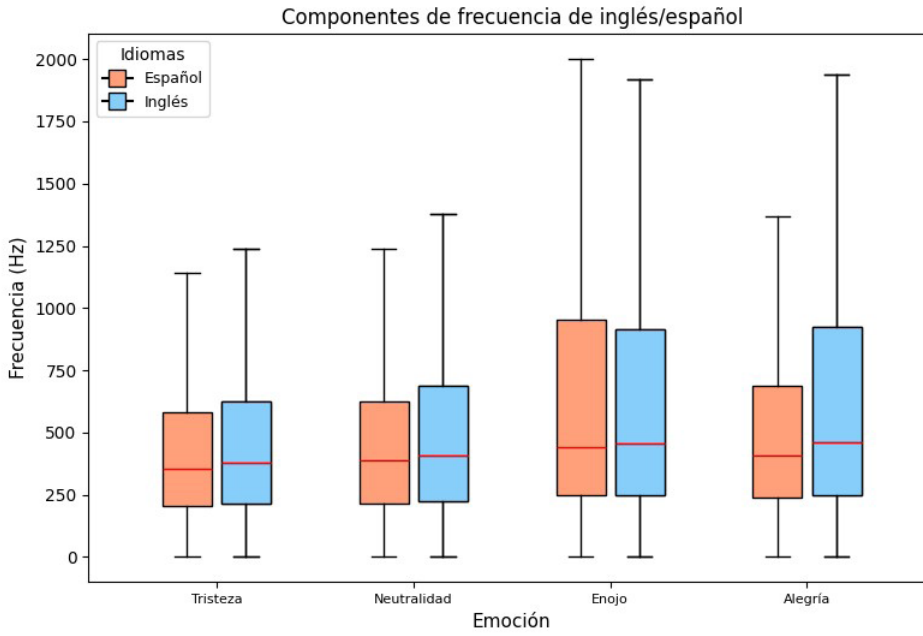
Figura 3. Señales de voz en español de las cuatro emociones principales en función de la frecuencia



Fuente: elaboración propia.

La comparación entre inglés y español de las distribuciones de frecuencias que se realizó en la Figura 4 muestra que el inglés tiene una mayor distribución de frecuencias en casi todas las emociones. También se observa que los componentes en frecuencia van de 0 a 2000 Hz, y las emociones de enojo y alegría tienen un mayor rango de distribución respecto a tristeza y neutralidad. Se aprecia una leve diferencia en la distribución, la cual es más notable en el caso de la emoción de alegría.

Figura 4. Distribución de las frecuencias de cada emoción en inglés-español



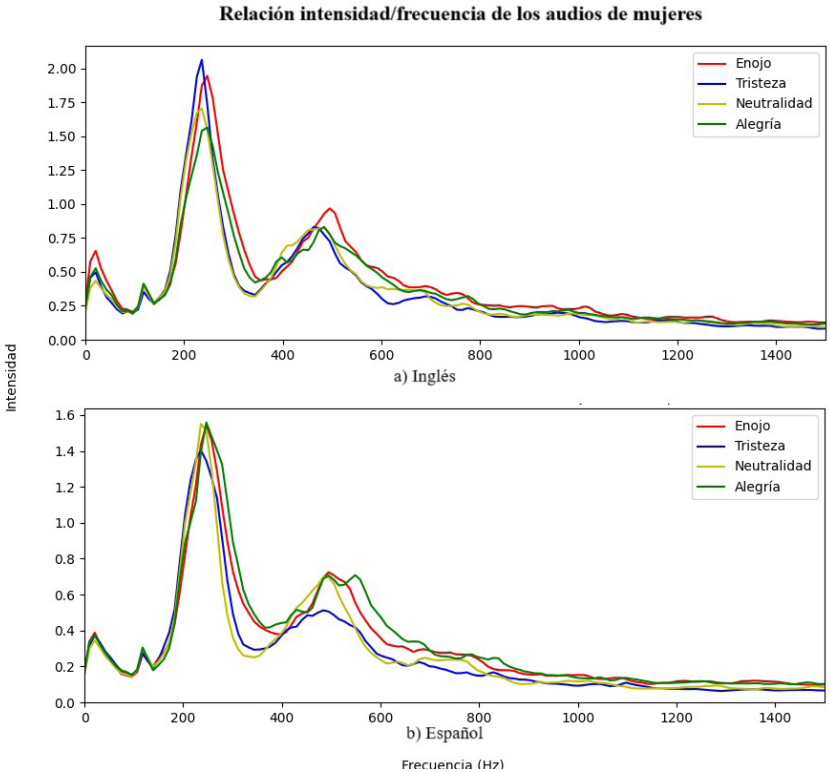
Fuente: elaboración propia.

Las características en frecuencia de la voz, especialmente la Frecuencia fundamental (F0), cambian de acuerdo con las características de la persona, como la edad o el género, mismas que concuerdan con el estudio realizado por Traunmüller y Eriksson (1995) para Europa y China, donde la F0 es de 120 y 210 Hz para hombres y mujeres, respectivamente. En las Figuras 5 y 6 se muestra una relación intensidad/frecuencia por género en español e inglés. En la Figura 5 se puede identificar la frecuencia F0 como el punto máximo de cada señal, aproximadamente superior a 200 Hz en ambos idiomas. En español la emoción de tristeza se logra distinguir del resto, mientras que las demás señales parecen ser similares. Sin embargo, para el inglés no coincide con los resultados de Shen et al. (2011).

Decires, Vol. 26, No. 35, diciembre de 2025, pp. 7-28
DOI: 10.22201/cepe.14059134e.2025.26.35.479

Recibido: 29 de julio de 2025
Aceptado: 22 de agosto de 2025

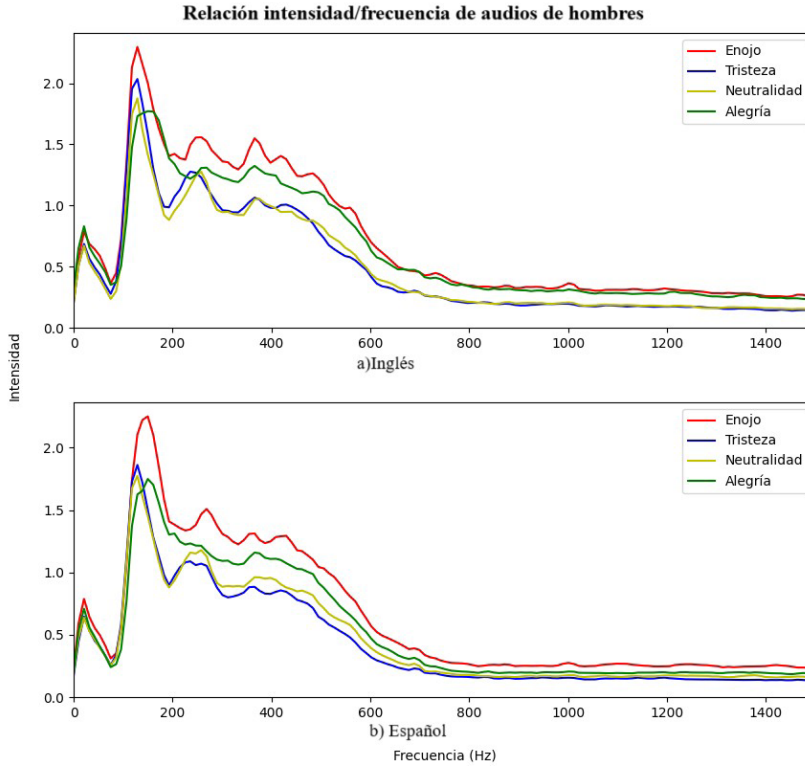
Figura 5. Señal de voz de emociones en mujeres en función de la frecuencia



Fuente: elaboración propia.

En la Figura 6 se observa que la F0 se encuentra por debajo de los 200 Hz, con componentes significativos y constantes en el rango de 200 a 400 Hz. A diferencia de las señales femeninas mostradas en la figura anterior, en este caso es posible diferenciar las señales correspondientes a cada emoción con mayor claridad a partir de su intensidad, destacando particularmente la emoción de enojo, que presenta una mayor diferenciación respecto a las demás. En el caso de los hombres no se observa una diferencia significativa entre los idiomas. En contraste, las mujeres muestran una variación más notable, especialmente en la intensidad, ya que al hablar en inglés utilizaron un mayor nivel de energía.

Figura 6. Señal de voz de emociones en hombres en función de la frecuencia



Fuente: elaboración propia.

c) Algoritmos de IA

En el lenguaje de programación Python se implementaron diferentes algoritmos de clasificación a partir del corpus de oraciones en inglés y español. Los algoritmos SVM y RF fueron seleccionados para generar modelos de aprendizaje supervisado y clasificar los audios del corpus de acuerdo con el tipo de emociones.

Las variables de entrada para ambos algoritmos corresponden a las características del espectro de frecuencias extraídas de fragmentos de audio de 93 milisegundos asociados a cada emoción. Para el aprendizaje de los algoritmos (SVM y RF) para clasificar las emociones, se utilizó el 67% del total de los datos, mientras que el 33% restante se empleó para comprobar su precisión en la etapa de evaluación.

En la Tabla 2 se pueden observar los resultados obtenidos en la clasificación de las emociones al implementar el algoritmo SVM. Estos resultados indican que la

Recibido: 29 de julio de 2025
Aceptado: 22 de agosto de 2025

emoción de la alegría perteneciente a la *dataset* en español tiene una mejor puntuación. De igual manera, la emoción tristeza del *dataset* en inglés tiene la mejor puntuación.

Tabla 2. Resultados del rendimiento del algoritmo SVM para los *dataset* SES-ED y SES-SD

Algoritmo	Dataset	Emoción	Predichos	Total	ACC/PREC	F-SCORE
SVM	SES-SD	Alegría	39	83	**0.47**	**0.41**
		Enojo	39	83	0.47	0.41
		Neutralidad	30	99	0.30	0.35
		Tristeza	37	85	0.44	0.45
	SES-ED	Alegría	36	85	0.42	0.39
		Enojo	24	83	0.27	0.29
		Neutralidad	27	99	0.27	0.35
		Tristeza	41	75	**0.55**	**0.43**

Fuente: elaboración propia.

La Tabla 3 muestra los resultados obtenidos en el modelo clasificador utilizando el algoritmo RF. En ambas *dataset*, la emoción de la tristeza presenta un mejor desempeño; sin embargo, en la *dataset* SES-SD, la tristeza tiene un desempeño superior al del SES-ED.

Tabla 3. Resultados del rendimiento del algoritmo RF para los *dataset* SES-ED y SES-SD.

	Dataset	Emoción	Predichos	Total	ACC/PREC	F-SCORE
RF	SES-SD	Alegría	31	83	0.37	0.34
		Enojo	37	83	0.45	0.42
		Neutralidad	15	99	0.15	0.22
		Tristeza	46	85	**0.54**	**0.45**
	SES-ED	Alegría	37	85	0.44	0.38
		Enojo	27	90	0.30	0.34
		Neutralidad	20	99	0.20	0.26
		Tristeza	36	75	**0.48**	**0.38**

Fuente: elaboración propia.

DISCUSIÓN

Los resultados muestran que la construcción de un corpus de emociones específico para estudiantes de origen mexicano hispanohablantes que se expresan en inglés es viable, y su aplicación en modelos de clasificación como SVM y RF logra una precisión aceptable, especialmente en emociones básicas como alegría y tristeza. Este resultado responde directamente a la primera pregunta de investigación, ya que, conforme al modelo dimensional de emociones de Russell (1980), dichas emociones se ubican en cuadrantes opuestos del plano afectivo (alegría en el primero y tristeza en el tercero), lo que permite evaluar con mayor claridad la calidad del etiquetado emocional.

La comparación con corpus como CREMA-D, SAVEE o TESS (desarrollados con hablantes nativos) reveló diferencias en la distribución espectral de las emociones, lo que sugiere que los modelos preentrenados en estos conjuntos de datos podrían no presentar una generalización aceptable para hablantes no nativos. Esto refuerza la necesidad de corpus adaptados al contexto sociolingüístico de quienes los usan, como se planteó en la primera pregunta de investigación sobre la calidad del etiquetado emocional.

Además, los patrones acústicos observados (como mayor energía en frecuencias altas para enojo y alegría) coinciden parcialmente con estudios previos (Shen et al., 2011; Hashem et al., 2023), pero con variaciones según el género y el idioma, lo que indica que estos factores influyen en la expresividad emocional. Esto aporta evidencia empírica para mejorar el etiquetado y el diseño de modelos de IA en contextos educativos multilingües.

CONCLUSIONES

Este estudio logró construir un corpus de oraciones emocionales específico para estudiantes de origen mexicano hispanohablantes que se expresan en inglés, etiquetado con representatividad lingüística y cultural. Este conjunto de datos permitió entrenar modelos de aprendizaje automático como Support Vector Machines (SVM) y Random Forest (RF), los cuales fueron utilizados tanto para evaluar la calidad del etiquetado emocional como para analizar la expresión vocal de emociones entre los idiomas español e inglés, lo que dio respuesta a las dos preguntas de investigación planteadas.

Respecto a la primera pregunta, los resultados evidencian que los algoritmos de IA pueden utilizarse como herramientas objetivas para validar la coherencia entre las etiquetas emocionales y las características acústicas de los audios. No

Recibido: 29 de julio de 2025
Aceptado: 22 de agosto de 2025

obstante, se identificó la necesidad de mejorar la calidad del etiquetado para incrementar la precisión de los modelos, lo cual puede abordarse en futuras investigaciones mediante técnicas semisupervisadas, validación cruzada con personas expertas o procedimientos automatizados más robustos.

En relación con la segunda pregunta de investigación, el análisis espectral reveló diferencias significativas entre las emociones, pero no entre los idiomas español e inglés en cuanto a su expresión vocal. Este hallazgo sugiere que, aunque el reconocimiento emocional debe considerar el perfil lingüístico de cada hablante, en el caso de estudiantes de México hispanohablantes, la manifestación vocal de las emociones se mantiene consistente incluso al utilizar un segundo idioma, como el inglés. Los modelos lograron clasificar con mayor precisión emociones como alegría y tristeza, mientras que emociones como enojo o sorpresa presentaron menor exactitud, lo que sugiere que es necesario ajustar los métodos de procesamiento de señales y segmentación acústica.

Este trabajo no solo aporta un corpus original y contextualizado, sino que también proporciona evidencia empírica sobre la importancia de adaptar los modelos de IA al contexto sociolingüístico de quienes los usan. Esto representa un paso relevante hacia el desarrollo de sistemas inclusivos y efectivos para el reconocimiento emocional en entornos educativos multilingües.

Finalmente, se recomienda que futuras investigaciones se orienten a:

- → Integrar modelos de aprendizaje profundo, como redes neuronales convolucionales (CNN).

- → Optimizar los procesos de etiquetado emocional mediante técnicas semisupervisadas o validación cruzada con personas expertas.

BIBLIOGRAFÍA

Abdul, Z. K. y Al-Talabani, A. K. (2022). Mel frequency cepstral coefficient and its applications: A review. *Ieee Access*, 10, 122136-122158.

Ashraf, M. R., Jana, Y., Umer, Q., Jaffar, M. A., Chung, S. y Ramay, W. Y. (2023). BERT-based sentiment analysis for low-resourced languages: A case study of Urdu language. *IEEE Access*, 11, http://dx.doi.org/10.1109/ACCESS.2023.3322101

Aljuhani, R. H., Alshutayri, A. y Alahdal, S. (2021). Arabic Speech Emotion Recognition from Saudi Dialect Corpus. *IEEE Access, 9*, 127081-127085. https://doi.org/10.1109/ACCESS.2021.3110992

Althari, G. y Alsulmi, M. (2022). Exploring Transformer-Based Learning for Negation Detection in Biomedical Texts. *IEEE Access, 10*, 83813-83825. https://doi.org/10.1109/ACCESS.2022.3197772

Amjad, M., Ashraf, N., Zhila, A., Sidorov, G., Zubiaga, A. y Gelbukh, A. (2021). Threatening language detection and target identification in Urdu tweets. *IEEE Access*, 9, 128302-128313.

Antoniou, N., Katsamanis, A., Giannakopoulos, T. y Narayanan, S. (2023). Designing and evaluating speech emotion recognition systems: A reality check case study with IEMOCAP. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1-5). *IEEE*.

Cao, H., Cooper, D.G., Keutmann, M.K., Gur, R.C., Nenkova, A. y Verma, R. (2014). Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE Trans. Affect. Comput.* 5 (4), 377–390

Egger, M., Ley, M. y Hanke, S. (2019). Emotion recognition from physiological signal analysis: A review. *Electronic Notes in Theoretical Computer Science, 343*, 35-55.

Gonzalez-Gomez, L. J., Hernandez-Munoz, S. M., Borja, A., Azofeifa, J. D., Noguez, J. y Caratozzolo, P. (2024). Analyzing Natural Language Processing Techniques to Extract Meaningful Information on Skills Acquisition from Textual Content. *IEEE Access, 12*, 139742-139757. https://doi.org/10.1109/ACCESS.2024.3465409

Hashem, A., Arif, M. y Alghamdi, M. (2023). Speech Emotion Recognition Approaches: A Systematic Review. *Speech Communication, 154*.

Kalateh, S., Estrada-Jimenez, L. A., Nikghadam-Hojjati, S. y Barata, J. (2024). A Systematic Review on Multimodal Emotion Recognition: Building Blocks, Current State, Applications, and Challenges. *IEEE Access, 12*, 103976-104019. https://doi.org/10.1109/ACCESS.2024.3430850

Kim, C.-G., Hwang, Y.-J. y Kamyod, C. (2022). A Study of Profanity Effect in Sentiment Analysis on Natural Language Processing Using Ann. *Journal of Web Engineering, 21*(3), 751-766. https://doi.org/10.13052/jwe1540-9589.2139

Laserna, M. S. y Torres, V. Á. (2022). ¿De qué hablamos cuando divulgamos sobre lingüística? Análisis de un corpus de textos divulgativos y aplicaciones al estudio terminológico de la semántica léxica. *ELUA Estudios de Lingüística Universidad de Alicante*, (38), 73-98. https://doi.org/10.14198/elua.22384

Mares, A., Diaz-Arango, G., Perez-Jacome-Friscione, J., Vazquez-Leal, H., Hernandez-Martinez, L., Huerta-Chua, J. y Dominguez-Chavez, A. (2025). Advancing Spanish Speech Emotion Recognition: A Comprehensive Benchmark of Pre-Trained Models. *Applied Sciences, 15*(8), 4340.

Meftah, A. H., Qamhan, M. A., Seddiq, Y., Alotaibi, Y. A. y Selouani, S. A. (2021). King Saud University emotions corpus: construction, analysis, evaluation, and comparison. *IEEE Access, 9*, 54201-54219.

Recibido: 29 de julio de 2025
Aceptado: 22 de agosto de 2025

Mifrah, S., y Benlahmar, E. H. (2022). Sentence-Level Sentiment Classification A Comparative Study Between Deep Learning Models. *Journal of ict Standardization, 10*(2), 339-352. https://doi.org/10.13052/jicts2245-800X.10213

Nassif, A. B., Shahin, I., Attili, I., Azzeh, M. y Shaalan, K. (2019). Speech recognition using deep neural networks: A systematic review. *IEEE Access, 7*, 19143-19165.

Nasution, A. H. y Onan, A. (2024). ChatGPT Label: Comparing the Quality of Human-Generated and LLM-Generated Annotations in Low-Resource Language NPL Tasks. *IEEE Access, 12*, 71876-71900. https://doi.org/10.1109/ACCESS.2024.3402809

Parry, J., Palaz, D., Clarke, G., Lecomte, P., Mead, R., Berger, M. y Hofer, G. (2019, September). Analysis of Deep Learning Architectures for Cross-Corpus Speech Emotion Recognition. *Interspeech* (1656-1660).

Russell, J. A. (1980). A Circumplex Model of Affect. *Journal of Personality and Social Psychology, 39*(6), 1161.

Rzayeva, Z. y Alasgarov, E. (2019). Facial emotion recognition using convolutional neural networks. En *2019 IEEE 13th international conference on application of information and communication technologies (AICT)* (pp. 1-5). IEEE.

Sen, O., Fuad, M., Islam, Md. N., Rabbi, J., Masud, M., Hasan, Md. K., Awal, Md. A., Ahmed Fime, A., Hasan Fuad, Md. T., Sikder, D. y Raihan Iftee, Md. A. (2022). Bangla Natural Language Processing: A Comprehensive Analysis of Classical, Machine Learning, and Deep Learning-Based Methods. *IEEE Access, 10*, 38999–39044. https://doi.org/10.1109/ACCESS.2022.3165563

Shen, P., Changjun, Z. y Chen, X. (2011, August). Automatic speech emotion recognition using support vector machine. En *Proceedings of 2011 international conference on electronic & mechanical engineering and information technology*. *IEEE,* vol. 2, pp. 621-625

Shim, M., Jin, M. J., Im, C. y Lee, S. (2019). Machine-learning-based classification between post-traumatic stress disorder and major depressive disorder using P300 features. *NeuroImage Clinical, 24*, 102001. https://doi.org/10.1016/j.nicl.2019.102001

Sultana, S., Rahman, M. S., Selim, M. R. y Iqbal, M. Z. (2021). SUST Bangla Emotional Speech Corpus (SUBESCO): An audio-only emotional speech corpus for Bangla. *PLoS ONE, 16*(4), e0250173. https://doi.org/10.1371/journal.pone.0250173

Traunmüller, H., y Eriksson, A. (1995). *The frequency range of the voice fundamental in the speech of male and female adults*. Manuscrito no publicado.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., y Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. Von Lux-

burg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, y R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 30). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Zehra, W., Javed, A. R., Jalil, Z., Khan, H. U. y Gadekallu, T. R. (2021). Cross corpus multi-lingual speech emotion recognition using ensemble learning. *Complex & Intelligent Systems*, 7(4), 1845-1854.

Zha, D., Bhat, Z. P., Lai, K. H., Yang, F., Jiang, Z., Zhong, S. y Hu, X. (2025). Data-centric artificial intelligence: A survey. *ACM Computing Surveys*, *57*(5), 1-42.