



Universidad
Internacional
de Valencia

Generación de Contenido Autoayuda y Apoyo Psicológico Personalizado

Titulación:
Máster en Inteligencia
Artificial
Curso académico
2023 – 2024

Alumno/a: Motta Valero Luis
Angel
D.N.I: 1117549110
Director/a de TFM: Cigales
Canga Jesús

Convocatoria:
Abril

15 abril 2024

De:
 Planeta Formación y Universidades

1. Preliminares

Resumen

En la sociedad contemporánea, el apoyo psicológico emerge como una necesidad vital, dada la creciente importancia del bienestar emocional. Este tipo de apoyo no solo mejora la calidad de vida individual, sino que también impacta en el tejido social y en la productividad económica, posicionando la atención a la salud mental como un pilar fundamental para construir sociedades más resilientes y saludables.

Numerosas contribuciones han surgido en este ámbito, destacando propuestas basadas en modelos de Aprendizaje Automático. Estos modelos buscan mejorar la interacción entre humanos y sistemas de inteligencia artificial mediante la generación de respuestas empáticas y el análisis multimodal de sentimientos. Sin embargo, la efectividad de estas propuestas enfrenta desafíos importantes, desde la necesidad de considerar aspectos que influyen en el estado emocional del individuo, como la intensidad de sentimiento expresado en su interacción con los demás y las distintas modalidades de comunicación que dan indicios del estado emocional del sujeto.

El trabajo final de máster se enfoca en el desarrollo de un sistema conversacional impulsado por inteligencia artificial, denominado EmpAI, con el objetivo principal de proporcionar apoyo psicológico personalizado a los usuarios mediante la generación de contenido autoayuda y respuestas empáticas. Se propone integrar tecnologías enfocadas en el Procesamiento de Lenguaje Natural para mejorar la comprensión del contexto emocional del usuario y ofrecer respuestas más efectivas. Accede al código desde el [repositorio de GitHub](#).

A pesar de las altas expectativas, los resultados del trabajo revelaron una baja calidad en las respuestas generadas por el modelo desarrollado, careciendo de coherencia y relevancia con respecto al input proporcionado por los usuarios. Estos hallazgos resaltan la necesidad de continuar investigando y refinando los modelos de inteligencia artificial para mejorar su capacidad de ofrecer apoyo psicológico personalizado de manera efectiva.

Abstract

In contemporary society, psychological support emerges as a vital need, given the growing importance of emotional well-being. This type of support not only improves individual quality of life, but also impacts the social fabric and economic productivity, positioning mental health care as a fundamental pillar for building more resilient and healthy societies.

Numerous contributions have emerged in this field, highlighting proposals based on machine learning models. These models seek to improve the interaction between humans and artificial intelligence systems by generating empathic responses and multimodal sentiment analysis. However, the effectiveness of these proposals faces significant challenges, from the need to consider aspects that influence the emotional state of the individual, such as the intensity of feeling expressed in their interaction with others and the different modes of communication that give clues to the emotional state of the subject.

The final master work focuses on the development of an artificial intelligence-driven conversational system, called EmpAI, with the main objective of providing personalized psychological support to users through the generation of self-help content and empathic responses. It proposes to integrate technologies focused on Natural Language Processing to improve the understanding of the user's emotional context and provide more effective responses. Access to the code from the GitHub repository.

Despite the high expectations, the results of the work revealed a low quality in the responses generated by the developed model, lacking coherence and relevance with respect to the input provided by the users. These findings highlight the need for further research and refinement of artificial intelligence models to improve their ability to deliver personalized psychological support effectively.

Palabras clave: Análisis multimodal; Apoyo psicológico; Bienestar emocional; Conversación personalizada; Empatía; Generación de contenido autoayuda; Procesamiento de Lenguaje Natural; Respuestas empáticas; Salud mental; Tecnologías de apoyo emocional.

2. Prefacio

Es con gran entusiasmo que presento este trabajo de investigación, el resultado de un arduo y apasionante proceso académico. La motivación que impulsó esta investigación surge de la profunda convicción en la importancia del tema abordado y su relevancia en el contexto actual. En la presente investigación se propuso explorar la generación de contenido autoayuda y apoyo psicológico personalizado desde un enfoque práctico, con el objetivo de contribuir al mejoramiento del estado emocional del usuario mediante respuestas empáticas.

Durante el desarrollo de este proyecto, se tuvo el privilegio de contar con el apoyo incondicional de diversas personas e instituciones, a quienes deseo expresar mi más sincero agradecimiento. En primer lugar, agradezco a la Universidad Internacional de Valencia – VIU, por brindar el espacio y los recursos necesarios para llevar a cabo este trabajo de investigación. Su compromiso con la excelencia académica ha sido fundamental para mi formación y desarrollo profesional.

Asimismo, quiero extender mi gratitud a los docentes que, con su dedicación y conocimiento, me acompañaron a lo largo de todo el proceso de formación. Sus orientaciones y enseñanzas han sido de inestimable valor para el éxito de este proyecto.

A mis estimados compañeros, les agradezco por su apoyo mutuo, comprensión y colaboración durante esta travesía académica. Su compañerismo y aliento han sido un pilar fundamental en los momentos de desafío.

A mi tutor, el docente Jesús Cigales, le estoy profundamente agradecido por su orientación profesional, su paciencia y su constante motivación. Su guía ha sido fundamental para el desarrollo y la finalización de este trabajo.

Finalmente, dedico mi esfuerzo a mis padres y hermanos, cuyo amor, apoyo incondicional y sacrificio han sido mi mayor inspiración y fortaleza. Su aliento y comprensión han sido el motor que me ha impulsado a alcanzar mis metas académicas y profesionales.

Índice

1. Preliminares	1
Resumen.....	1
Abstract.....	2
2. Prefacio.....	3
3. Introducción	7
4. Marco Teórico	8
Aprendizaje Automático	8
Aprendizaje Profundo.....	8
Procesamiento del Lenguaje Natural (NLP)	12
Redes Neuronales Recurrentes (RNN)	12
Modelo Transformers	16
AutoEncoder Variacional (VAE)	17
Modelos de Lenguaje	18
Preprocesamiento en el modelado de lenguaje.....	18
Data Collators	20
5. Estado del Arte	21
Generación de respuesta empática.....	21
Análisis Multimodal de Sentimiento	25
Traducción Multimodal	27
6. Material.....	30
EmpatheticDialogues Dataset	30
Mental health counseling conversation.....	3
Psych8k	4
Counsel Chat	5
Emotional Support Conversation – ESConv	7
Prompt-Aware margin Ranking (PAIR).....	8
EmpathetiCounseling Dataset	10
CMU-MOSI y CMU-MOSEI Datasets	10
API SeamlessM4T	13
7. Métodos.....	16
Elección del modelo idóneo.....	16
Funcionamiento general.....	22
Implementación.....	28
8. Resultados preliminares.....	31
RoBERTa-GPT2	31
CM-BERT.....	37
Propuestas de mejora	38
9. Conclusiones	39
Bibliografía.....	40
10. Anexos.....	43
Anexo I.....	43
Anexo II.....	44

Índice de figuras

Figura 1. Red neuronal de tres capas totalmente conectada.....	9
Figura 2. Funciones de activación.....	10
Figura 3. Red neuronal recurrente (RNN)	13
Figura 4. Red neuronal recurrente con una arquitectura codificador-decodificador	13
Figura 5. Red Neuronal Recurrente Bidireccional Profunda (Deep Bidirectional RNN).....	14
Figura 6. Unidades Long Short-Term Memory	14
Figura 7. Traducción automática con atención	16
Figura 8. Arquitectura del Modelo Transformers	16
Figura 9. Arquitecturas NVDM, para el modelado de documentos; y NASM, para la selección de preguntas-respuestas.....	17
Figura 10. Ejemplo del proceso de tokenización	19
Figura 11. Principales objetos Data Collator.....	20
Figura 12. Distribución de etiquetas de conversación dentro del conjunto de entrenamiento EMPATHETICDIALOGUES	1
Figura 13. Longitud mínima y máxima en Empathetic Dialogues Dataset	2
Figura 14. Longitud máxima y mínima de Mental Health Counseling Conversation Dataset4	
Figura 15. Longitud máxima y mínima de Psych8k Dataset	5
Figura 16. Longitud máxima y mínima de Counsel Chat Dataset	6
Figura 17. Longitud máxima y mínima en ESConv Dataset.....	8
Figura 18. Longitud máxima y mínima en PAIR Dataset	9
Figura 19. Longitud máxima y mínima de EmpathetiCounseling Dataset	10
Figura 20. Arquitectura codificador-decodificador de RoBERTa-GPT2	19
Figura 21. Estrategia 1: Modelo RoBERTa-GPT2	19
Figura 22. Estrategia 2: Modelo RoBERTa-DialoGPT	20
Figura 23. Estrategia 3: Modelo MentalRoBERTa-DialoGPT.....	20
Figura 24. Estrategia 4: Modelo DialoGPT	20
Figura 25. Arquitectura de CM-BERT y de su Atención Multimodal Enmascarada.....	21
Figura 26. Pasos de entrenamiento del modelo CM-BERT	22
Figura 27. Logo del sistema implementado.....	29
Figura 28. Interfaz de la App EmpAI	30
Figura 29. Resultados de entrenamiento – RoBERTa_EmpAI	32
Figura 30. Resultados de entrenamiento – MentalRoBERTa_EmpAI.....	33
Figura 31. Resultados de entrenamiento - GPT-2_EmpAI.....	34
Figura 32. Resultados de entrenamiento - DialoGPT_EmpAI.....	35
Figura 33. Resultados de entrenamiento - DialoGPT_EmpAI_FineTuned.....	36
Figura 34. Resultados de evaluación - CM-BERT	37

Índice de tablas

Tabla 1. Principales algoritmos de tokenización	19
Tabla 2. Resultados de la evaluación automática	22
Tabla 3. Resultados de evaluación del modelo RoBERTa-GPT2	23
Tabla 4. Comparación del rendimiento en el reconocimiento de palabras causantes de emociones	24
Tabla 5. Comparación entre las métricas clave de SPECTRA y el método SOTA anterior en cinco conjuntos de datos.	25
Tabla 6. Resultados experimentales de CM-BERT en el conjunto de datos CMU-MOSI...26	
Tabla 7. Comparación de resultados de SeamlessMT4 en multitarea X2T con respecto al SOTA	28
Tabla 8. Comparación del rendimiento del modelo PortaSpeech	29
Tabla 9. Dos ejemplos del conjunto de entrenamiento EMPATHETICDIALOGUES.....31	
Tabla 10. Las 3 palabras de contenido más utilizadas por el hablante oyente en cada etiqueta de emociones	1
Tabla 11. Ejemplo del conjunto Mental health conuseling conversation	3
Tabla 12. Ejemplo del conjunto de entrenamiento Psych8k	5
Tabla 13. Ejemplo del conjunto de entrenamiento Counsel chat	6
Tabla 14. Ejemplo del conjunto de entrenamiento ESConv	7
Tabla 15. Ejemplo del conjunto de entrenamiento PAIR	9
Tabla 16. Resumen de las estadísticas del conjunto de datos MOSI	12
Tabla 17. Resumen de las estadísticas del conjunto de datos CMU-MOSEI.....12	
Tabla 18. Comparación del rendimiento entre los modelos de Generación de Respuesta Empática	17
Tabla 19. Comparación del rendimiento entre los modelos de Análisis Multimodal de Sentimiento	17
Tabla 20. Casos en la entrada y salida del modelo	23
Tabla 21. Ejemplo de inferencia con la estrategia 3	36

3. Introducción

En el contexto actual, donde el bienestar emocional cobra una relevancia cada vez mayor, la demanda de sistemas de apoyo psicológico personalizado se ha incrementado significativamente. En este sentido, el desarrollo de herramientas tecnológicas capaces de ofrecer un acompañamiento empático y eficaz se convierte en un objetivo primordial.

La salud emocional influye directamente en la calidad de vida, siendo un componente esencial para el bienestar físico, como sostiene la médico-psiquiatra Marian Rojas Estapé (2018, pág. 159), la acumulación de emociones reprimidas por la falta de aceptación y relevancia, en algún momento puede desencadenar enfermedades psicosomáticas. Aunque la sociedad actual tiende a ver el bloqueo emocional como muestra de fortaleza, mostrarse débil y vulnerable a los demás en busca de apoyo es, en realidad, un acto valiente y efectivo para el bienestar propio.

Un estudio hecho por Schütz Balistieri & Mara de Melo Tavares (2013), demuestra que el respaldo emocional juega un papel crucial durante períodos de salud precaria. En estos momentos de dificultad amigos y familiares se convierten en verdaderos psicólogos personales, actuando como “personas vitamina” que entienden los silencios, miedos y preocupaciones del afectado sin miras a la crítica, transmitiendo serenidad y optimismo, en un efecto positivo que mitiga el dolor (Rojas Estapé, 2018, págs. 23-32).

El acto médico de un especialista en psicología, según Mondragón (2017), va más allá del tratamiento clínico y abraza la empatía como herramienta para aliviar cansancios, estimular esfuerzos continuos y superar escapes depresivos; comprende, además, el afecto compasivo desde la perspectiva del acompañamiento hacia el paciente.

En el marco del presente trabajo final de máster titulado "Generación de Contenido Autoayuda y Apoyo Psicológico Personalizado", y apoyado por la postura de los autores mencionados, se propone la creación de un sistema de conversación innovador que integre diversas tecnologías para brindar un soporte integral a los usuarios; más concretamente, comprende el desarrollo de un sistema de Inteligencia Artificial basado en un agente de diálogo para la generación de contenido autoayuda y apoyo psicológico personalizado. El objetivo es contribuir al mejoramiento del estado emocional del usuario mediante respuestas empáticas desde las diferentes modalidades que adopta el flujo conversacional entre las personas, dentro de las que incluye la modalidad lingüística y acústica (Zadeh A. , y otros, 2018), que facilite la comprensión del contexto y sentimientos presentes en la conversación.

Este sistema se cimenta en la combinación de tres pilares fundamentales: la Generación de Respuestas Empáticas, el Análisis Multimodal de Sentimientos y la Traducción Multimodal. La generación de respuestas empáticas se enfoca en ofrecer un apoyo emocional genuino mediante la comprensión y la empatía hacia los sentimientos y emociones de la persona. A su vez, el análisis multimodal de sentimientos permite capturar y comprender las expresiones emocionales del usuario a través de diferentes modalidades, como texto, voz y video. Por último, la traducción multimodal facilita la comunicación fluida en distintas modalidades – al igual que en distintos idiomas – garantizando que el sistema pueda llegar a una audiencia global y diversa.

4. Marco Teórico

La solución de tareas por parte de computadoras, basada en la toma de decisiones dada una predicción, donde se evidencia una interacción humano-máquina, como la generación de contenido, se da gracias al amplio campo denominado Inteligencia Artificial (IA), una disciplina que implementa una combinación de tecnologías de las ciencias de la computación para dotar a una máquina u ordenador de capacidades para reproducir funciones cognitivas asociadas a la inteligencia humana (Google Research, s.f.).

Aprendizaje Automático

Para lograr lo anterior, se hace uso de algoritmos que componen un subcampo de IA conocido como Aprendizaje Automático (Machine Learning en inglés – ML), un medio para la extracción de conocimiento analizando grandes cantidades de datos, de los que aprende y mejora con el tiempo a partir de la experiencia adquirida con los mismos, con el propósito de tomar decisiones que le permitan resolver el problema para el cual ha sido creado (Google Research, s.f.).

En resumen, un modelo de ML se clasifica acorde a la forma de los datos, naturaleza del problema en cuestión e inclusive, a la forma en que se podría afrontar dicho aprendizaje, entre: **Aprendizaje supervisado**, donde el conjunto de datos se estructura mediante atributos (X) y clases (Y) asociada a estos, que será la etiqueta objetivo a predecir; **Aprendizaje no supervisado**, cuyo objetivo es aprender a identificar patrones en los datos, los cuales son no etiquetados, por ello, son clasificados en grupos denominados “clústeres”; **Aprendizaje semi-supervisado** como una combinación entre los anteriores, existiendo etiquetas en algunos ejemplos del conjunto de datos (Google Research, s.f.).

Aprendizaje Profundo

Una aplicación del ML inspirada en las neuronas cerebrales y sus conexiones es el Aprendizaje Profundo, que engloba todos los métodos, arquitecturas y aplicaciones que implican representaciones de redes neuronales, las cuales se componen de funciones tanto lineales (mediante una multiplicación matricial), como no lineales (una función de activación no lineal en específico), realizando una propagación hacia adelante o **Forward propagation** de las activaciones de la capa $\ell - 1$ a la capa ℓ de la red para inicializar los pesos de las conexiones W^ℓ a un valor distinto de cero a partir de pequeños valores aleatorios utilizando una distribución normal $\frac{N(0,1)}{\sqrt{n}}$, donde n es el número de conexiones en una unidad de activación (neurona) (Cambridge University Press, 2022). El pseudocódigo para el algoritmo de Forward Propagation se provee a continuación.

Algoritmo. Forward Propagation

Dado los pesos iniciales W^1, \dots, W^L

Dado un vector de ejemplo de datos x^1

Por cada capa $\ell = 1, \dots, L$ hacer:

$$x^{\ell+1} = f^{\ell}(W^{\ell T} x^{\ell})$$

Seguidamente, se optimizan estos pesos aplicando el método de **Backpropagation**, que consiste en el cálculo de gradientes de una función de pérdida con respecto a todos los parámetros en cada capa ℓ , para encontrar un mínimo local dando pasos en la dirección del descenso más pronunciado y acorde a un factor o ratio de aprendizaje (**Cambridge University Press, 2022**), de modo que hace uso del algoritmo de *Descenso del gradiente*:

Algoritmo. Descenso del Gradiente

Dado un punto inicial $x \in \text{dom}f$

Repetir:

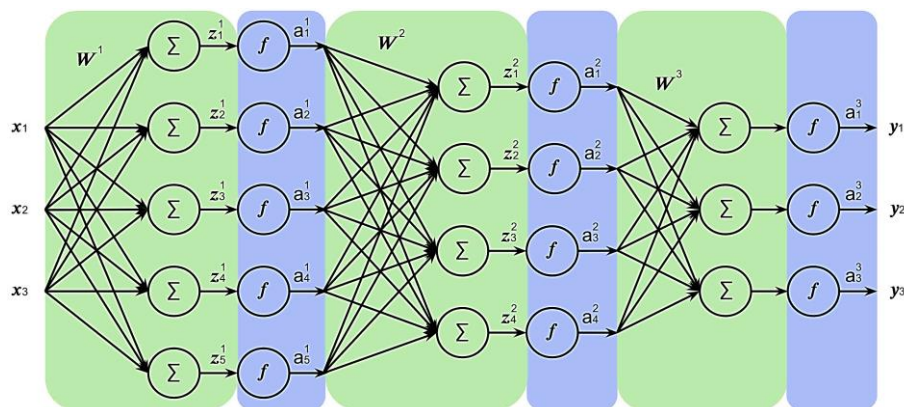
Determina la dirección del descenso $-\nabla f(x)$

Elige un escalar α

Actualiza $x := x - \alpha \nabla f(x)$

Hasta que se cumpla el criterio de parada

Figura 1. Red neuronal de tres capas totalmente conectada



Nota. Se compone de una multiplicación matricial W (verde) y una aplicación puntual de una función no lineal f (azul). Fuente: (Cambridge University Press, 2022)

Son muy pocos los problemas que poseen un comportamiento lineal, con los cual, se debe aplicar un tipo de activación sobre el resultado de la multiplicación matricial en cada capa de una red neuronal; estas funciones efectuadas sobre una preactivación z son diferenciables en $f: \mathbb{R} \mapsto \mathbb{R}$ (**Cambridge University Press, 2022**). Las **funciones de activación** no lineales típicas incluyen: la función Sigmoide utilizada habitualmente en la regresión logística para construir un clasificador tomando la sigmoidea de la regresión lineal, asignando la entrada z a $(0, 1)$ mediante

$$f(z) = \frac{1}{1 + e^z};$$

la función Tangente hiperbólica (Tanh), que mapea la entrada z a $(-1, 1)$ mediante

$$f(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}};$$

la función Rectified Linear Unit (ReLU), comúnmente aplicada sobre las capas ocultas de la red neuronal, y asigna los valores negativos a cero, de acuerdo con la consigna

$$g(z) = \max(0, z);$$

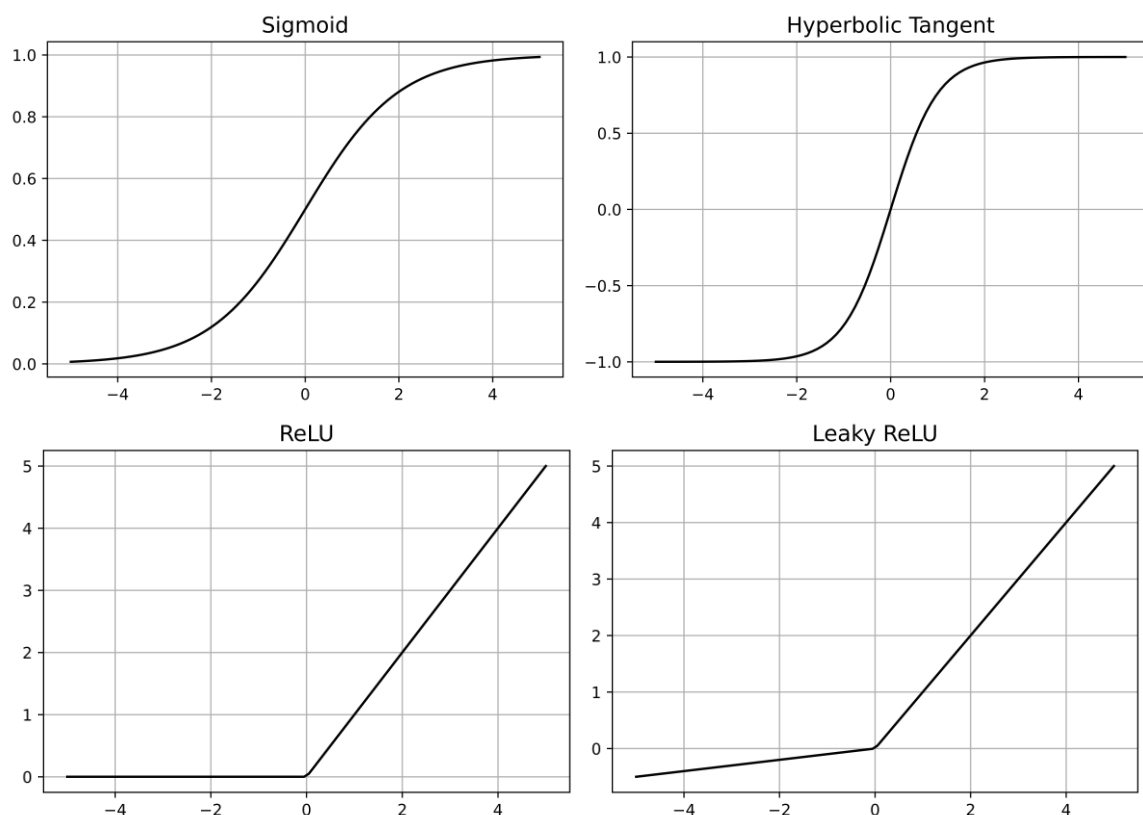
la función Leaky ReLU, que suaviza la restricción de la función ReLU respecto a los valores z negativos, y se define, para $\alpha \geq 0$, por

$$g(z) = z_+ - \alpha z_- = \max(0, z) - \alpha \max(0, -z);$$

y la función Softmax, especializada para aplicarse sobre la última capa L o capa de salida para problemas de clasificación multiclase, asignando un vector $z \in \mathbb{R}^k$ a un vector $f(z) \in [0, 1]^k$ cuya suma es 1: $\sum_{c=1}^k f(z)_c = 1$. La función softmax para la clase i viene dada por:

$$f^L(z^L)_i = \frac{e^{z_i^L}}{\sum_{c=1}^k e^{z_c^L}}$$

Figura 2. Funciones de activación



Fuente: (Cambridge University Press, 2022)

Dada la predicción de una salida para una entrada cuya etiqueta real es conocida, se debe validar la precisión del modelo respecto a este resultado, la misma se efectúa mediante la implementación de una **función de pérdida** o función de coste sobre un conjunto de datos de validación. Considere la etiqueta real y para una entrada x y llámela verdad absoluta (ground truth), al igual que una salida $f(x) = \hat{y} = \mathbf{a}^L$ (\mathbf{a}^L es la activación \mathbf{a} hecha sobre la última capa L); el papel que cumple la función de pérdida es comparar la salida \hat{y} con la etiqueta y para una entrada x y el objetivo es minimizar dicha pérdida, que no garantiza un mínimo global (**Cambridge University Press, 2022**); así, función de pérdida para un único ejemplo de entrada y etiqueta de salida es $\mathcal{L}(y, F(x))$ y la pérdida media sobre todos los ejemplos, o coste, es:

$$\frac{1}{m} \sum_{i=1}^m \mathcal{L}(y^i, \hat{y}^i)$$

Las funciones de pérdida también se utilizan al momento de optimizar los pesos W propagando hacia delante las entradas x a través de la red para obtener las etiquetas de salida $\hat{y}^i = F(x^i, W)$, así:

$$\min_W \frac{1}{m} \sum_{i=1}^m \mathcal{L}(y^i, F(x^i, W))$$

Adicionalmente, sobre la implementación de la red se puede añadir un término de regularización común $\mathcal{R}(W)$ a la función de pérdida para preferir modelos sencillos y evitar el sobreajuste (overfitting). Las funciones de pérdida comunes son el error cuadrático medio, con $\mathcal{L}(y^i, \hat{y}^i)$ definido por:

$$\mathcal{L}(y^i, \hat{y}^i) = (y^i - \hat{y}^i)^2,$$

y, la función de pérdida logística, que se define por:

$$\mathcal{L}(y^i, \hat{y}^i) = -y^i \log(\hat{y}^i) - (1 - y^i) \log(1 - \hat{y}^i)$$

Que es un caso especial para la clasificación binaria $k = 2$:

$$\mathcal{L}(y^i, \hat{y}^i) = - \sum_{c=1}^k I\{y^i = c\} \log p(y = c | x^i, W)$$

para k clases, donde I es la función indicadora tal que $I\{true\} = 1$ e $I\{false\} = 0$, y $p(y = k | x^i, W)$ es un coeficiente softmax. Para el caso especial de la regresión logística, el error cuadrático medio no es convexo, mientras que la pérdida de regresión logística es convexa.

Una práctica común y que mejora la convergencia del descenso del gradiente es la **normalización** de los datos de entrada x al score estándar mediante $x = \frac{x - \mu}{\sigma}$ donde μ es la media y σ^2 es la varianza; similarmente, la técnica de *batch normalization* normaliza cada

lote de entrada para cada capa de la red. Una **inicialización** común es una distribución normal con media cero y varianza

$$\sigma^2 = \frac{2}{n_{l-1} + n_l},$$

donde n_{l-1} y n_l son el número de unidades de activación en las capas anterior y actual de los pesos, respectivamente (Cambridge University Press, 2022).

Procesamiento del Lenguaje Natural (NLP)

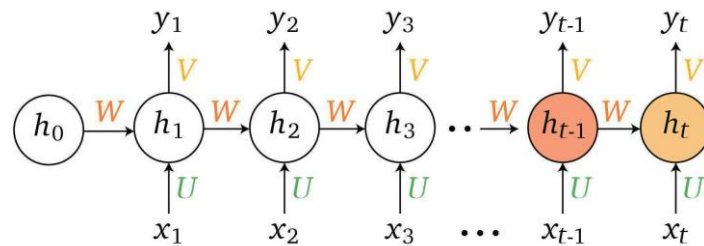
Una aplicación de modelos de Deep Learning que destaca se encuentra dentro del uso de representaciones de secuencias temporales, que hacen uso de los denominados Modelos de Lenguaje Natural (NLP, por su siglas en inglés), compartiendo pesos en el tiempo desde una representación del lenguaje enmarcado dentro del paradigma del Procesamiento del Lenguaje Natural, que es el uso del lenguaje humano por parte de un ordenador, definiendo una distribución de probabilidades sobre secuencias de palabras, caracteres o bytes en un lenguaje natural (Goodfellow, Bengio, & Courville, 2016).

Uno de los modelos de lenguaje más sencillos se denomina **Bag of Words**, un unigrama que almacena la frecuencia $TF(x, d)$ del término x en un único documento d ; en el mismo, las palabras se normalizan a minúsculas y se separan eliminando sus sufijos, también se eliminan las palabras vacías comunes (por ejemplo, un, una, el, la, etc.). Una representación que difiere de la anterior al conservar la información referente al orden de las palabras en una frase es un **vector de características**; sin embargo, esto lo limita con respecto a frases que son semánticamente idénticas. Un modelo 2-gram o bi-grama (una secuencia de 2 palabras adyacentes), como lo es un **modelo de Markov**, no modela dependencias a largo plazo, es decir, la probabilidad de una palabra en un enunciado depende únicamente de la palabra anterior, ignorando prefijos que podrían aportar mayor información (**Cambridge University Press, 2022**); esta limitación es solucionada por el modelo de Redes Neuronales Recurrentes (RNN, por sus siglas en inglés).

Redes Neuronales Recurrentes (RNN)

Las redes neuronales recurrentes son diseñadas para modelar dependencias a largo plazo mediante el procesamiento de vectores con secuencias de entrada x_1, \dots, x_t a través de unidades ocultas h_0, h_1, \dots, h_t para formar vectores con salidas y_1, \dots, y_t que comparten parámetros en el tiempo, permitiendo que ambos vectores en un mismo ejemplo tengan longitudes diferentes, donde también se conserva el orden de las palabras.

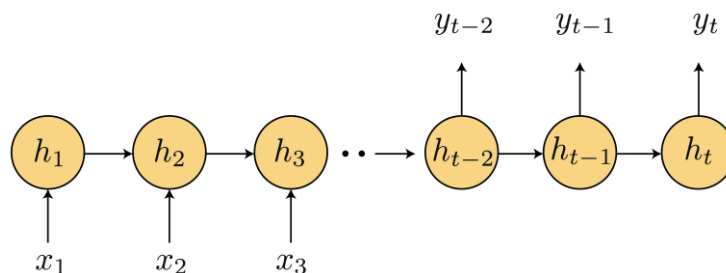
Figura 3. Red neuronal recurrente (RNN)



Nota. Proceso de Forward propagation. Fuente: (Cambridge University Press, 2022)

Las redes neuronales recurrentes asignan secuencias de entrada a secuencias de salida de longitudes variables. Esta asignación puede ser de uno a muchos, de muchos a uno o de muchos a muchos. También se puede utilizar un mapeo de muchos a muchos en una arquitectura de codificador-decodificador, utilizada en tareas como la traducción automática, en la que la secuencia de entrada, que es una frase en un idioma, se codifica y luego se descodifica en una frase de salida en otro idioma.

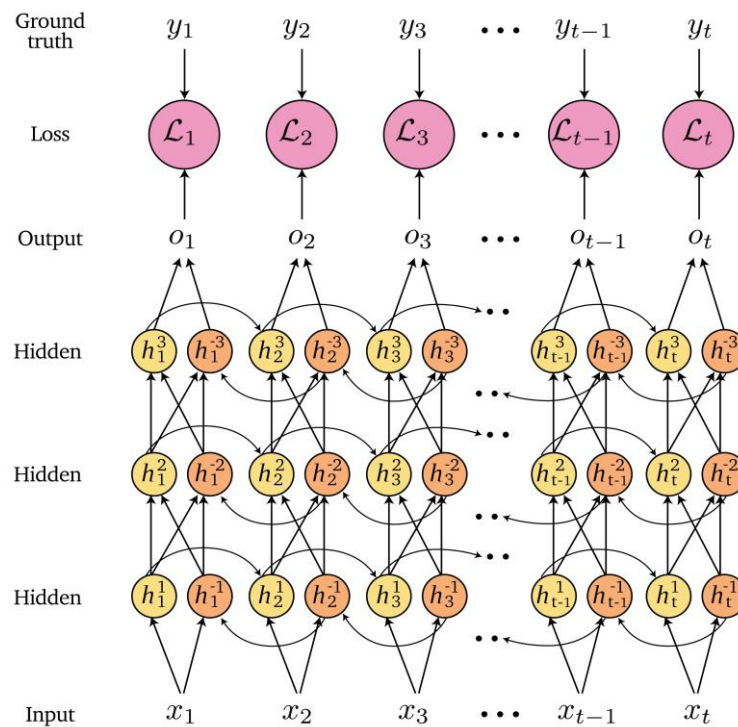
Figura 4. Red neuronal recurrente con una arquitectura codificador-decodificador



Nota. Mapeo de muchos a muchos. Fuente: (Cambridge University Press, 2022)

Las RNN de tipo bidireccionales combinan una RNN que avanza en el tiempo desde el principio de la secuencia con otra que retrocede en el tiempo desde el final de la secuencia (Goodfellow, Bengio, & Courville, 2016), de este modo, modelan las dependencias secuenciales hacia delante y hacia atrás, y las RNN profundas utilizan múltiples capas ocultas. Las RNN simples son difíciles de entrenar, ya que las señales de error que retroceden en el tiempo explotan o desaparecen. Por eso, las unidades ocultas se sustituyen por puertas simples que se entrenan fácilmente. El apilamiento de múltiples capas ocultas bidireccionales da como resultado una RNN bidireccional profunda.

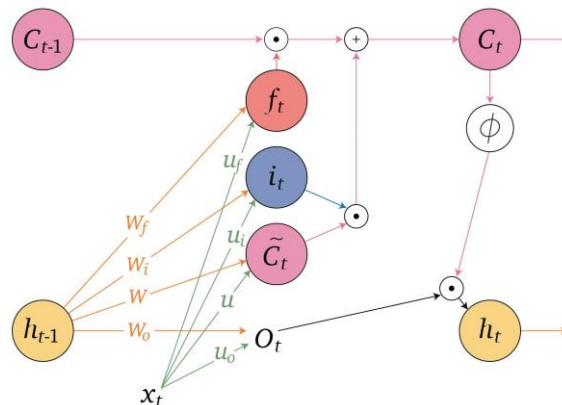
Figura 5. Red Neuronal Recurrente Bidireccional Profunda (Deep Bidirectional RNN)



Fuente: (Cambridge University Press, 2022)

El tipo de modelo secuencial RNN conocido como **Long Short-Term Memory – LSTM**, consiste en unidades ubicadas una tras otra, donde las salidas del estado oculto h_t y de la celda de memoria c_t de una unidad sirven como entradas del estado oculto y de la celda de memoria de la siguiente unidad, introduciendo, además, bucles propios con un flujo del gradiente en largos periodos, cuyos pesos varían dependiendo del contexto, permitiendo que la escala temporal de integración cambie en función de la secuencia de entrada (**Goodfellow, Bengio, & Courville, 2016**). El LSTM consiste en una puerta de olvido f_t , una puerta de entrada i_t , una puerta de salida o_t , y una memoria candidata \tilde{c}_t , como se muestra:

Figura 6. Unidades Long Short-Term Memory



Fuente: (Cambridge University Press, 2022)

Una tarea adicional en el modelado de secuencias es el **reconocimiento automático del habla**, que consiste en convertir una señal acústica que contiene un enunciado hablado en lenguaje natural en la correspondiente secuencia de palabras, al crear una función f_{ASR}^* que calcule la secuencia lingüística $y = (y_1, y_2, \dots, y_N)$ más probable dada la secuencia acústica $X = (x^1, x^2, \dots, x^T)$, mediante una distribución condicional P^* (**Goodfellow, Bengio, & Courville, 2016**):

$$f_{ASR}^*(X) = \arg \max_y P^*(y|X = X)$$

Una aplicación de las unidades LSTM se evidencia en modelos de secuencia a secuencia (seq2seq) dentro de su composición de codificador-decodificador, útiles en tareas que requieren considerar secuencias enteras, u oraciones, como entrada y salida. El codificador puede ser una red bidireccional y profunda, que codifica la secuencia de entrada (x_1, \dots, x_s) en un vector de contexto como salida $z = \text{codificador}(x_1, \dots, x_s)$, que es recibido por el decodificador como entrada en un primer vector de estado oculto para generar una secuencia de salida $(y_1, \dots, y_t) = \text{decodificador}(z)$. Los modelos codificador y decodificador se entrenan de extremo a extremo de forma que

$$(y_1, \dots, y_t) = \text{decodificador}(\text{codificador}(x_1, \dots, x_s)).$$

Modelos como **Seq2seq** han sido recientemente mejorados incorporando mecanismos de atención entre diferentes partes de la secuencia de salida sobre partes de la secuencia de entrada, de modo que el decodificador considera la secuencia del codificador en su totalidad a diferencia del proceso secuencial original (**Cambridge University Press, 2022**); así mismo, la entrada a las unidades ocultas del decodificador en una LSTM bidireccional es un conjunto de vectores de contexto c_i para cada paso de tiempo i , calculados como:

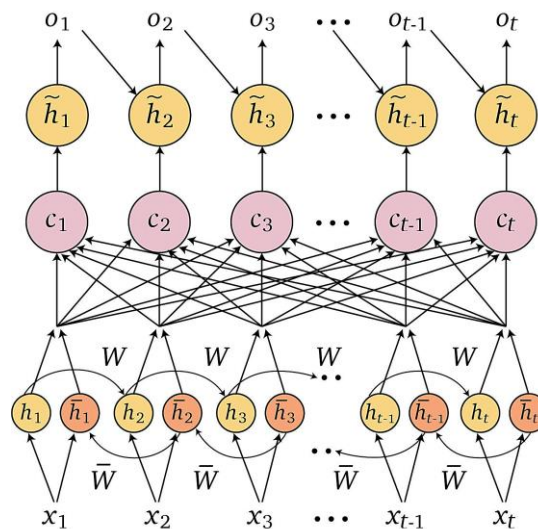
$$c_i = \sum_j \alpha_{i,j} [h_j; \bar{h}_j]^T$$

Los pesos $\alpha_{i,j}$ suman 1, $\sum_j \alpha_{i,j} = 1$, y son la cantidad de atención que la palabra de salida o_i presta a la palabra de entrada x_j :

$$\alpha_{i,j} = \frac{\exp(s_{i,j})}{\sum_k \exp(s_{i,k})}$$

donde $s_{i,j}$ es una función de las unidades ocultas del codificador $[h_j; \bar{h}_j]^T$ y las unidades ocultas del decodificador \bar{h}_{t-1} .

Figura 7. Traducción automática con atención

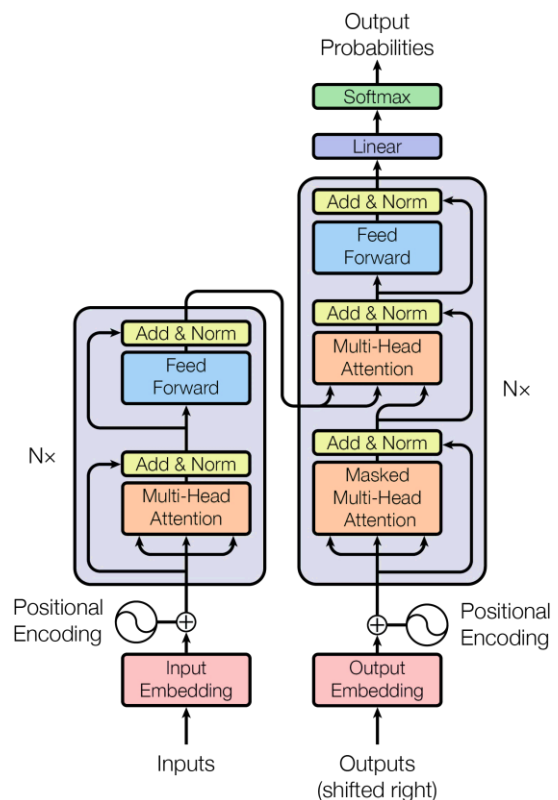


Fuente: (Cambridge University Press, 2022)

Modelo Transformers

Una estrategia que mejora el proceso de codificación, haciendo que cada palabra de una secuencia en el codificador considere el efecto de todas las demás palabras de la secuencia, es emplear mecanismos de autoatención, en estos se basa la arquitectura Transformers o Transformadores, los cuales se prescinden de usar RNNs o CNNs (Redes Neuronales Convolucionales) y, así, reducir su tiempo de entrenamiento. El modelo consiste en una pila de codificadores conectados a una pila de decodificadores: cada codificador consta de una capa de autoatención y una red neuronal que pasa su salida como entrada al siguiente codificador de la pila; cada decodificador pasa su salida como entrada al siguiente decodificador de la pila, y estos contienen una capa de autoatención, seguida de una capa de atención codificador-decodificador, seguida de una red neuronal. La entrada al primer codificador es una incrustación de palabra y una incrustación de posición.

Figura 8. Arquitectura del Modelo Transformers



Fuente: (Vaswani et al., 2017)

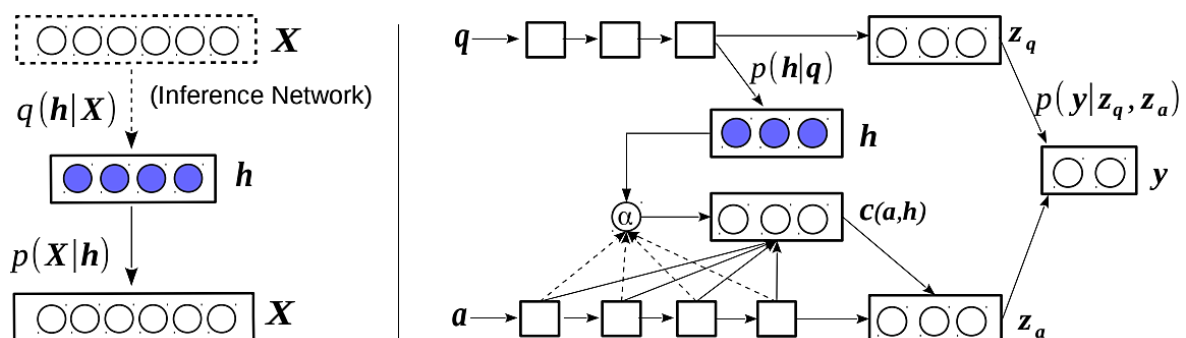
Entre cada una de las 6 capas que componen el codificador, se tiene una subcapa con un mecanismo de autoatención multicabezal y otra subcapa compuesta de un mecanismo simple de posición. Además de estas, el decodificador inserta una tercera subcapa en la misma capa, que realiza la atención multicabezal sobre la salida de la pila del codificador, empleando conexiones residuales alrededor de cada subcapa, seguidas de una normalización de capas (Vaswani, y otros, 2017).

AutoEncoder Variacional (VAE)

El AutoEncoder Variacional (VAE) es un modelo generativo probabilístico basado en una arquitectura codificador-decodificador, aprende modelos de variables latentes profundos y modelos de inferencia correspondientes. El codificador mapea la variable de entrada a un espacio latente que corresponde a los parámetros de una distribución variacional, produciendo múltiples muestras diferentes que provienen de la misma distribución. Por su parte, el decodificador mapea las muestras latentes de la distribución variacional a la distribución de probabilidad de la variable de salida con el objetivo de reconstruir la variable de entrada original a partir de la muestra latente (Kingma & Welling, 2019).

En tareas de Procesamiento del Lenguaje Natural el VAE se utiliza para modelar la distribución de probabilidad de las secuencias de texto, combinando técnicas de autoencoders y métodos probabilísticos para generar datos nuevos y reconstruir datos existentes con variabilidad. Algunas de las aplicaciones específicas son: el modelado de documento variacional neural (NVDM), caso de aprendizaje no supervisado que combina una representación de documento estocástica continua con un modelo generativo de bolsa de palabras (Bag of Words) para generar todas las palabras de un documento de forma independiente en una representación de su contenido semántico; y el modelo de selección de respuesta neural (NASM), un caso de aprendizaje supervisado que emplea una capa de representación estocástica dentro de un mecanismo de atención para extraer la semántica entre un par de pregunta y respuesta mediante dos LSTM diferentes (Miao, Yu, & Blunsom, 2016).

Figura 9. Arquitecturas NVDM, para el modelado de documentos; y NASM, para la selección de preguntas-respuestas



Fuente: (Miao, Yu, & Blunsom, 2016)

Modelos de Lenguaje

La modelización del lenguaje se puede dividir en dos enfoques principales: causal y enmascarada. Los modelos de lenguaje causales se destacan en la generación de texto y permiten aplicaciones creativas como la creación de aventuras de texto personalizadas o asistentes de codificación inteligente como Copilot o CodeParrot. Estos modelos predicen el siguiente token en una secuencia y solo tienen acceso a los tokens previos, como en el caso de GPT-2 (Hugging Face).

Por otro lado, la modelización del lenguaje enmascarada se centra en predecir tokens enmascarados en una secuencia y permite al modelo atender a los tokens en ambas direcciones. Esta metodología es ideal para tareas que requieren una comprensión contextual completa de toda una secuencia, como lo demuestran modelos como BERT y RoBERTa (Hugging Face). Durante el entrenamiento en ambos enfoques, es esencial realizar un preprocesamiento adecuado del conjunto de datos, convirtiendo el texto en la codificación comprensible para el modelo, y crear lotes de ejemplos mediante el uso de un Data Collator.

Preprocesamiento en el modelado de lenguaje

El entrenamiento de modelos de lenguaje requiere un tratamiento previo de los datos, convirtiendo el texto en valores numéricos, un proceso conocido como codificación y que se divide en dos etapas fundamentales: la tokenización, seguida de la conversión a IDs de entrada. La tokenización implica un preprocesamiento que consta de cuatro pasos realizados por un componente llamado tokenizador, los cuales incluyen la normalización, la pre-tokenización, el modelado y el postprocesamiento. Este enfoque garantiza que el texto se divida en unidades significativas llamadas tokens, que luego son convertidos en índices esenciales para el procesamiento del lenguaje natural y el modelado de lenguaje en particular, formando un tensor con ellos para alimentar al modelo. El proceso inverso que convierte los índices nuevamente en cadena de texto se conoce como decodificación.

Antes de dividir un texto en subtokens (de acuerdo a su modelo), el tokenizador realiza dos pasos: *normalización* y *pre-tokenización*. El paso de normalización involucra una limpieza general, como la remoción de espacios en blanco innecesario, transformar a minúsculas, y/o remoción de acentos. Seguidamente, es necesario separar los textos en entidades más pequeñas, como palabras; ahí es donde el paso de pre-tokenización entra en juego, que para un tokenizador basado en palabras (word-based), esta división se hace separando en espacios en blanco y puntuación, para lo cual, esas palabras serán las fronteras de los subtokens que el tokenizador aprende durante su entrenamiento (Hugging Face).

Posteriormente, la tokenización continúa con dos pasos adicionales: modelado y postprocesamiento. Durante el modelado, el tokenizador construye un vocabulario inicial

basado en los subtokens obtenidos previamente. Este vocabulario es crucial para representar el texto de entrada de manera eficiente. Luego, en el paso de postprocesamiento, se lleva a cabo la codificación de los subtokens en IDs de entrada numéricos, que son comprensibles para el modelo de IA. Este proceso finaliza la tokenización y prepara los datos para ser procesados por el modelo de lenguaje. La combinación de estos cuatro pasos garantiza una representación adecuada del texto en formato numérico, permitiendo al modelo comprender y manipular la información de manera efectiva durante el entrenamiento y la inferencia.

Figura 10. Ejemplo del proceso de tokenización



Después de haber examinado brevemente el proceso de tokenización de varios tokenizadores distintos, es momento de profundizar en los algoritmos subyacentes. Los tres principales algoritmos de tokenización por subpalabra (subword tokenization) son los siguientes: BPE, utilizado en modelos como GPT-2, RoBERTa, BART y DeBERTa; WordPiece, implementado por BERT, DistilBERT y MobileBERT; y Unigram, empleado por T5, AIBERT, XLNet y otros. A continuación, se proporciona una concisa descripción de cómo opera cada uno de estos algoritmos.

Tabla 1. Principales algoritmos de tokenización

Algoritmo	BPE	WordPiece	Unigram
Entrenamiento	Comienza a partir de un pequeño vocabulario y aprende reglas para fusionar tokens	Comienza a partir de un pequeño vocabulario y aprende reglas para fusionar tokens	Comienza de un gran vocabulario y aprende reglas para remover tokens
Etapas de Entrenamiento	Fusiona los tokens correspondientes a	Fusiona los tokens correspondientes al par con	Remueve todos los tokens en el vocabulario

	los pares más comunes	el mejor puntaje basado en la frecuencia del par, privilegiando pares donde cada token individual es menos frecuente	que minimizarán la función de pérdida (loss) calculado en el corpus completo
Aprende	Reglas de fusión y un vocabulario	Sólo un vocabulario	Un vocabulario con puntaje para cada token
Codificación	Separa una palabra en caracteres y aplica las fusiones aprendidas durante el entrenamiento	Encuentra la subpalabra más larga comenzando del inicio que está en el vocabulario, luego hace lo mismo para el resto de las palabras	Encuentra la separación en tokens más probable, usando los puntajes aprendidos durante el entrenamiento

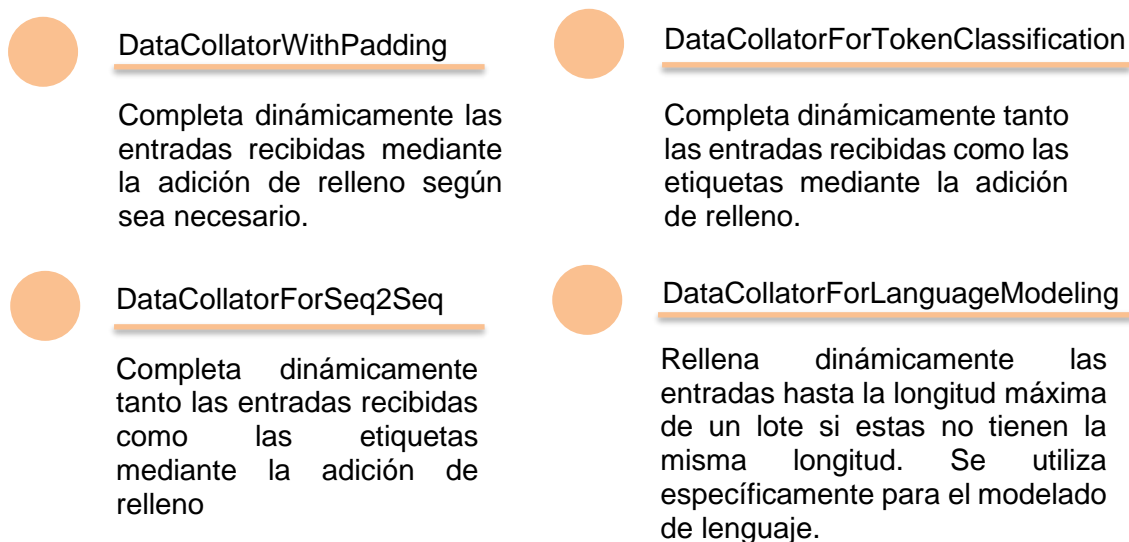
Fuente: (Hugging Face)

Data Collators

Los Data Collators, o recopiladores de datos, son objetos encargados de crear lotes a partir de una lista de elementos del conjunto de datos, los cuales son del mismo tipo que los elementos presentes en los conjuntos de entrenamiento (train_dataset) o evaluación (eval_dataset). Para generar estos lotes, los recopiladores de datos pueden aplicar diversos procesamientos, como el relleno. Algunos, como DataCollatorForLanguageModeling, también realizan aumentos aleatorios de datos, como el enmascaramiento aleatorio.

Un ejemplo básico de recopilador de datos es DefaultDataCollator, que simplemente intercala lotes de objetos tipo dict y maneja claves nombradas de manera especial: "label" para un único valor (int o float) por objeto; y "label_ids" para una lista de valores por objeto (Hugging Face). Entre los principales recopiladores de datos se encuentran:

Figura 11. Principales objetos Data Collator



5. Estado del Arte

Una práctica común en las personas es compartir la experiencia de episodios relevantes como medio de recuperación emocional o, simplemente, tener la oportunidad de expresar sus sentimientos asociados al mismo, cuya respuesta por parte del oyente influye en el estado emocional siguiente en un sentido positivo o negativo (Zech & Rimé, 2005); por tal motivo, contar con la adecuada inteligencia emocional ofreciendo una respuesta asertiva conlleva al beneficio de la persona afectada dentro de una conversación empática, definida desde un ámbito de la psicología como el resultado de la alineación e interacción conscientes entre la cognición y el afecto de la empatía (Zhou, Zheng, Wang, Zhang, & Huang, 2023) y explorar diversas formas de aplicar la empatía emerge como un paso crucial hacia sistemas de diálogo que buscan similitud con los intercambios humanos (Sabour, Zheng, & Huang, 2022).

Por lo anterior, la propuesta resultante combina tres tareas de NLP con las cuales sea apta para un público diverso y global, tratando de evitar que se discrimine su uso a usuarios específicos debido a su situación física o su lengua natal; estas son: Generación de Respuesta Empática, Análisis Multimodal de Sentimientos y Traducción Multimodal. A continuación, se presenta el estado del arte de los enfoques en cuestión.

Generación de respuesta empática

La empatía, ampliamente empleada en el asesoramiento psicológico, hace presencia como un aspecto fundamental en la comunicación entre individuos, con lo cual, al considerar la particularidad dinámica de las emociones y el conocimiento de sentido común a lo largo de una conversación, (Lin, y otros, 2022) proponen **SEEK** (Serial Encoding and Emotion-Knowledge interaction), una estrategia de codificación de grano fino sensible al flujo dinámico de la emoción en las conversaciones para predecir la característica emoción-intento de respuesta para la generación de diálogos empáticos. Extensos experimentos con EmpatheticDialogues demuestran que SEEK supera a las fuertes líneas de base tanto en las evaluaciones automáticas como en las manuales.

Análogamente y dos años antes, (Ren, y otros, 2020) proponen un modelo adversarial multi-resolución **EmpDG**, para generar respuestas más empáticas. EmpDG explota tanto las emociones a nivel de diálogo de grano grueso como las emociones a nivel de token de grano fino, lo que ayuda a captar mejor los matices de la emoción del usuario. Además, introducimos un marco de aprendizaje interactivo adversarial que explota la retroalimentación del usuario para identificar si las respuestas generadas evocan la percepción de la emoción en los diálogos.

Siguiendo la idea de imitación de las emociones del usuario en respuestas empáticas, así como la variación de éstas dentro de una misma conversación, **(Peng, y otros, 2020)** realizan dos aportaciones dentro de su modelo **MIME** (MIMicking Emotions). En primer lugar, introduce un nuevo enfoque para la generación de empatía que codifica el contexto y las emociones, y utiliza el muestreo estocástico de emociones y la imitación de emociones para generar respuestas que sean apropiadas y empáticas para enunciados positivos o negativos. En segundo lugar, mediante amplios experimentos de ablación de características, arrojamamos luz sobre el papel que desempeñan la imitación de emociones y la agrupación de emociones en la tarea de generar respuestas empáticas.

En lo que respecta al conocimiento de sentido común mencionado, **(Sabour, Zheng, & Huang, 2022)** aprovechan este enfoque bajo una propuesta que obtiene más información sobre la situación del usuario y la utiliza para mejorar la comprensión cognitiva, por consiguiente, la expresión de empatía en las respuestas generadas por su modelo **CEM** (Commonsense-aware Empathetic Chatting Machine).

Los cuatro modelos anteriores y que comparten un enfoque en específico, fueron evaluados por su respectivo autor en el conjunto de referencia ampliamente utilizado para la generación de respuestas empáticas **EMPATHETICDIALOGUES**, los resultados demuestran una mejora en diferentes métricas de calidad de generación entre cada modelo, tales como la precisión (Accuracy – Acc), la perplejidad como medida en el gado de calidad del modelo de generación y las distinciones 1 y 2 para medir la proporción de unigramas y bigramas distintos en los resultados generados:

Tabla 2. Resultados de la evaluación automática

Modelo	Acc	Perplejidad	Distinc-1	Distinc-2
MIME	0.2938	37.08	0.31	1.03
EmpDG	0.3003	37.77	0.59	2.48
CEM	0.3644	37.03	0.66	2.99
SEEK	0.4185	37.09	0.73	3.23

La generación de respuestas empáticas por parte de un sistema basado en agentes de diálogo es resultado del reto de reconocer los sentimientos del interlocutor y responder en consecuencia, requiriendo identificar la emoción del usuario y expresar adecuadamente su empatía. **(Liu, Maier, Minker, & Ultes, 2021)** proponen **RoBERTa-GPT2** para la generación de diálogos empáticos, un modelo codificador-decodificador en el cual el autocodificador preentrenado RoBERTa actúa como codificador, además, se incluye un Extractor de Conocimientos y Conceptos Emocionales de Sentido Común (CKECE) a partir del contexto de diálogo con el fin de permitir la capacidad de empatía y sentido común del decodificador GPT-2 de OpenAI.

Los resultados de la evaluación de RoBERTa-GPT2, tomando como precisión la exactitud de la emoción predicha, empleando el método CKECE y sin el mismo, son mostrados junto con la evaluación del modelo preentrenado, que fueron entrenados empleando el conjunto de datos EMPATHETICDIALOGUES.

Tabla 3. Resultados de evaluación del modelo RoBERTa-GPT2

Modelo	Precisión de la Emoción	Perplejidad	Distinc-1	Distinc-2
RoBERTa w/o GPT-2	0.3439	-	-	-
RoBERTa-GPT2 w/o CKECE	0.5262	14.97	1.62	10.47
RoBERTa-GPT2	0.5151	13.57	2.04	11.68

Para comprender con mayor detalle en qué consiste el modelo **RoBERTa** (Robustly Optimized BERT Pretraining Approach), (Liu et al, 2019) exponen una mejora en el rendimiento del preentrenamiento BERT, alcanzando una puntuación de 88,5 en la tabla de clasificación pública GLUE, con un nuevo estado del arte en 4 de las 9 tareas GLUE: MNLI, QNLI, RTE y STS-B e igualando los resultados del estado del arte en SQuAD y RACE. Entre las modificaciones realizadas por los autores, y enfocadas en el proceso de entrenamiento, se encuentran:

- Aumento en el tiempo de entrenamiento, tamaño de lotes (batch size) y longitud del conjunto de datos.
- Eliminar el objetivo de predicción de la siguiente frase.
- Mayor longitud en las secuencias.
- Cambio dinámico sobre el patrón de enmascaramiento en los datos.

Una comparación mejor referenciada de los resultados previamente ilustrados pierde el sentido si se ignora el fundamento y rendimiento del modelo base **BERT** (Bidirectional Encoder Representations from Transformers), diseñado por (**Devlin, Chang, Lee, & Toutanova, 2019**) como un modelo de representación del lenguaje para preentrenar representaciones bidireccionales profundas a partir de texto sin etiquetar. Obtiene una puntuación GLUE hasta el 80,5%, MultiNLI hasta el 86,7%, prueba F1 de respuesta a preguntas SQuAD v1.1 hasta el 93,2 y prueba F1 de SQuAD v2.0 hasta el 83,1.

Tras haber explorado los modelos BERT y RoBERTa, dos pilares fundamentales en el campo del procesamiento del lenguaje natural, surge la necesidad de avanzar hacia la aplicación específica en el ámbito de la salud mental. En este contexto, (Ji et al, 2021) presentan los modelos **MentalBERT** y **MentalRoBERTa**, dos modelos de lenguaje enmascarado preentrenados dirigidos a la detección temprana de trastornos mentales y la ideación suicida a partir del análisis de contenido en plataformas sociales. Estos modelos representan una evolución significativa al ofrecer representaciones lingüísticas preentrenadas adaptadas al dominio de la salud mental, lo que proporciona una herramienta

poderosa para la identificación y el tratamiento de problemas de salud mental en un contexto social. El entrenamiento de estos modelos se realizó utilizando cuatro GPU Nvidia Tesla V100, con un tamaño de lote de 16 por GPU y una evaluación cada 1.000 pasos durante 624.000 iteraciones [1]. Este enfoque, respaldado por recursos informáticos adecuados, demuestra su eficacia en la detección temprana de trastornos mentales y la ideación suicida en plataformas sociales.

Tomando en consideración la capacidad de la propuesta innovadora de RoBERTa-GPT2 para generar respuestas empáticas en diálogos, es menester considerar el modelo **DialoGPT** (Dialogue Generative Pre-trained Transformer) en el contexto de la generación de diálogos. DialoGPT, desarrollado por (Zhang et al, 2020) expande las capacidades de GPT-2 y se entrena con datos de Reddit para abordar los desafíos en la generación de respuestas conversacionales. Este modelo se centra en generar texto natural y relevante para la solicitud. Para su entrenamiento, se empleó el programador de tasa de aprendizaje Noam con 16000 pasos de calentamiento, y la tasa de aprendizaje se ajusta según la pérdida de validación. El entrenamiento se detiene cuando la pérdida de validación no mejora, con hasta 5 épocas para modelos pequeños y medianos, y un máximo de 3 épocas para modelos grandes.

Volviendo con los modelos generacionales de respuestas empáticas, una solución hecha por parte de (Kim, Kim, & Kim, 2021) en su modelo **GEE** (Generative Emotion Estimator) aborda la conversación desde la identificación de la palabra causante de la emoción del usuario y actuar en consecuencia reflejando dicha palabra en la generación de respuestas. La primera cuestión es lograda utilizando lo que denominan un “estimador generativo” sobre enunciados sin etiqueta a nivel de palabra inspirados en la cognición social; la generación de respuesta centrada en la palabra inferida es realizada con un método basado en la pragmática, como afirman los autores.

La comparación del rendimiento en el reconocimiento de palabras causantes de emociones entre GEE, aleatorio, RAKE, EmpDG y BERT en el conjunto de evaluación EMOCAUSE se muestra en la tabla.

Tabla 4. Comparación del rendimiento en el reconocimiento de palabras causantes de emociones

Modelo	Top-1 Recall	Top-3 Recall	Top-5 Recall
Humano	41.3	81.1	95.0
Aleatorio	10.7	30.6	48.5
EmpDG	13.4	36.2	49.3
RAKE	12.7	35.8	55.0
BERT-Attention	13.8	40.6	61.2
GEE	17.3	48.1	68.4

Análisis Multimodal de Sentimiento

Son diversos los enfoques propuestos que tratan de simular el carácter multimodal complejo de la comunicación humana en su máxima expresión, ocurriendo una interrelación entre la modalidad lingüística (uso de palabras), modalidad visual (gestos corporales y faciales) y modalidad acústica (cambios de tonos en el habla); comprender los sentimientos y diversidad de emociones que intervienen en una conversación demanda el análisis de estas modalidades y su forma de interaccionar, un desafío importante para la Inteligencia Artificial (IA) en el desarrollo del entendimiento del lenguaje (Zadeh A. , y otros, 2018). Como aporte al cubrimiento de esta tarea, algunos autores han realizado modificaciones a los modelos tratados BERT y RoBERTa o se han inspirado en el mismo:

(Yu, y otros, 2023) Proponen **SPECTRA** (Speech-text dialog Pre-training for spoken dialog understanding with ExpliCiT cRoss-Modal Alignment), considerado por ellos como el primer modelo de preentrenamiento de diálogo voz-texto para la comprensión de diálogos hablados. La arquitectura se compone de un codificador de texto (RoBERTa) y un codificador del habla (inspirado en WavLM) a los cuales, durante el preentrenamiento, se introducen respectivamente las entradas emparejadas de texto y voz previamente convertidas en incrustaciones unimodales, que luego son concatenadas en una misma entrada a un módulo de fusión de modalidades del modelo, obteniendo representaciones fusionadas para el preentrenamiento habla-texto. Los autores preentrenaron SPECTRA sobre el conjunto de datos de diálogo voz-texto en escenas del mundo real Spotify100k, ilustrando los resultados del estado del arte en la siguiente tabla para distintos conjuntos de datos y tareas en los que ha sido evaluado el modelo.

Tabla 5. Comparación entre las métricas clave de SPECTRA y el método SOTA anterior en cinco conjuntos de datos.

Tarea	Dataset	Métrica	SOTA previo	SPECTRA
Análisis Multimodal de Sentimiento	MOSI	Acc ₂	84.40 (MIB)	87.50
	MOSEI	Acc ₂	86.20 (BBFN)	87.34
Reconocimiento de Emociones en Conversación	IEMOCAP	Acc	66.52 (M2FNET)	67.94
Comprensión de la Lengua Hablada	MIntRec	Acc ₂₀	72.16 (MAG-BERT)	73.48
Seguimiento del Estado del Diálogo	SpokenWoz	JGA	20.90 (WavLM)	21.96

A diferencia del anterior, (Yang, Xu, & Gao, 2020) presentan **CM-BERT** (Cross-Modal BERT), una adaptación multimodal de BERT, que utiliza la información de la modalidad de audio para ayudar a la modalidad de texto a ajustar dinámicamente el peso de las palabras y afinar el modelo BERT preentrenado, mediante la atención multimodal enmascarada. El

método es evaluado en los conjuntos de datos públicos de análisis multimodal de sentimientos CMU-MOSI y CMU-MOSEI, empleando las métricas Acc_7 (accuracy o precisión de 7 clases), Acc_2 (accuracy de 2 clases), score $F1$ (puntuación sobre clasificación binaria, al igual que Acc_2), MAE (Mean Absolute Error o error absoluto medio) y la correlación $Corr$ de las predicciones del modelo con las etiquetas reales.

Los resultados son mostrados tras una evaluación en modalidad texto, audio y video, a excepción de BERT que se realizó en modalidad texto.

Tabla 6. Resultados experimentales de CM-BERT en el conjunto de datos CMU-MOSI

Modelo	Acc_7	Acc_2	F1	MAE	$Corr$
EF-LSTM	33.7	75.3	75.2	1.023	0.608
LMF	32.8	76.4	75.7	0.912	0.668
MFN	34.1	77.4	77.3	0.965	0.632
MARN	34.7	77.1	77.0	0.968	0.625
RMFN	38.3	78.4	78.0	0.922	0.681
MFM	36.2	78.1	78.1	0.951	0.662
MCTN	35.6	79.3	79.1	0.909	0.676
MuIT	40.0	83.0	82.8	0.871	0.698
BERT (Text)	41.5	83.2	83.2	0.784	0.7748
CM-BERT	44.9	84.5	84.5	0.729	0.791

La implementación de modelos que pretenden ser entrenados para el reconocimiento multimodal de emociones basada en datos provenientes del habla con una arquitectura de aprendizaje auto supervisado (Self-Supervised Learning – SSL) preentrenada tipo BERT es explorada por (Siriwardhana, Reis, Weerasekera, & Nanayakkara, 2020), partiendo de un modelo denominado por los autores como **Speech-BERT** y el modelo RoBERTa. Se han evaluado dos posibles mecanismos de fusión para combinar los dos modelos SSL, cuyo rendimiento sobre los conjuntos de datos IEMOCAP, CMU-MOSEI y CMU-MOSI fue comparado en un modelo final propuesto.

Traducción Multimodal

Los anteriores modelos se enfocan en la predicción de emociones bajo el análisis multimodal de sentimiento con la comprensión del contexto de un diálogo; sin embargo, la interacción completa multimodal entre humano y agente conversacional afronta la necesidad de tomar en consideración la posibilidad de una comunicación multimodal junto a la correspondiente traducción intermodalidad, tanto en la tarea de traducción voz-texto (Speech-to-Text, S2T) como la traducción texto-voz (Text-to-Speech, T2S).

(**Ma, Pino, & Koehn, 2020**) estudiaron la manera de adaptar los métodos de traducción simultánea de texto (Simultaneous text translation – SimulMT) a traducción simultánea del habla (Simultaneous speech translation – SimulST) de extremo a extremo haciendo uso de la arquitectura S-Transformer, que logra un rendimiento competitivo en el conjunto de datos MuST-C. En el codificador de **SimulMT to SimulST**, se aplica una atención bidimensional después de las capas CNN y se introduce una penalización de distancia para sesgar la atención hacia las dependencias de corto alcance. Adicionalmente, se investigó dos tipos de mecanismos de traducción simultánea: la atención multicabezal monotónica, que es un caso de política flexible; y el modelo prefijo a prefijo, un caso de política fija.

La traducción S2S, S2T, T2S y T2T admitidas en un único modelo, así como el reconocimiento automático del habla (ASR) para hasta 100 idiomas, es la promesa de (**Meta AI, INRIA, UC Berkeley, 2023**) con **SeamlessM4T** (Massively Multilingual & Multimodal Machine Translation), que siguió un proceso de entrenamiento con un millón de horas de datos de audio de habla abierta para aprender representaciones del habla auto supervisadas con w2v-BERT 2.0 (word2vec BERT):

1. Creación de un modelo multimodal de traducción automática.
2. Instauración de SeamlessAlign, un corpus multimodal de traducciones del habla alineadas automáticamente.
3. Innovación en la traducción en voz y texto desde y hacia el inglés por un sistema multilinguaje.

Los autores exponen los resultados de su propuesta sobre las pruebas Fleurs y Flores para las tareas X2T (ASR, traducción S2T y traducción T2T) tanto hacia el inglés (X-eng) como desde este idioma (eng-X), en comparación con los modelos de traducción directa del SOTA.

Tabla 7. Comparación de resultados de SeamlessMT4 en multitarea X2T con respecto al SOTA

Modelo	Tamaño	Traducción S2T (\uparrow BLEU)			
		Fleurs X-eng (n = 81)	Fleurs Eng-X (n = 88)	CoVoST 2 X-eng (n = 21)	CoVoST 2 Eng-X (n = 15)
XLS-R-2B-S2T	2.6B		x	22.1	27.8
Whisper-Large-v2	1.5B	17.9	x	29.1	x
AudioPaLM-2-8B-AST	8.0B	19.7	x	37.8	x
SeamlessM4T-Medium	1.2B	20.9	19.2	29.8	26.6
SeamlessM4T-Large	2.3B	24.0	21.5	34.1	30.6

Modelo	Tamaño	ASR (\downarrow WER)		Traducción S2T (\uparrow chrF+ +)	
		Fleurs (n = 77)	Fleurs-54 (n = 54)	Flores X-eng (n = 95)	Flores Eng-X (n = 95)
NLLB-3.3B	3.3B	x	x	60.7	49.6
Whisper-Large-v2	1.5B	41.7	43.7	x	x
MMS-L61-noLM-LSAH	1.0B	x	31.0	x	x
MMS-L1107-CCLM-LSAH	1.0B	x	18.7	x	x
SeamlessM4T-Medium	1.2B	21.9	22.0	55.4	48.4
SeamlessM4T-Large	2.3B	23.1	23.7	60.8	50.9

El alcance de una calidad a nivel humano en la traducción T2S implica definir en qué se fundamenta y cómo juzgar esa calidad, según **(Tan, y otros, 2022)**, para ello, establecen tal definición basándose en la importancia estadística de la medida subjetiva e introducen directrices apropiadas para juzgarla, como punto de partida para en el desarrollo de un sistema llamado **NaturalSpeech** que, sostienen, alcanza la calidad a nivel humano en el conjunto de datos de referencia LJSpeech con una puntuación de opinión media comparativa (CMOS) de -0.01 con respecto a las grabaciones humanas y un nivel **p** de **0.05** sobre la prueba de rango con signo de Wilcoxon. El modelo utiliza un autocodificador variacional (VAE) para la generación de texto, con varios módulos clave para mejorar la capacidad del previo a partir del texto y reducir la complejidad del posterior a partir del habla, incluyendo:

- Preentrenamiento de fonemas,
- Modelado de duración diferenciable,
- Modelado bidireccional previo/posterior
- Mecanismo de memoria en VAE.

Un modelo generativo T2S denominado **PortaSpeech** es propuesto por **(Ren, Liu, & Zhao, 2022)**. La promesa descrita por los autores es la alta calidad en la generación de un habla diversa con detalles naturales y una prosodia rica con las siguientes características en la arquitectura y entrenamiento:

- Arquitectura principal empleando una VAE ligera con una priorización mejorada seguida de una postred basada en flujos con entradas condicionales fuertes.
- Mecanismo de compartición de parámetros agrupados en las capas de acoplamiento afín de la postred.
- Implementación de un codificador lingüístico con alineación mixta que combina la alineación dura a nivel de palabra y la alineación suave a nivel de fonema.

Los resultados experimentales y comparación del rendimiento sonoro (MOS-Q y MOS-P), la latencia de inferencia, peak memory (Peak Mem.) y el factor de tiempo real RTF (segundos necesarios para que el sistema sintetice un segundo de audio) se muestran en la tabla.

Tabla 8. Comparación del rendimiento del modelo PortaSpeech

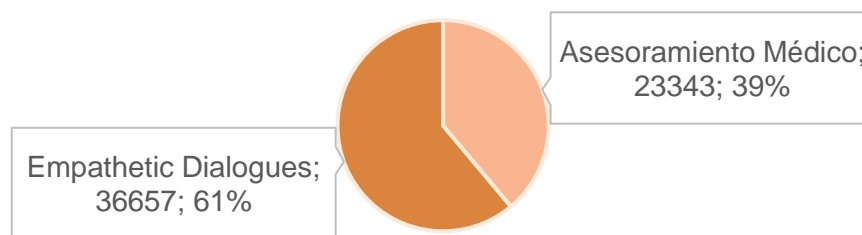
Modelo	MOS-P	MOS-Q	RTF	Peak Mem.
GT	4.52 ± 0.07	4.41 ± 0.06	/	/
GT (voc.)	4.48 ± 0.08	4.15 ± 0.07	/	/
Tacotron 2	3.85 ± 0.07	3.80 ± 0.08	0.115	61.78MB
TransformerTTS	3.87 ± 0.06	3.82 ± 0.07	0.955	118.66MB
FastSpeech	3.63 ± 0.08	3.72 ± 0.08	0.0198	115.20MB
FastSpeech 2	3.72 ± 0.07	3.83 ± 0.06	0.0200	124.80MB
Glow-TTS	3.61 ± 0.07	3.88 ± 0.08	0.0196	116.40MB
BVAE-TTS	3.80 ± 0.06	3.72 ± 0.06	0.0169	90.10MB
PortaSpeech (normal)	3.89 ± 0.06	3.92 ± 0.06	0.0216	83.60MB
PortaSpeech (pequeño)	3.82 ± 0.06	3.86 ± 0.06	0.0208	39.30MB

6. Material

En el desarrollo del sistema de Generación de Contenido Autoayuda y Apoyo Psicológico Personalizado propuesto, es fundamental contar con datasets robustos y herramientas de traducción multimodal apropiadas. Para este fin, se utilizarán varios recursos importantes que se describen a continuación, como son: un conjunto de datos en inglés nombrado EmpathetiCounseling, construido a partir de algunos ejemplos de EmpatheticDialogues y otros datasets de asesoramiento médico con ejemplos compuestos por pares de consulta-respuesta en torno a un contexto de la salud mental; adicionalmente, los conjuntos para AMS CMU-MOSI y CMU-MOSEI y la API SeamlessM4T.

El número de ejemplos de EmpatheticDialogues y los datasets de asesoramiento, dentro del conjunto de datos construido, se encuentran en la siguiente relación:

Figura 9. Distribución del conjunto EmpathetiCounseling



Por su parte, los datasets enfocados en el asesoramiento médico son:

- Mental health counseling conversation
- Psych8k
- Counsel Chat
- Emotional-Support-Conversation (ESConv)
- Prompt-Aware margIn Ranking (PAIR)

EmpatheticDialogues Dataset

Se utilizará como principal enfoque para el entrenamiento de los modelos de respuesta empática, esto debido a los resultados experimentales que demuestran la superioridad de los modelos entrenados con este dataset en comparación con modelos entrenados con conversaciones de internet a gran escala (Rashkin, Smith, Li, & Boureau, 2019), como indican sus autores. Con EmpatheticDialogues, se busca mejorar la capacidad del sistema para comprender y responder a las emociones del usuario de manera más efectiva. Algunos ejemplos que destacan los autores, extraído del paper original, se muestran en la tabla 8.

Tabla 9. Dos ejemplos del conjunto de entrenamiento EMPATHETICDIALOGUES

<p>Label: Afraid Situation: Speaker felt this when... “I’ve been hearing noises around the house at night” Conversation: Speaker: I’ve been hearing some strange noises around the house at night. Listener: oh no! That’s scary! What do you think it is? Speaker: I don’t know, that’s what’s making me anxious. Listener: I’m sorry to hear that. I wish I could help you figure it out</p>	<p>Label: Proud Situation: Speaker felt this when... “I finally got that promotion at work! I have tried so hard for so long to get it!” Conversation: Speaker: I finally got promoted today at work! Listener: Congrats! That’s great! Speaker: Thank you! I’ve been trying to get it for a while now! Listener: That is quite an accomplishment and you should be proud!</p>
--	--

Nota. El primer trabajador (Speaker) recibe una etiqueta de emoción y escribe su propia descripción de una situación en la que se ha sentido así, luego cuenta su historia en una conversación con un segundo trabajador (Listener). Fuente: (Rashkin, Smith, Li, & Boureau, 2019)

El conjunto de datos EMPATHETICDIALOGUES (ED) posee originalmente las siguientes características o columnas (a la izquierda el tipo de dato; a la derecha un ejemplo):

conv_id: string.	conv_id: 'hit:1_conv:2',
utterance_idx: int32.	utterance_idx: 5,
context: string.	context: 'afraid',
prompt: string.	prompt: ' i used to scare for darkness',
speaker_idx: int32.	speaker_idx: 2,
utterance: string.	utterance: ' i virtually thought so.. and i used to get sweatings',
selfeval: string.	selfeval: '4 3 4_3 5 5',
tags: string.	tags: "

De las anteriores, la columna ‘context’ indica la emoción expresada por el Speaker, ‘prompt’ es su enunciado y ‘utterance’ es la respuesta del Listener. Las columnas restantes son omitidas a causa de su irrelevancia para el entrenamiento.

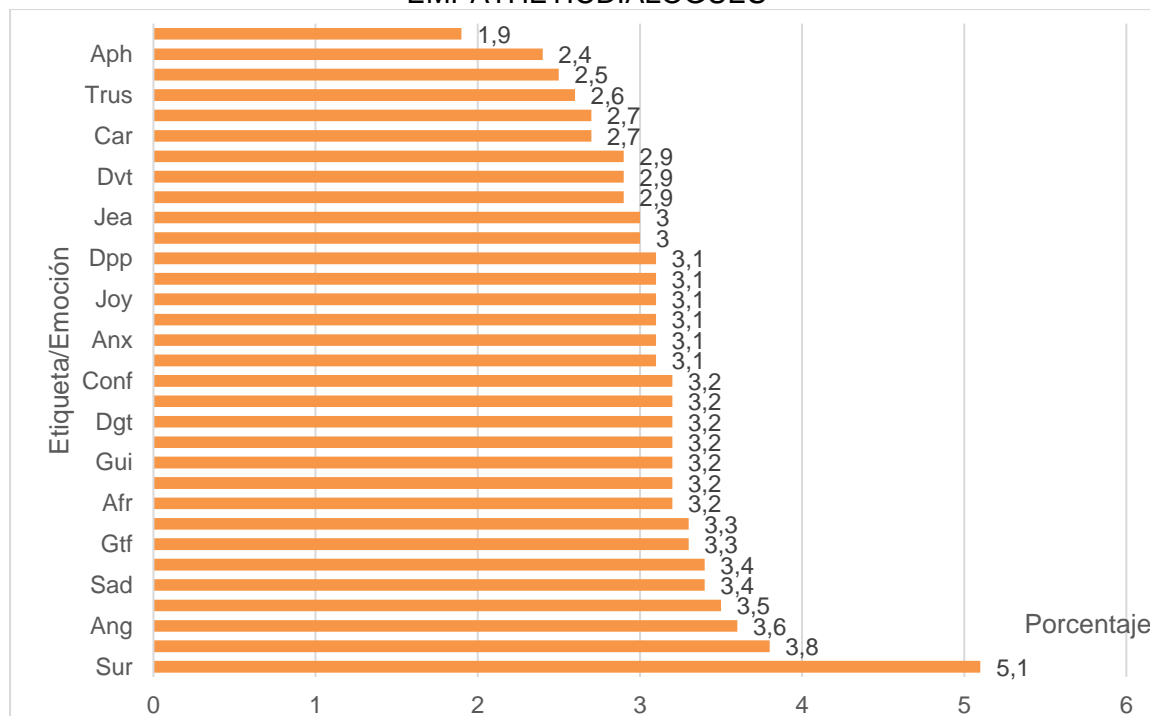
A continuación, se presenta la **etiqueta** de las emociones consideradas dentro del conjunto de entrenamiento EMPATHETICDIALOGUES y las 3 palabras de contenido más utilizadas por el *hablante/oyente*; además de la distribución porcentual de las etiquetas.

Tabla 10. Las 3 palabras de contenido más utilizadas por el hablante oyente en cada etiqueta de emociones

Surprised (Sur) got, shocked, really that's, good, nice	Excited (Exc) going, wait, i'm that's, fun, like	Angry (Ang) mad, someone, got oh, would, that's	Proud (Pro) got, happy, really that's, great, good
Sad really, away, get sorry, oh, hear	Annoyed (Ann) get, work, really that's, oh, get	Grateful (Gtf) really, thankful, i'm that's, good, nice	Lonely (Lnl) alone, friends, i'm i'm, sorry, that's
Afraid (Afr) scared, i'm, night oh, scary, that's	Terrified (Trf) scared, night, i'm oh, that's, would	Guilty (Gui) bad, feel, felt oh, that's, feel	Impressed (Imp) really, good, got that's, good, like
Disgusted (Dgt) gross, really, saw oh, that's, would	Hopeful (Hop) i'm, get, really hope, good, that's	Confident (Conf) going, i'm, really good, that's, great	Furious (Fur) mad, car, someone oh, that's, get
Anxious (Anx) i'm, nervous, going oh, good, hope	Anticipating (Ant) wait, i'm, going sounds, good, hope	Joyful (Joy) happy, got, i'm that's, good, great	Nostalgic (Nos) old, back, really good, like, time
Disappointed (Dpp) get, really, work oh, that's, sorry	Prepared (Pre) ready, i'm, going good, that's, like	Jealous (Jea) friend, got, get get, that's, oh	Content (Cont) i'm, life, happy good, that's, great
Devastated (Dvt) got, really, sad sorry, oh, hear	Embarrassed (Emb) day, work, got oh, that's, i'm	Caring (Car) care, really, taking that's, good, nice	Sentimental (Sent) old, really, time that's, oh, like
Trusting (Tru) friend, trust, know good, that's, like	Ashamed (Ash) feel, bad, felt oh, that's, i'm	Apprehensive (Aph) i'm, nervous, really oh, good, well	Faithful (Fth) i'm, would, years good, that's, like

Fuente: (Rashkin, Smith, Li, & Boureau, 2019)

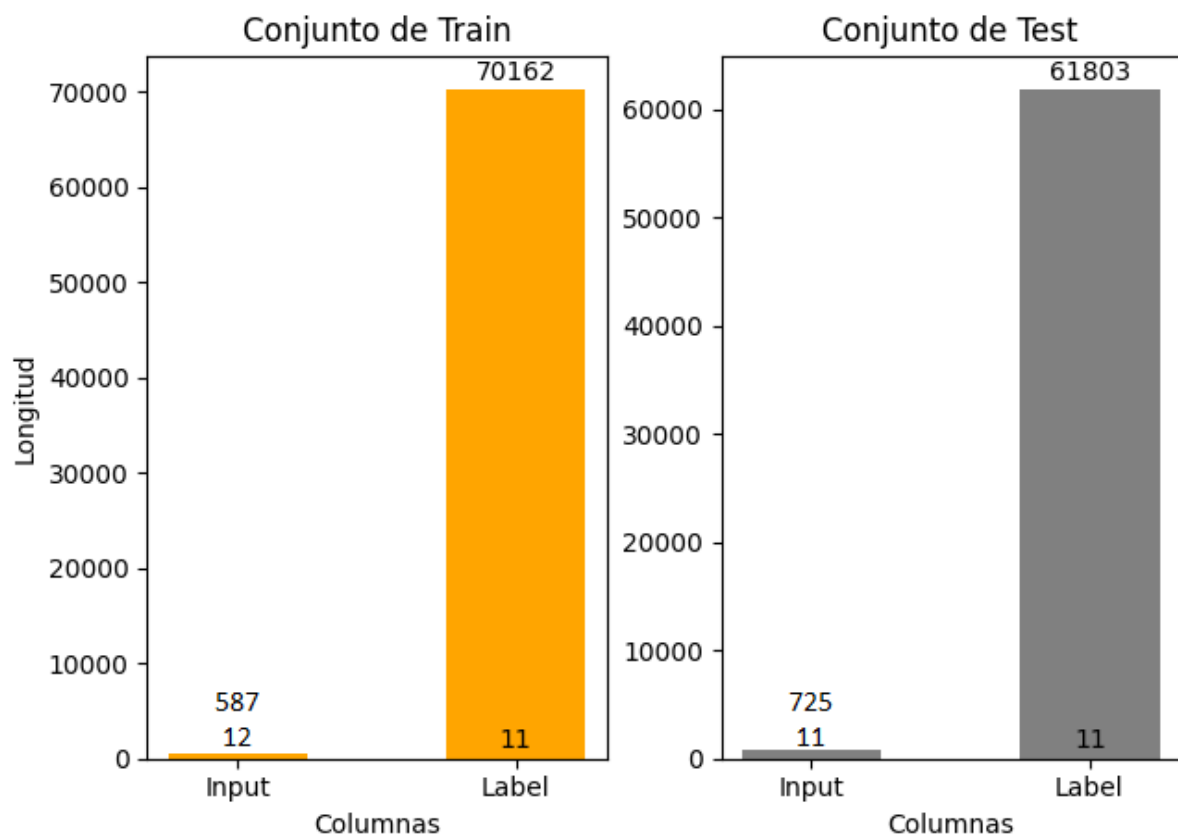
Figura 12. Distribución de etiquetas de conversación dentro del conjunto de entrenamiento EMPATHETICDIALOGUES



A pesar de que los autores mencionan que el conjunto de datos consta exactamente de 24,850 conversaciones sobre descripciones de situaciones (recopiladas de 810 participantes diferentes), dividido en aproximadamente un 80% de entrenamiento, un 10% de validación y un 10% de prueba; el dataset trabajado cuenta con 99646 ejemplos, de los cuales se tomaron 29326 equivalentes a un 80% para entrenamiento y 7331 para un 20% como datos de validación, sumando 36657 ejemplos de tuplas entrada-etiqueta. La entrada hace referencia al enunciado dado por el Speaker y es la entrada que recibe el modelo, cuya columna se renombra como 'input'; por otra parte, la etiqueta se refiere a la respuesta del Listener, que es la cadena de texto que trata de ser predicha.

Tal como se procede con el resto de conjuntos que componen al dataset construido EmpathetiCounseling, únicamente se consideran ambas columnas mencionadas y se renombran como 'input' y 'label', eliminando los ejemplos con respuesta None o de longitud ≤ 10 (vacías o compuestas por 1 o 2 palabras). De esta manera, la máxima y mínima longitud de cadenas de texto presentes en el conjunto de datos EmpatheticDialogues preprocesado por cada columna, se muestra en la gráfica.

Figura 13. Longitud mínima y máxima en Empathetic Dialogues Dataset



Mental health counseling conversation

Este conjunto de datos contiene una variedad de temas de salud mental, con respuestas proporcionadas por psicólogos cualificados, recopilando preguntas y respuestas de dos plataformas en línea de asesoramiento y terapia. Está diseñado para mejorar los modelos lingüísticos, especialmente en la generación de textos para brindar asesoramiento psicológico, teniendo como propósito principal apoyar la generación de consejos o sugerencias en respuesta a preguntas relacionadas con la salud mental. Como indican su autor, (Amod), los datos fueron limpiados meticulosamente para incluir solo conversaciones relevantes; de esta forma, cada instancia es abarcada por sus columnas 'Context' y 'Response':

Context: Cadena que contiene la pregunta formulada por un usuario.

Response: Cadena que contiene la respuesta correspondiente dada por un psicólogo.

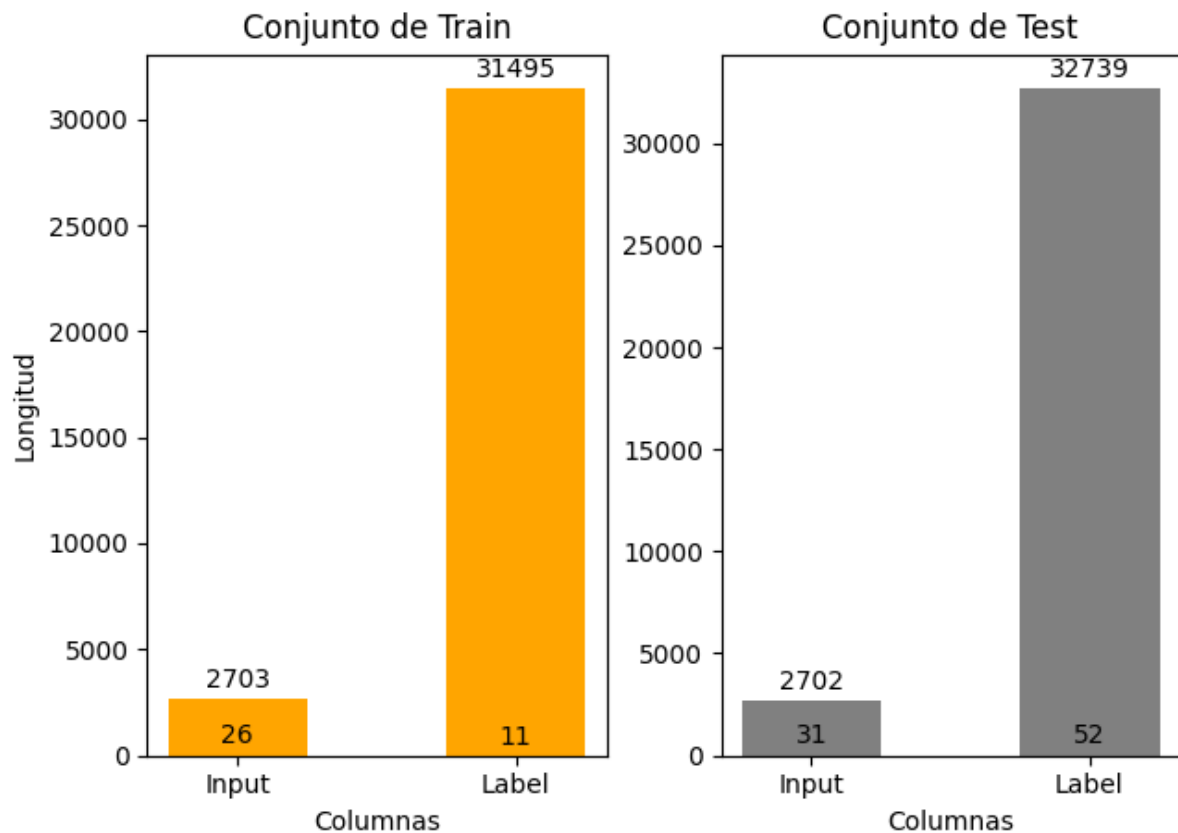
Tras realizar el preprocesamiento, se obtuvo un total de 3505 instancias, divididas en 2804 para entrenamiento (80%) y 701 de validación (20%). Una de las instancias que componen el dataset es la siguiente:

Tabla 11. Ejemplo del conjunto Mental health counseling conversation

Context (input)	I'm going through some things with my feelings and myself. I barely sleep and I do nothing but think about how I'm worthless and how I shouldn't be here.\n I've never tried or contemplated suicide. I've always wanted to fix my issues, but I never get around to it.\n How can I change my feeling of being worthless to everyone?
Response (label)	If everyone thinks you're worthless, then maybe you need to find new people to hang out with.Seriously, the social context in which a person lives is a big influence in self-esteem.Otherwise, you can go round and round trying to understand why you're not worthless, then go back to the same crowd and be knocked down again.There are many inspirational messages you can find in social media. \xa0Maybe read some of the ones which state that no person is worthless, and that everyone has a good purpose to their life.Also, since our culture is so saturated with the belief that if someone doesn't feel good about themselves that this is somehow terrible.Bad feelings are part of living. \xa0They are the motivation to remove ourselves from situations and relationships which do us more harm than good.Bad feelings do feel terrible. \xa0 Your feeling of worthlessness may be good in the sense of motivating you to find out that you are much better than your feelings today.

Parece ser una respuesta bastante larga, sin embargo, existen instancias con respuestas desde una longitud igual a 24 caracteres, como muestra la gráfica de longitudes máxima y mínima.

Figura 14. Longitud máxima y mínima de Mental Health Counseling Conversation Dataset



Psych8k

El conjunto de datos de entrenamiento, Psych8k, ha sido presentado por (Liu et al, 2023) acompañando su propuesta de modelo de lenguaje amplio (LLM) ChatCounselor para apoyo en salud mental. Se basa en 260 entrevistas en profundidad, cada una de una hora de duración, recopiladas para evaluar la calidad de las respuestas de asesoramiento, tomando el Banco de asesoramiento como referencia evaluativa.

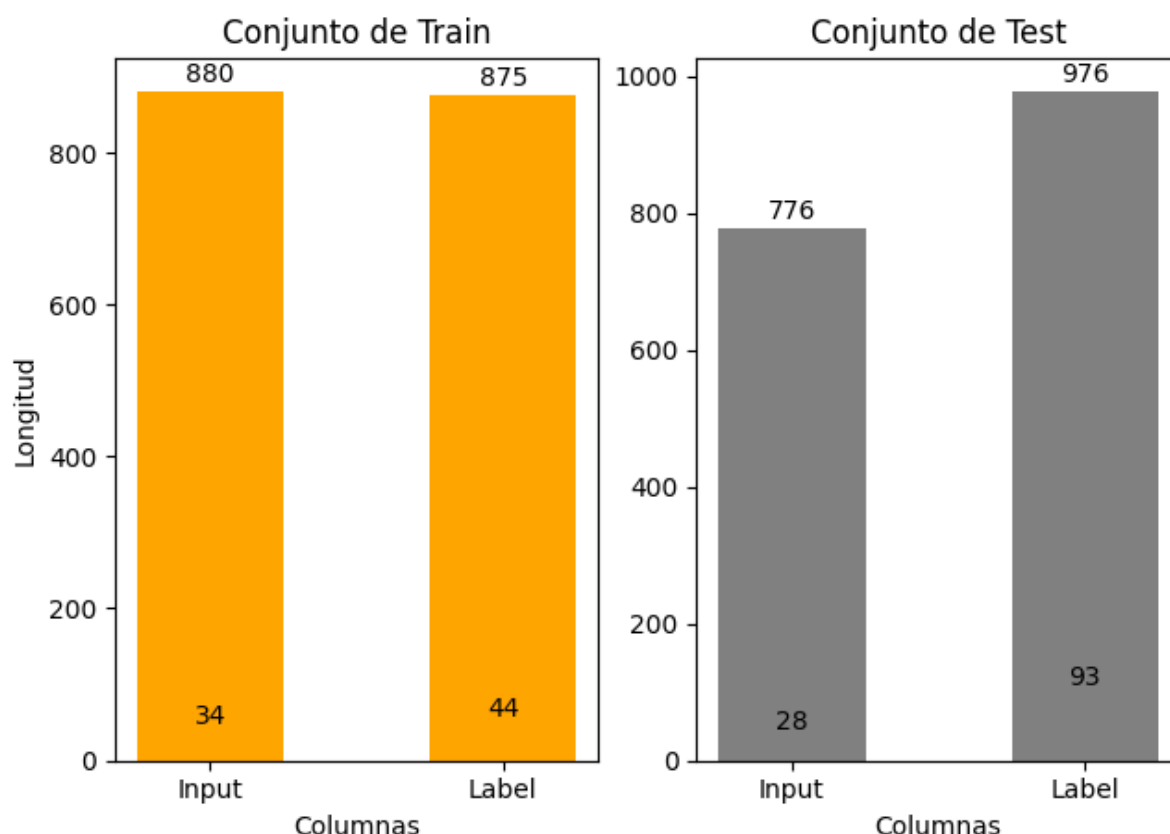
Las 260 conversaciones reales en grabaciones de asesoramiento, en inglés, se transcribieron y utilizaron para construir los conjuntos de datos de entrenamiento y prueba. Las conversaciones abarcan una amplia gama de temas, incluyendo emociones, familia, relaciones, desarrollo profesional, estrés académico, entre otros.

Luego de realizar el preprocesamiento, se pudo aprovechar las 8187 filas que constituyen el conjunto original, distribuidas en 6549 para entrenamiento y 1638 para validación, abarcando las columnas 'input' y 'output' que, tras su renombramiento, son respectivamente 'input' y 'label'. A continuación se muestra un ejemplo del conjunto de datos y la longitud máxima y mínima entre los mismos.

Tabla 12. Ejemplo del conjunto de entrenamiento Psych8k

input	Lately, I've been feeling a bit off. I've noticed that I've been having trouble concentrating at work,...
output (label)	Thank you for sharing your feelings and experiences with me. It's important to recognize when we are feeling off and seek help to better understand what's happening...

Figura 15. Longitud máxima y mínima de Psych8k Dataset

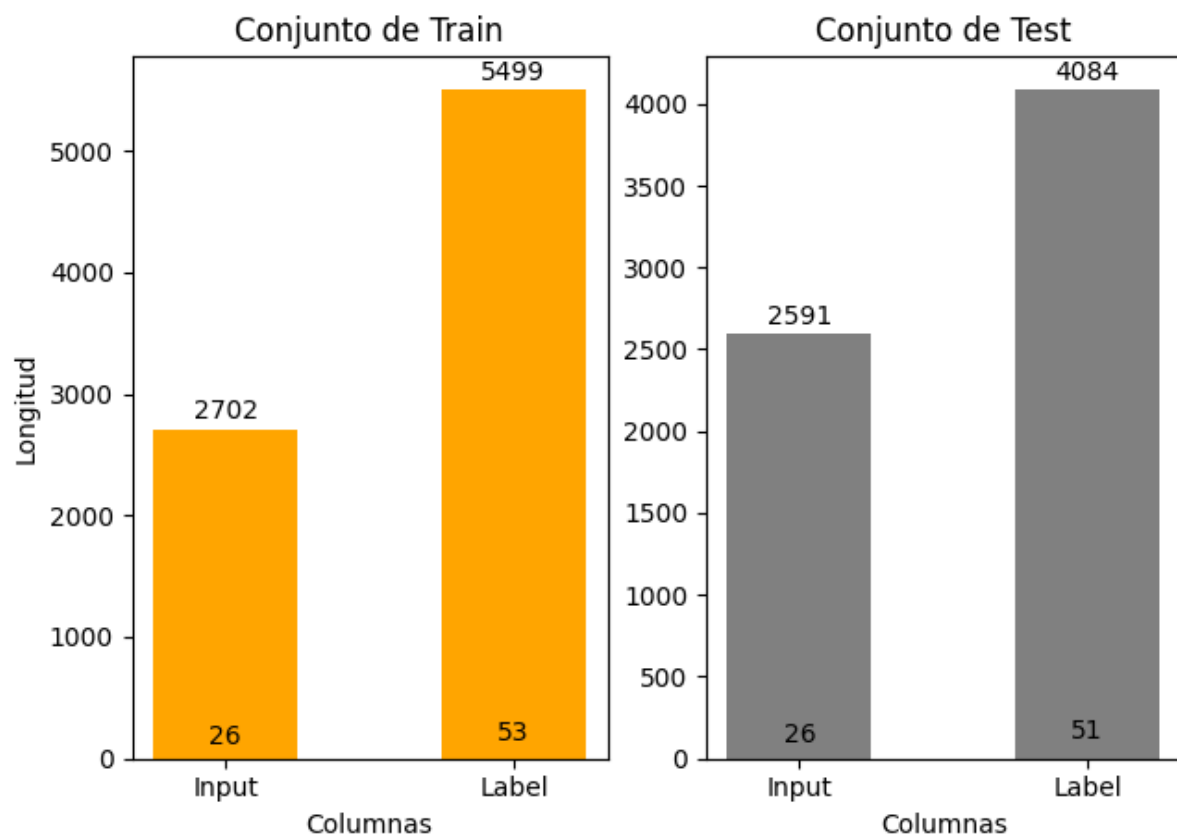


Counsel Chat

El dataset Counsel-chat es una recopilación hecha en colaboración entre (Bertagnolli, Lord, Lee, & Ström, 2020) con respuestas de terapeutas de alta calidad a preguntas sobre salud mental de pacientes reales. De acuerdo con el relato de los autores, los datos fueron inicialmente extraídos de www.counselchat.com, pero tras establecer contacto con los fundadores del sitio, se obtuvo acceso a la totalidad de los datos para el proyecto. La construcción del conjunto de datos surge de la necesidad de abordar la escasez de datos de salud mental de alta calidad disponibles para su estudio en Procesamiento del Lenguaje Natural (PNL), ya que la mayoría de los conjuntos de datos se centran en foros como Reddit, que no reflejan el lenguaje utilizado por los asesores. El objetivo principal del dataset ha

sido llenar este vacío al proporcionar datos de asesores que interactúan con pacientes necesitados, destacado por contar con respuestas de terapeutas verificados, lo que garantiza un nivel de calidad razonablemente alto, aunque no todas las respuestas sean excelentes, se sabe que provienen de expertos en el dominio.

Figura 16. Longitud máxima y mínima de Counsel Chat Dataset



La anterior gráfica muestra la longitud máxima y mínima que pueden alcanzar las instancias que componen el conjunto de datos, las cuales son en total 2127, repartidas en 1701 para entrenamiento y 426 para validación, a lo largo de 10 columnas, considerando únicamente 'questionText' y 'answerText'

questionText: El cuerpo de la pregunta del individuo a los consejeros.

answerText: La respuesta del terapeuta a la pregunta

Para la anteriores, un ejemplo representativo es el siguiente:

Tabla 13. Ejemplo del conjunto de entrenamiento Counsel chat

questionText (input)	Sometime when small thing happen, like losing a comb, I get furious. I even start throwing things. Do I have a problem?
-------------------------	---

answerText (label)	I don't like the word "problem". It is such a strong word. However I would say that you have an issue that needs to be addressed and controlled before it directed at the wrong person and gets totally out of control. I have found that anger management classes has been helpful for my clients. Try and see if it will help. You will be surprised at what you learn about your self.
-----------------------	---

Emotional Support Conversation – ESConv

La tarea de Conversación de Apoyo Emocional (ESC), es sustentada bajo la construcción de un conjunto de datos Emotional Support Conversation – ESConv por (Liu et al, 2021), con anotaciones ricas especialmente en estrategias de apoyo, recopiladas en un modo de búsqueda de ayuda y apoyo. Se evalúan modelos de diálogo avanzados en su capacidad para brindar apoyo emocional, destacando la importancia de las estrategias de apoyo y la utilidad de ESConv en la formación de sistemas de apoyo emocional. Además, se introduce ESConv para la investigación de habilidades de apoyo emocional en sistemas de diálogo, asegurando la calidad de las conversaciones para garantizar la eficacia del apoyo emocional.

Originalmente, el dataset cuenta con 1300 filas en torno a las columnas 'seeker_question1', 'problem_type', 'supporter_question2', 'emotion_type', 'dialog', 'survey_score', 'seeker_question2', 'experience_type', 'situation', 'supporter_question1'; sin embargo, a excepción de 'dialog', los demás atributos son removidos. La columna en mención, consiste en un diálogo de distinta longitud por cada ejemplo entre dos actores, denominados 'seeker' (la persona que busca ayuda) y 'supporter' (quien brinda el apoyo necesario).

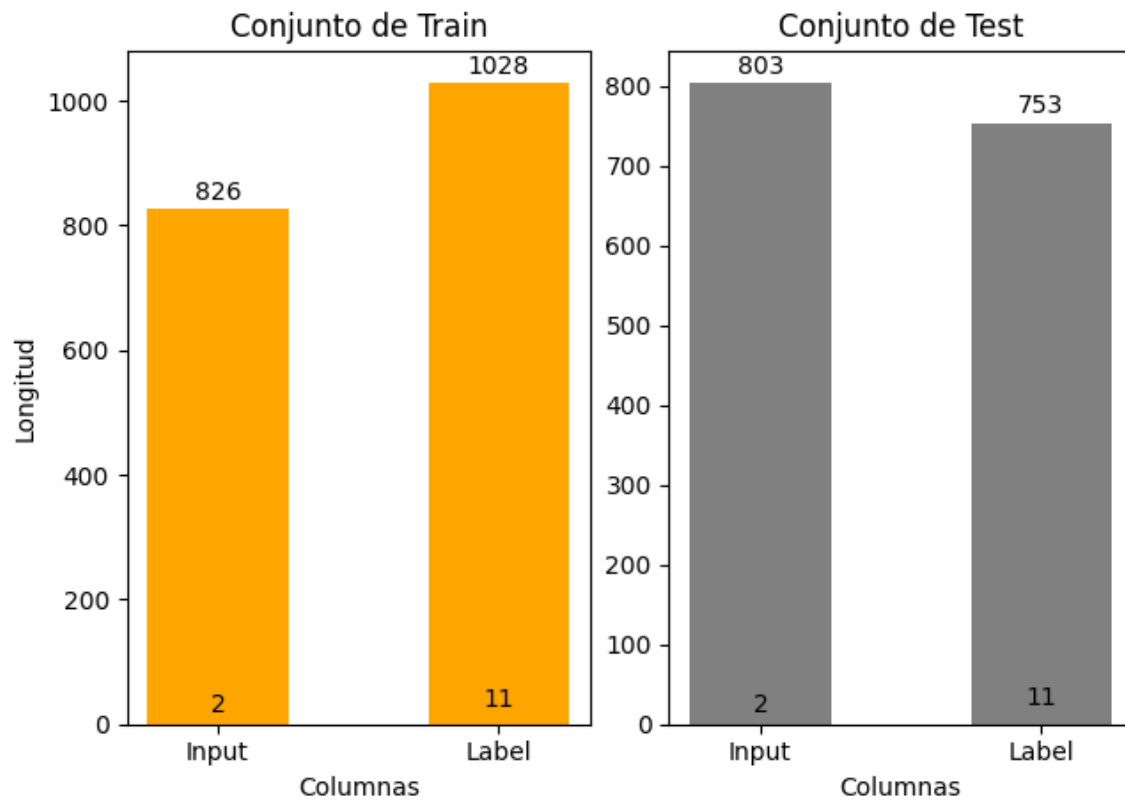
Con el fin de tener la misma estructura que el resto de conjuntos de datos que conforman EmpathetiCounseling, cada línea de diálogo es separada formando las tuplas entrada-etiqueta, tomando únicamente las conversaciones que son iniciadas por el actor "seeker"; de este modo, resultan 7003 ejemplos, de los cuales, 5602 son destinados como conjunto de entrenamiento y los 1401 restantes conforman el conjunto de validación.

Un ejemplo del dataset construido a partir de ESConv, así como la longitud máxima y mínima por columna, se muestran en la tabla y gráfica, respectivamente:

Tabla 14. Ejemplo del conjunto de entrenamiento ESConv

input	I was trying to stop the fight, I just said bit loud, "Stop all this nonsense at my place." Then they all started backfiring me.
label	oh,jst give them time to reflect on what happen, they will definitely reach out to you, but you also have to make an effort to forgo the situation because they see you raising your voice on them as a sign of command of seniority.

Figura 17. Longitud máxima y mínima en ESConv Dataset



Prompt-Aware margin Ranking (PAIR)

Para presentar el nuevo marco de puntuación de reflexión llamado Prompt-Aware margin Ranking (PAIR), (June Min, Pérez-Rosas, Resnicow, & Mihalcea, 2022) han recopilado un conjunto de datos que refleja diferentes niveles de habilidades de escucha reflexiva, donde cada interacción está en inglés e incluye un mensaje del cliente y respuestas de orientación que muestran diferentes niveles de habilidad de reflexión: calidad baja-lq, media-mq y alta-hq. Todos los modelos propuestos con la utilización del enfoque, incluidas las líneas de base, emplean la arquitectura RoBERTa e inician con pesos preentrenados de mental-roberta-base, como especifican los autores. Esta elección se debe a la similitud del dominio que caracteriza al conjunto de datos con el corpus de preentrenamiento utilizado para mental-roberta-base, que contiene publicaciones sobre salud mental de Reddit, emparejando solicitudes de consejo con respuestas; según explican los autores, experimentos preliminares muestran mejoras en el rendimiento general con los pesos preentrenados.

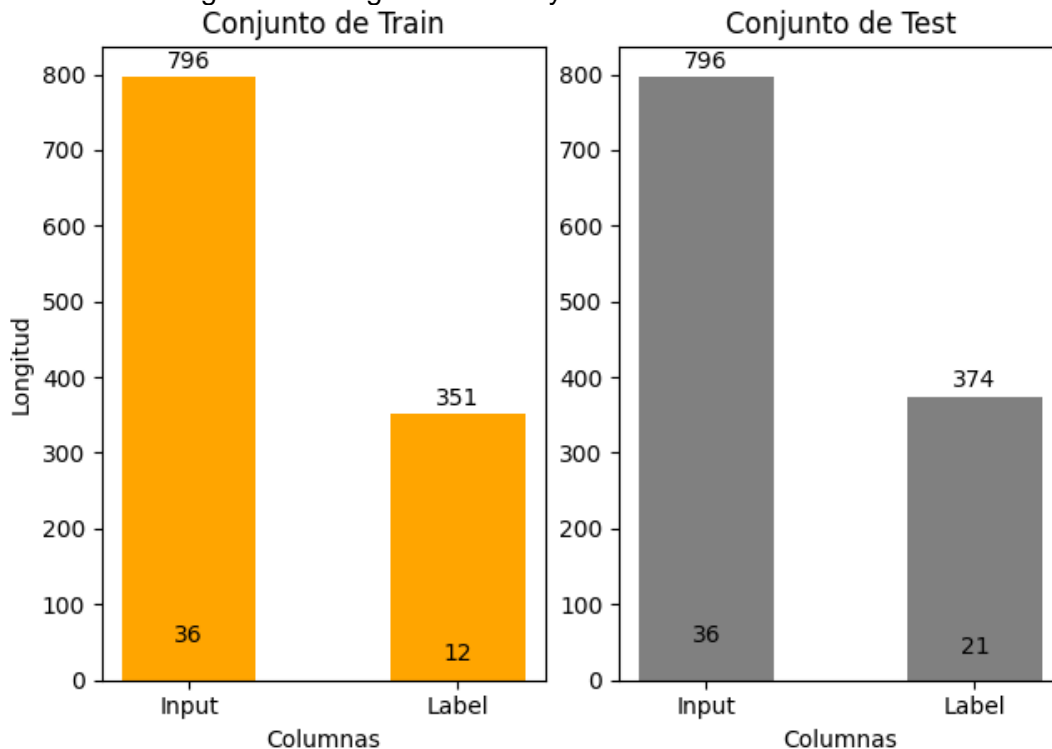
El conjunto de datos abarca 318 elementos, cada uno de los cuales está compuesto por una sentencia del cliente bajo la columna 'prompt', acompañada por 8 diferentes respuestas de acuerdo con los niveles de habilidad reflexiva: 'hq1', 'hq2', 'mq1', 'lq1', 'lq2', 'lq3', 'lq4', 'lq5'; el siguiente ejemplo ilustra en más detalle a qué hace referencia lo anterior.

Tabla 15. Ejemplo del conjunto de entrenamiento PAIR

prompt	I know I am too big, and I probably should exercise more and eat better, but I am so busy. I've got school, homework, and my job at the mall, so I don't see anywhere to fit it in. Plus, I can't afford any of those gyms. And none of my friends want to exercise with me. They're lazier than I am.
hq1	You are starting to think it's time to do something about your weight, and you know exercise and eating a little better would help. But fitting it in, between school and work, seems almost impossible. The gym isn't an option, and you can't think of any friends who would work out with you. But it is something you are starting to think about.
hq2	You have put a lot of effort into losing weight but it has not paid off. You are starting to feel a bit desperate. Diets don't seem to work for you. You are looking for something different that might last.\r\n
mq1	You don't know how you'd fit exercise into your schedule.
lq1	It's free to exercise at home. Maybe ride your bike or walk. Ask one of your parents to help. Try to eat some salads and fruit. Bring snacks to work and school.
lq2	Do you have a cheap gym near you, like a Planet Fitness?
lq3	You always have time for something, and about the lack of money, just work harder or even if you lack money, start training in public places.
lq4	Your feelings are valid, but I would advise that you take some time in your day for self care. It is important that you are happy with yourself. Where there is a will there is a way.
lq5	Start with small steps, start exercising at home and then progress. And really consider if you really don't have the time or just need to adjust your priorities.

Se decidió separar cada ejemplo en 8 tuplas, ofreciendo mayores alternativas de respuestas para una misma sentencia de entrada, ampliando el dataset a 2544 ejemplos, que luego se redujeron a 2521 tras su limpieza, distribuidos en 2016 elementos para entrenamiento y 505 para validación; la longitud máxima y mínima de los mismos se muestran en la gráfica.

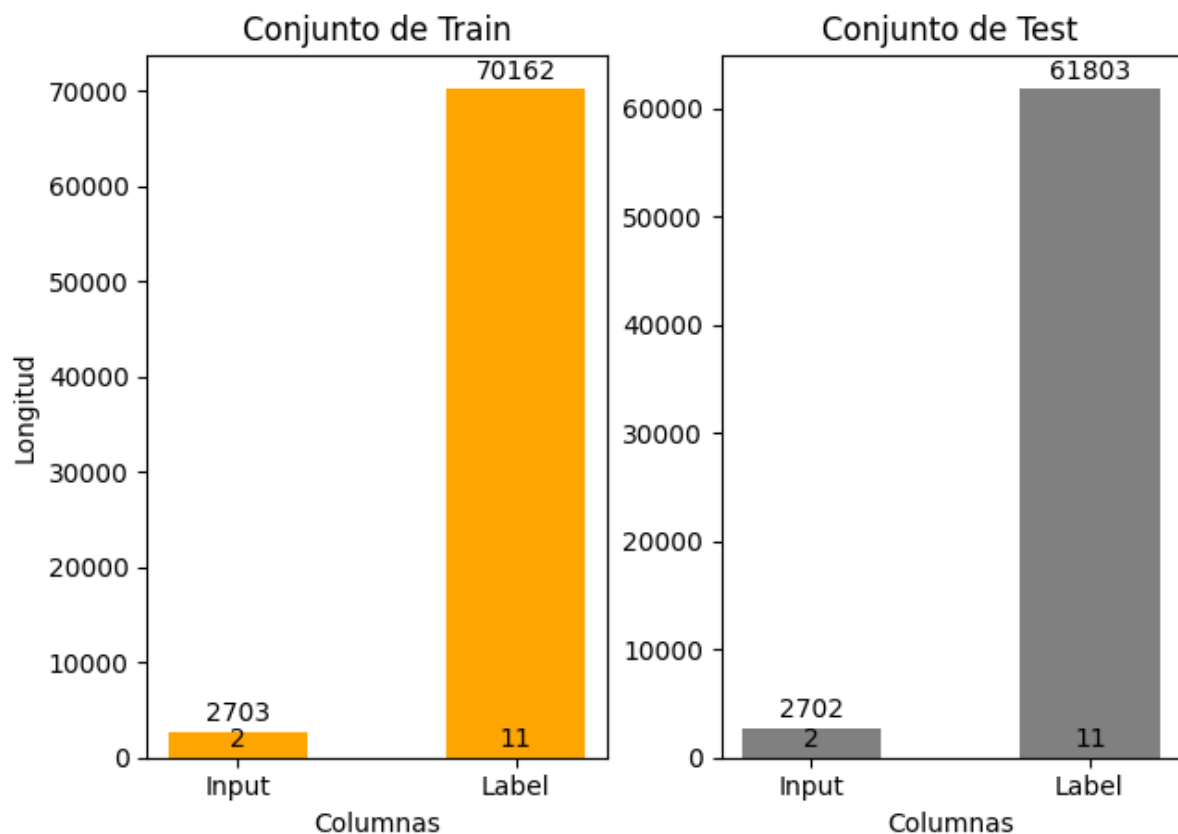
Figura 18. Longitud máxima y mínima en PAIR Dataset



EmpathetiCounseling Dataset

Una vez culminado el proceso de limpieza y preparación de los anteriores conjuntos de datos, se han concatenado para producir el dataset EmpathetiCounseling con 60000 elementos de diálogo repartidos en 48000 ejemplos de entrenamiento y 12000 ejemplos de validación. Por lo tanto, la longitud máxima y mínima

Figura 19. Longitud máxima y mínima de EmpathetiCounseling Dataset



Se debe aclarar que para los conjuntos que cuentan con una columna asociada a una etiqueta emocional, como EmpatheticDialogues y Counsel Chat, el proceso de split para dividir entre datos de entrenamiento y datos de validación se ha procurado realizarse de manera equitativa por cada etiqueta.

CMU-MOSI y CMU-MOSEI Datasets

Estos conjuntos de datos proporcionarán la base para el análisis multimodal de sentimientos. Al realizar un entrenamiento de ajuste sobre el modelo CM-BERT, estos datasets permitirán identificar la intensidad del sentimiento expresado por el usuario en

diferentes modalidades, como texto y audio, enriqueciendo así la capacidad del sistema para interpretar el contexto emocional de las interacciones.

Entrando un poco en detalle, CMU-MOSI (Multimodal Opinion of Sentiment Intensity) se enfoca en opiniones y sentimientos en un contexto de películas (Zadeh, Zellers, Pincus, & Morency, 2017), mientras que CMU-MOSEI (Multimodal Opinion-level Sentiment and Emotion Intensity) amplía este enfoque al incluir también la intensidad emocional en las opiniones (Zadeh A. , y otros, 2018), lo que lo convierte en un recurso más completo para la investigación en análisis de sentimientos y emociones. Ambos conjuntos de datos están disponibles públicamente para la investigación en el sitio web de CMU Multicomp Lab; la siguiente es una descripción comparativa de ambos conjuntos de datos, con información extraída del sitio y artículos oficiales.

1. Tipo de Datos.

- CMU-MOSI: Se centra en opiniones y sentimientos expresados en videos de monólogos, rigurosamente anotado con etiquetas de subjetividad, intensidad de sentimiento, características visuales anotadas por fotograma y por opinión, y características de audio anotadas por milisegundo.
- CMU-MOSEI: Aborda análisis de sentimientos y reconocimiento de emociones en videos de interacciones sociales y opiniones. El conjunto de datos está equilibrado en cuanto a género. Todas las frases pronunciadas se eligen aleatoriamente entre varios temas y vídeos de monólogos. Los vídeos se han transcrito y puntuado correctamente.

2. Tamaño del Conjunto de Datos.

- CMU-MOSI: Contiene 2199 clips de videos cortos, obtenidos de diversos usuarios de YouTube. Cada vídeo de opinión está anotado con un grado de sentimiento en el rango: [-3: strongly negative, -2 negative, -1 weakly negative, 0 neutral, +1 weakly positive, +2 positive, +3 strongly positive].
Además de la opción “uncertain” para el caso de no estar seguro.
- CMU-MOSEI: Es el conjunto de datos más grande, con múltiples modalidades y más de 23,500 muestras compuestas por vídeos de frases de más de 1.000 oradores en línea de YouTube. Sus anotaciones son congruentes con las de MOSI, en una escala sentimental de: [negative, weakly negative, neutral, weakly positive, positive]. Y etiquetas de emociones: [happiness, sadness, anger, disgust, surprise, fear].

3. Modalidades de Datos.

- CMU-MOSI: Presentado como el primer corpus anotado a nivel de opinión para el análisis del sentimiento y la subjetividad en vídeos en línea, se centra principalmente en datos unimodales de video y texto.
- CMU-MOSEI: Presentado como el mayor conjunto de datos de análisis de sentimientos y reconocimiento de emociones hasta la fecha, ofrece una variedad de modalidades, incluyendo video, texto y audio.

4. Estructura de los datos.

- CMU-MOSI: El conjunto de observaciones multimodales incluye:
 - Transcripción del habla, gestos y características visuales y de audio automáticas.
 - Segmentación subjetiva a nivel de opinión.
 - Anotaciones sobre la intensidad del sentimiento.
 - Alineación entre palabras y características visuales y acústicas
- CMU-MOSEI: Los datos se presentan en formato de vídeo con un hablante delante de la cámara, de los que se extrae:
 - Transcripción manual del habla.
 - Fotogramas de los vídeos completos a 30 Hz.
 - Características acústicas relacionadas con las emociones y tono de voz.

5. Complejidad de Análisis.

- CMU-MOSI: Menos complejo debido a la naturaleza de los monólogos, lo que puede simplificar algunas tareas de análisis.
- CMU-MOSEI: Puede ser más complejo debido a la diversidad de contextos y modalidades de datos, lo que requiere enfoques de análisis más sofisticados.

Tabla 16. Resumen de las estadísticas del conjunto de datos MOSI

Número total de segmentos	3702
Número total de segmentos de opinión	2199
Número total de segmentos objetivos	1503
Número total de vídeos	93
Número total de oradores distintos	89
Número medio de segmentos de opinión en vídeo	23.2
Duración media de los segmentos de opinión (segundos)	4.2
Número medio de palabras por segmento de opinión	12
Número total de palabras en los segmentos de opinión	26,295
Número total de palabras únicas en los segmentos de opinión	3,107
Número total de palabras en segmentos de opinión que aparecen al menos 10 veces en el conjunto de datos	557

Fuente: (Zadeh, Zellers, Pincus, & Morency, 2017)

Tabla 17. Resumen de las estadísticas del conjunto de datos CMU-MOSEI

Número total de frases	23453
Número total de vídeos	3228
Número total de oradores distintos	1000
Número total de temas distintos	250
Número medio de frases en un vídeo	7.3
Duración media de las frases en segundos	7.28
Número total de palabras en las frases	447143

Número total de palabras únicas en las frases	23026
Número total de palabras que aparecen al menos 10 veces en el conjunto de datos	3413
Número total de palabras que aparecen al menos 20 veces en el conjunto de datos	1971
Número total de palabras que aparecen al menos 50 veces en el conjunto de datos	888

Fuente: (Zadeh A. , y otros, 2018)

API SeamlessM4T

Esta API ofrecerá funcionalidades esenciales para la traducción multimodal y multilingüaje. Al integrar SeamlessM4T, se podrá convertir texto a voz, voz a texto y realizar traducciones en tiempo real en múltiples idiomas. Esto garantizará que el sistema sea accesible y útil para usuarios de diferentes regiones y con diversas preferencias lingüísticas y comunicativas.

En conjunto, este recurso con los modelos descritos en la sección “Métodos”, proporcionarán las herramientas necesarias para desarrollar un sistema de apoyo psicológico personalizado que sea sensible, efectivo y adaptable a las necesidades individuales de cada usuario. A continuación, se describirá detalladamente cada uno de sus elementos y su contribución al método de desarrollo del sistema.

1. **Expansión de la cobertura de idioma.** SeamlessM4T es un modelo basado en inteligencia artificial (IA) desarrollado por Meta (anteriormente Facebook) que se especializa en la traducción de texto a voz y viceversa en cerca de 100 idiomas diferentes. Esta capacidad multilingüaje lo distingue como una herramienta versátil para superar barreras lingüísticas en diversas aplicaciones y contextos de comunicación. Recientes avances han ampliado la cobertura de idiomas y mejorado las capacidades multimodales de traducción automática de voz.
2. **Limitaciones actuales.** A pesar de estos avances, los sistemas de traducción automática de voz a gran escala aún carecen de características clave que proporcionen una comunicación fluida, similar al diálogo humano.
3. **Modelos mejorados.** La API SeamlessM4T proporciona servicios de traducción de alta calidad y precisión, permitiendo la conversión fluida entre diferentes idiomas en tiempo real. Ofrece funciones avanzadas de reconocimiento y síntesis de voz para una experiencia de usuario óptima. Una versión mejorada del modelo original es SeamlessM4T v2, entrenada con más datos de idiomas de recursos limitados. Además, incorpora el marco UnitY2 actualizado.

4. **Funcionalidades de traducción.** El conjunto de modelos soporta las siguientes tareas.

- Speech-to-speech translation (S2ST)
- Speech-to-text translation (S2TT)
- Text-to-speech translation (T2ST)
- Text-to-text translation (T2TT)
- Automatic speech recognition (ASR)

La API incluye modelos como SeamlessExpressive y SeamlessStreaming. La primera conserva estilos vocales y prosodia para la traducción speech-to-speech, mientras que la segunda genera traducciones de baja latencia sin esperar las oraciones completas, al recibir entradas de audio y retornar salidas con modalidad audio/texto.

5. **Modo de Uso.** Los desarrolladores pueden integrar fácilmente la API SeamlessM4T en aplicaciones y sistemas existentes mediante llamadas de API simples y claras. Esto permite la implementación de traducción multilinguaje en una amplia variedad de contextos y plataformas. Para saber cómo integrar los modelos Seamless a un sistema, siga el siguiente [tutorial de implementación](#).

6. **Evaluación de rendimiento.** Se emplean métricas automáticas novedosas y adaptadas para evaluar prosodia, latencia y robustez. Además, se realizan evaluaciones humanas centradas en la preservación de significado, naturalidad y expresividad.

7. **Seguridad y responsabilidad.** Se implementan medidas para garantizar el uso seguro y responsable de los modelos, como la detección y mitigación de toxicidad añadida, evaluación sistemática de sesgo de género y mecanismos de marca de agua para mitigar el impacto de deepfakes.

8. **Formación de la API.** Se combinan componentes de SeamlessExpressive y SeamlessStreaming para formar Seamless, el primer sistema disponible públicamente que permite la comunicación expresiva y multilinguaje en tiempo real.

Para usar la API se inicializa un objeto Translator y se usa su función 'predict' para obtener una salida traducida del mensaje original, como muestra el pseudocódigo.

Algoritmo. Uso de la API SeamlessM4T

Input: mensaje en idioma y modalidad original, output: mensaje traducido

```
# Inicializa un objeto Translator con un modelo multitarea, vocoder en la GPU.
model_name = "seamlessM4T_v2_large"
vocoder_name = "vocoder_v2" if model_name == "seamlessM4T_v2_large"
else "vocoder_36langs"

traductor = Translator(
    model_name,
    vocoder_name,
    device=torch.device("cuda:0"),
    dtype=torch.float16,
)
```

Traducción del mensaje a la modalidad e idioma deseado

```
tgt_lang = idioma_objetivo
mensaje = "dirección/del/archivo" #Audio o directamente el mensaje textual
text_output, speech_output = traductor.predict(
    input=mensaje,
    task_str=tarea,
    tgt_lang=tgt_lang,
    src_lang=idioma_origen)
```

7. Métodos

El presente trabajo final de máster se centra en el desarrollo de un sistema conversacional innovador para la generación de contenido autoayuda y apoyo psicológico personalizado, como indica el propio título. Este sistema tiene como objetivo principal proporcionar un apoyo emocional efectivo y adaptado a las necesidades individuales de cada usuario.

Para lograr este propósito, se combinarán diversas técnicas y tecnologías avanzadas, descritas en orden de importancia. En primer lugar, se implementará la generación de respuestas empáticas, donde el sistema demostrará comprensión y empatía hacia los sentimientos y las emociones expresadas por el usuario. Esta característica es fundamental para establecer una conexión emocional significativa y ofrecer un apoyo genuino.

Además, el sistema integrará un análisis multimodal de sentimientos, permitiendo interpretar no solo el texto, sino también otros elementos comunicativos como tono de voz, expresiones faciales y gestos corporales. Esta capacidad mejorará la precisión en la comprensión del estado emocional del usuario, facilitando una respuesta más adecuada y personalizada.

Un último aspecto destacado de este sistema es su capacidad de traducción multimodal y multilinguaje, que permitirá la comunicación fluida en distintos idiomas y modalidades. Estas características se basan en los principios fundamentales de un sistema de apoyo psicológico eficiente, que incluye la empatía y escucha activa, la comunicación abierta, la personalización, la disponibilidad y consistencia, así como la confidencialidad y privacidad; garantizando su disposición para una amplia variedad de usuarios, independientemente de su idioma nativo y la modalidad de comunicación, entre otras cuestiones de su situación de vida actual.

Elección del modelo idóneo

Para el cumplimiento del propósito descrito en el ámbito de la generación de contenido autoayuda y apoyo psicológico personalizado, la elección del modelo de inteligencia artificial idóneo es fundamental para garantizar la eficacia y la precisión del sistema. Por tal motivo, a continuación, se presenta una comparativa entre los modelos discutidos en el apartado "Estado del Arte". Esta comparación tiene como objetivo analizar los resultados y el rendimiento de los distintos modelos para dos de las tres tareas principales del sistema: Generación de Respuestas Empáticas y Análisis Multimodal de Sentimientos; en cuanto a la Traducción Multimodal y Multilinguaje, se ha definido aprovechar el potencial de la API SeamlessM4T por amplia gama de funciones que combinan los distintos usos de esta tarea en un único recurso. Adicionalmente, este análisis servirá como guía para evaluar y seleccionar el modelo más adecuado para cada tarea del sistema, considerando sus fortalezas y limitaciones en términos de rendimiento y las métricas de evaluación.

Tabla 18. Comparación del rendimiento entre los modelos de Generación de Respuesta Empática

Modelo	Precisión de la Emoción	Perplejidad	Distinc-1	Distinc-2
MIME	0.2938	37.08	0.31	1.03
EmpDG	0.3003	37.77	0.59	2.48
CEM	0.3644	37.03	0.66	2.99
SEEK	0.4185	37.09	0.73	3.23
RoBERTa w/o GPT-2	0.3439	-	-	-
RoBERTa-GPT2 w/o CKECE	0.5262	14.97	1.62	10.47
RoBERTa-GPT2	0.5151	13.57	2.04	11.68

Tabla 19. Comparación del rendimiento entre los modelos de Análisis Multimodal de Sentimiento

Modelo: SPECTRA						
Tarea	Dataset	Métrica	Resultado			
Análisis Multimodal de Sentimiento	MOSI	Acc ₂	87.50			
	MOSEI	Acc ₂	87.34			
Reconocimiento de Emociones en Conversación	IEMOCAP	Acc	67.94			
Comprensión de la Lengua Hablada	MIIntRec	Acc ₂₀	73.48			
Seguimiento del Estado del Diálogo	SpokenWoz	JGA	21.96			
Modelo: CM-BERT						
Tarea	Dataset	Acc ₇	Acc ₂	F1	MAE	Corr
Análisis Multimodal de Sentimiento	MOSI	44.9	84.5	84.5	0.729	0.791

Tras realizar un análisis exhaustivo del estado del arte en el campo de la generación de contenido autoayuda y apoyo psicológico personalizado, en conjunto con la comparativa anterior, se ha determinado la implementación de los siguientes modelos para cada una de las tres tareas de Procesamiento de Lenguaje Natural, como se describe a continuación.

1. **Generación de Respuestas Empáticas.** Por razones de simplicidad de los primeros modelos en relación al proceso de ajuste o fine-tuning (los pasos a seguir para el entrenamiento se especifican en los respectivos repositorios de GitHub, estando ya definidos los valores óptimos de cada hiperparámetro) en comparación

con la propuesta de (Liu, Maier, Minker, & Ultes, 2021), así como la superioridad de este último en términos de rendimiento y resultados en las distintas métricas, para abordar esta tarea, se opta por desarrollar un modelo inspirado en el denominado **RoBERTa-GPT2**, tratando de simular su funcionamiento y alcances. Este modelo será entrenado específicamente para generar respuestas empáticas y de asesoramiento psicológico que sigan una modalidad textual, brindando así un soporte emocional efectivo a los usuarios; sin embargo, se siguen 4 estrategias para determinar el mejor modelo para esta tarea, que se detallarán más adelante.

2. **Análisis Multimodal de Sentimientos.** Para llevar a cabo el análisis multimodal de sentimientos, se utilizará el modelo **CM-BERT**. Esta elección se fundamenta en la capacidad del modelo para interpretar y procesar diferentes modalidades de entrada, como texto, audio y video, proporcionando una comprensión más profunda y precisa del estado emocional del usuario, a partir de elementos comunicativos como el tono de voz, expresiones faciales y gestos corporales.
3. **Traducción Multimodal y Multilenguaje.** Para facilitar la comunicación en distintos idiomas y modalidades, se integrará la API de **SeamlessM4T**. Esta herramienta permitirá la traducción entre texto-a-texto, voz-a-texto y texto-a-voz en una amplia variedad de idiomas, garantizando así la accesibilidad y la inclusión para usuarios de diferentes regiones y culturas.

Es importante destacar que el proceso de fine-tuning de los modelos se llevará a cabo con conjuntos de datos específicos. Para el modelo RoBERTa-GPT2, se utilizará el dataset EmpathetiCounseling construido como se ha descrito en el apartado 6 “Material”, el cual es preprocesado para adaptarse a las necesidades del sistema. Por otro lado, el proceso de entrenamiento del modelo CM-BERT se asemeja al enunciado en los respectivos repositorios de los modelos generadores de respuesta empática que no fueron elegidos, empleándose los conjuntos de datos CMU MOSI y CMU-MOSEI para la tarea de Análisis Multimodal de Sentimiento.

Esta estrategia de implementación garantiza la eficacia y la precisión del sistema, permitiendo así ofrecer un apoyo psicológico personalizado y de alta calidad a los usuarios que lo necesiten. Sin embargo, para una mejor comprensión de los métodos y la metodología implementada, a continuación, se profundizará en los aspectos clave alusivos a la arquitectura, proceso de entrenamiento y Fine-tuning, así como los resultados de evaluación del mismo.

RoBERTa-GPT2

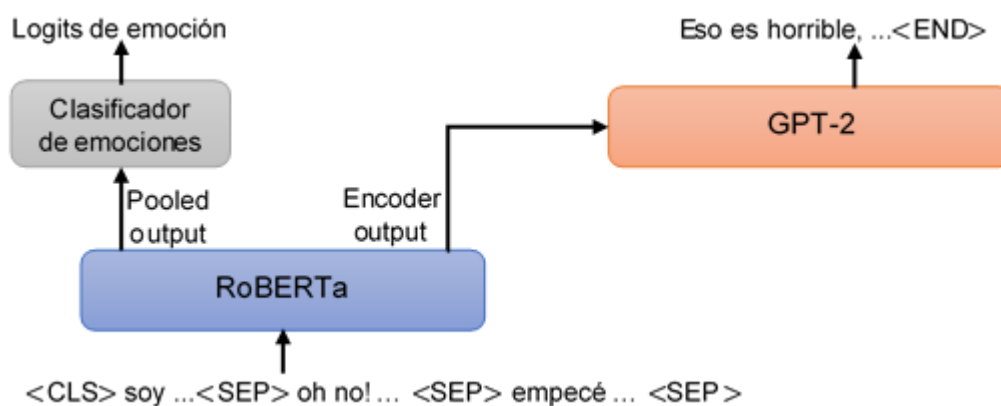
En la propuesta, los autores presentan una arquitectura codificador-decodificador RoBERTa-GPT2 para la generación de diálogos empáticos. El codificador RoBERTa, preentrenado, funciona como el codificador, mientras que el decodificador GPT-2,

autorregresivo y preentrenado, actúa como el decodificador. Además, se introduce un Extractor de Conocimientos y Conceptos Emocionales de Sentido Común (CKECE), que se encarga de extraer los conceptos relevantes de la historia del diálogo para permitir la capacidad de empatía y sentido común del decodificador GPT-2.

El CKECE utiliza dos fuentes de conocimiento: ConceptNet y NRC_VAD, junto con la herramienta de extracción de palabras clave KeyBERT. Se extraen palabras clave del contexto del diálogo mediante KeyBERT y luego se filtran los conceptos de sentido común y los conceptos emocionales más relevantes utilizando ConceptNet y NRC_VAD.

Es importante aclarar que, para la construcción del modelo similar a la propuesta, no se incorporará el extractor de conocimiento, es decir, se tratará de una arquitectura codificador-decodificador RoBERTa-GPT2 sin CKECE, como muestra la siguiente figura extraída del artículo original y modificada para fines ilustrativos.

Figura 20. Arquitectura codificador-decodificador de RoBERTa-GPT2

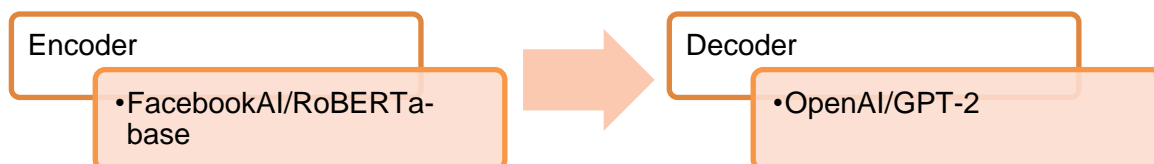


Nota. Modificación de la imagen original, extraída de (Liu, Maier, Minker, & Ultes, 2021)

Como se ha mencionado, se siguen 4 estrategias para la construcción del modelo generador de respuestas de diálogo combinando distintos modelos en la arquitectura de modelo codificador-decodificador, así:

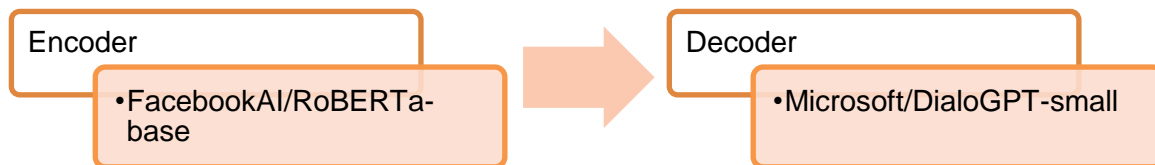
1. Modelo RoBERTa-GPT2 tal como fue hecho por los autores. Se entrenan RoBERTa y GPT-2 por separado, partiendo de los hiperparámetros usados por los autores: `learning_rate=1e-5` y `batch_size=16`.

Figura 21. Estrategia 1: Modelo RoBERTa-GPT2



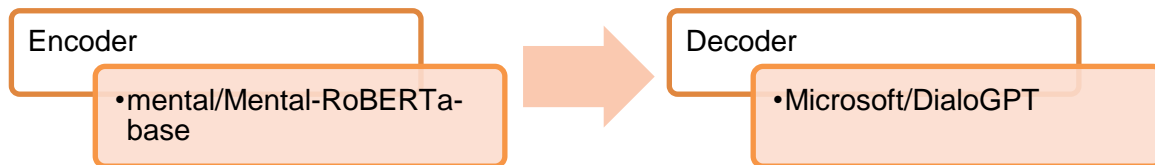
2. Modelo RoBERTa-DialoGPT que reemplaza a GPT-2 por su versión extendida para tareas de generación de diálogos. A diferencia de la anterior, esta estrategia adopta a DialoGPT de Microsoft como decodificador.

Figura 22. Estrategia 2: Modelo RoBERTa-DialoGPT



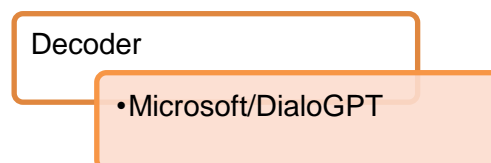
3. Modelo MentalRoBERTa-DialoGPT que combina un codificador enfocado en temas de salud mental y un decodificador específico para tareas de carácter conversacional. Al igual que las anteriores estrategias, ambos modelos son entrenados previamente.

Figura 23. Estrategia 3: Modelo MentalRoBERTa-DialoGPT



4. Modelo DialoGPT que únicamente hace uso del decodificador. Esta metodología se emplea de manera analítica con respecto al impacto que podría tener un modelo encoder en la capacidad de generación de diálogos del modelo DialoGPT debido a las diferencias en la tokenización, la compatibilidad del vocabulario y la transferencia de conocimiento limitada.

Figura 24. Estrategia 4: Modelo DialoGPT



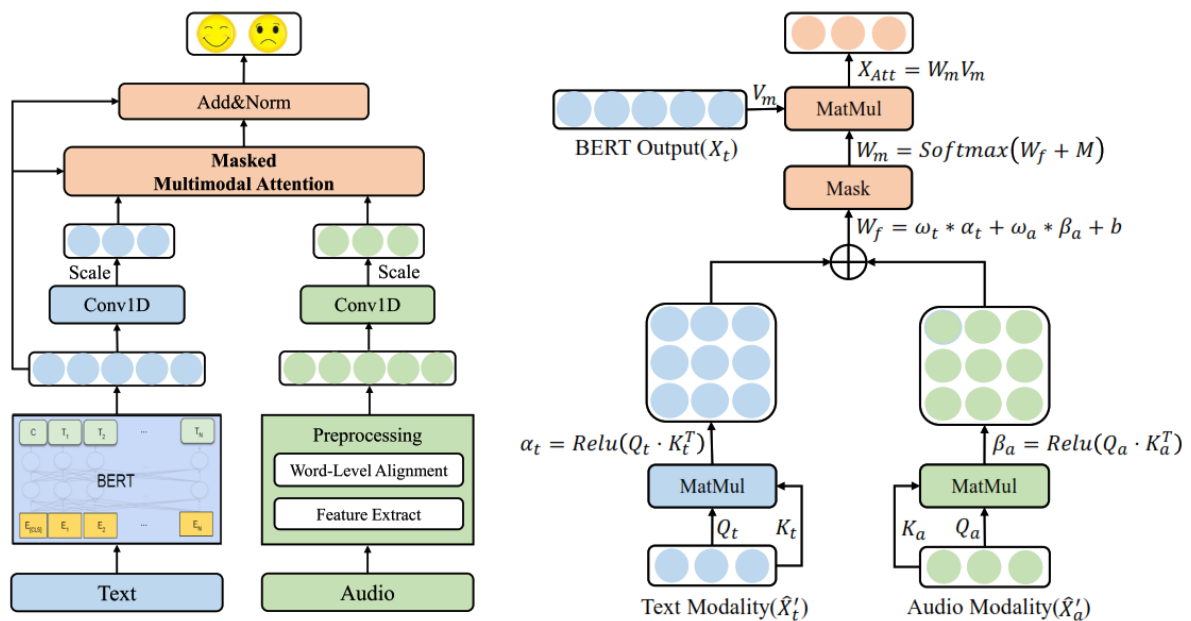
Para cada estrategia, se entrenan independientemente los modelos RoBERTa, MentalRoBERTa, GPT-2 y DialoGPT con el dataset EmpathetiCounseling para ser ajustados en temas de respuestas empáticas y asesoramiento médico y psicológico. Luego cada modelo se toma para la construcción de la arquitectura codificador-decodificador de acuerdo con su respectiva estrategia, cada una de las cuales es posteriormente evaluada para identificar aquella que ofrece un mejor rendimiento en relación a las respuestas empáticas y de asesoramiento o apoyo psicológico.

CM-BERT

En su investigación, los autores introducen Cross-Modal BERT, una adaptación multimodal de BERT (CM BERT), que integra información tanto de texto como de audio para mejorar el modelo BERT preentrenado. En la arquitectura de CM-BERT representada en la figura 25 (izquierda), la entrada consta de dos componentes: una secuencia de texto de tokens de palabras y características de audio de alineación a nivel de palabra. Inicialmente, la secuencia de texto se procesa a través del modelo BERT preentrenado, y la salida de la última capa del codificador se utiliza como características del texto.

Por otro lado, el núcleo de CM BERT se presenta como una arquitectura de atención multimodal enmascarada, diseñada para incorporar información de la modalidad de audio en el proceso de ajuste del peso de las palabras y refinamiento del modelo BERT preentrenado; figura 25 (derecha).

Figura 25. Arquitectura de CM-BERT y de su Atención Multimodal Enmascarada



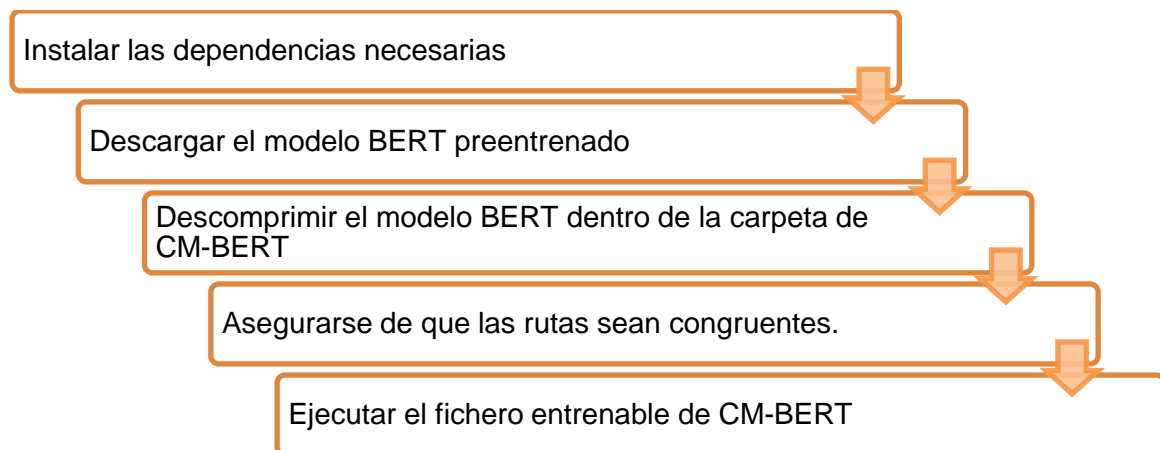
Nota. A la izquierda se muestra la arquitectura CM-BERT y a su derecha está la estructura que se encarga de la Atención Multimodal Enmascarada. Fuente: (Yang, Xu, & Gao, 2020)

La evaluación se llevó a cabo en los conjuntos de datos CMU-MOSI y CMU-MOSEI. Los autores dividieron 52, 10 y 31 vídeos en conjuntos de entrenamiento, validación y prueba, respectivamente, asegurando la ausencia del hablante en estos conjuntos y el equilibrio entre datos positivos y negativos. Esto resultó en 1284, 229 y 686 enunciados para cada conjunto. CMU-MOSEI, al igual que CMU-MOSI, es un conjunto de datos multimodal de análisis de sentimientos y emociones, compuesto por 23,454 vídeos de críticas.

El modelo BERT preentrenado utilizado en CM-BERT es la versión `bert-base-uncased`, compuesta por 12 bloques transformadores. Se configuraron las tasas de aprendizaje de las capas codificadoras y del resto de las capas para evitar el sobreajuste. Los parámetros de la capa de incrustación se mantuvieron congelados para mejorar el rendimiento. El entrenamiento del modelo CM-BERT se realizó con un tamaño de lote de 24, una longitud máxima de secuencia de 50 y tres épocas, utilizando el optimizador *Adam* con la función de pérdida *mean - squarererror*.

Entrando en detalle en el proceso de entrenamiento para el cumplimiento de la tarea de Análisis Multimodal de Sentimiento dentro del sistema de generación de contenido autoayuda y apoyo psicológico personalizado, primeramente, se accedió al [repositorio de GitHub](#) y, luego de haberlo clonado en un cuaderno de Colab, se siguió las instrucciones recomendadas en el mismo:

Figura 26. Pasos de entrenamiento del modelo CM-BERT



Adicional a estos, se alojó el modelo entrenado en un repositorio de Hugging Face para su posterior uso en el debido momento. Para este modelo no se enfatizó mucho en su entrenamiento debido al nivel de importancia inferior a la tarea de Generación de Respuesta Empática para el presente trabajo final de máster, ilustrando, en términos generales, la sencillez en el entrenamiento de los modelos dedicados a la tarea mencionada y que han sido anteriormente comparados con la arquitectura RoBERTa-GPT2 trabajada.

Funcionamiento general

Habiendo discutido en mayor profundidad la consistencia y funcionalidad de los modelos que conforman el sistema completo de generación de contenido autoayuda y apoyo psicológico personalizado, se describe el método general para el cumplimiento de las tareas a las cuales se hizo mención.

El novedoso sistema conversacional propuesto para la generación de contenido autoayuda y apoyo psicológico personalizado consiste en un agente de diálogo, desarrollado como una aplicación móvil que brinda al usuario la oportunidad de alternar entre modalidad textual, acústica y visual acorde a sus necesidades y preferencias, recibiendo una respuesta escrita o hablada gracias a la traducción texto-a-texto, texto-a-voz y voz-a-texto y entre 100 idiomas distintos, luego de que el modelo procese la entrada del usuario reconociendo la emoción implícita en el mensaje y analizando la intensidad de sentimiento congruente con el tono de voz, expresiones faciales y gestos corporales, para dar una respuesta empática acorde al contexto y contenido del mismo, que ofrezca el acompañamiento emocional y psicológico requerido por el usuario.

El funcionamiento de los componentes del sistema se divide en 12 casos de acuerdo con la modalidad e idioma de la entrada (input) al modelo y su respectiva salida (output), considerando que el modelo de generación de respuestas empáticas se entrena con un conjunto de datos compuesto por ejemplos textuales en inglés (dataset Empathetic Dialogues); como se muestra en la tabla.

Tabla 20. Casos en la entrada y salida del modelo

Modalidad		Idioma
Entrada	Salida	
Texto	Texto	Inglés
		Otro
	Audio	Inglés
		Otro
Audio	Texto	Inglés
		Otro
	Audio	Inglés
		Otro

El modelo predeterminado responderá en inglés y en modalidad elegida aleatoriamente para permitir alternancia entre respuestas textuales y acústicas; sin embargo, si la intensidad de sentimiento expresado en el mensaje del usuario (analizado por CM-BERT) o la longitud de la respuesta superan cierto umbral, se establece la respuesta siguiendo la modalidad de audio. Esto se debe a que, en momentos de crisis emocional o al compartir noticias emocionantes, es más reconfortante escuchar palabras de aliento que simplemente leerlas. Además, se brinda al usuario la flexibilidad de personalizar la modalidad y el idioma de respuesta según sus preferencias y requisitos.

Esta estrategia se alinea con la teoría de la comunicación emocional de Mehrabian (Mehrabian & Ferris, 1967), que destaca que una gran parte de la comunicación se transmite a través del tono de voz y las expresiones faciales, más que a través del texto escrito. Mehrabian argumentó que solo un pequeño porcentaje de la comunicación emocional se transmite a través de las palabras, mientras que el tono de voz y el lenguaje corporal juegan un papel significativo en la expresión de emociones. Por lo tanto, en situaciones como las indicadas, la comunicación oral puede ser más efectiva para transmitir empatía y apoyo, lo cual es más efectivo para transmitir emociones que el texto escrito.

El funcionamiento general del sistema y los modelos que lo componen, siempre y cuando el usuario no haya configurado una modalidad única, se ilustra en el siguiente pseudocódigo.

Algoritmo. Generación de respuesta empática multimodal y multilinguaje

Input: mensaje del usuario, output: respuesta del generador

```
# Guarda el mensaje, la intensidad de sentimiento predicha y umbrales para  
dicha intensidad y la longitud de la salida generada
```

```
entrada = mensaje_usuario
```

```
umbral_sentim =  $\alpha$ 
```

```
umbral_salida =  $\beta$ 
```

```
intens_sentim = CM-BERT(entrada)
```

```
# Traducción al idioma inglés si el idioma de la entrada es distinto
```

```
Si idioma_origen != 'inglés':
```

```
    entrada = SeamlessM4T(input=entrada,  
                           task_str=tarea  
                           src_lang=idioma_origen,  
                           tgt_lang="eng")
```

```
# Procesa el mensaje y genera la respuesta
```

```
mensaje_tokenizado = RobertaTokenizer(entrada)
```

```
mensaje_codificado = RobertaModel(mensaje_tokenizado).encoder
```

```
respuesta = GPT2Model(mensaje_codificado)
```

```
salida = GPT2Tokenizer(respuesta).decoder
```

```
# Retorna la respuesta al idioma personalizado por el usuario
```

```
Si idioma_origen != 'inglés':
```

```
    salida = SeamlessM4T(input=salida,  
                          task_str=tarea  
                          src_lang="eng",  
                          tgt_lang=idioma_origen)
```

```
# Se determina la modalidad de la respuesta
```

```
modalidad_audio = False
modalidad = random('texto', 'audio')
Si (intens_sentim > umbral_sentim) o (modalidad == 'audio'):
    modalidad_audio = True
```

```
# Se traduce a la modalidad definida si es necesario
```

```
Si modalidad_audio o (len(salida) > umbral_salida):
    salida_audio = SeamlessM4T(input=salida,
                                task_str=tarea,
                                src_lang="eng",
                                tgt_lang=idioma_origen)
    salida = salida_audio
```

Considere lo siguiente del contenido presente en el pseudocódigo:

- La entrada para la función del modelo CM-BERT, así como su funcionamiento, depende de la modalidad empleada por el usuario.
- El hiperparámetro *task_str* de la API SeamlessM4T es una cadena o string que indica la tarea de traducción, según esta lista: 's2st', 's2tt', 't2st', 't2tt'.
- Los hiperparámetros *src_lang* y *tgt_lang* de la API SeamlessM4T indican el idioma de origen y el idioma objetivo, respectivamente. Reciben una cadena o string con la abreviatura del idioma en cuestión.

A continuación, se especifican los modelos y el procedimiento involucrados en cada caso presentado cuando la modalidad de entrada es lingüística o acústica, con el propósito de desglosar detalladamente el pseudocódigo anterior para una mejor comprensión del mismo (bajo la modalidad visual el sistema se comporta semejante a la acústica).

1. **Entada: Texto, Salida: Texto, Idioma: inglés.** Este caso se caracteriza por los siguientes aspectos.

Modelo:

- RoBERTa-GPT2

Procedimiento:

- Mensaje textual del usuario.
- El texto es recibido por el codificador RoBERTa para su procesamiento, prediciendo la emoción presente.
- La salida anterior será la entrada para el decodificador GPT-2, que genera una respuesta empática textual.

2. **Entada: Texto, Salida: Texto, Idioma: Otro.** El idioma del sistema es configurado por el usuario previamente, siendo distinto al inglés. Aquí hace presencia la API de SeamlessM4T, así.

Modelo:

- RoBERTa-GPT2.
- SeamlessM4T

Procedimiento:

- Mensaje textual del usuario en un idioma distinto al inglés.
- El texto es recibido y procesado por la API SeamlessM4T, otorgando una salida en idioma inglés gracias a su traducción texto-a-texto (t2tt).
- El nuevo mensaje es recibido por el codificador RoBERTa para su procesamiento, prediciendo la emoción presente.
- La salida anterior será la entrada para el decodificador GPT-2, que genera una respuesta empática textual.
- La respuesta se traduce nuevamente al idioma original por la API SeamlessM4T.

3. **Entada: Texto, Salida: Audio, Idioma: inglés.** Se ha elegido la modalidad de audio por el usuario o el modelo. Los aspectos generales son:

Modelo:

- RoBERTa-GPT2
- SeamlessM4T
- CM-BERT

Procedimiento:

- Mensaje textual del usuario.
- El texto es recibido por el codificador RoBERTa para su procesamiento, prediciendo la emoción presente.
- La salida anterior será la entrada para el decodificador GPT-2, que genera una respuesta empática textual.
- Análogamente, CM-BERT analiza el contenido del mensaje retornando la intensidad de sentimiento presente, el cual ha superado el umbral establecido o, por el contrario, la longitud de respuesta ha sido mayor al mismo.
- El carácter aleatorio de la respuesta cambia a la modalidad de audio.
- La salida textual es recibida por SeamlessM4T para aplicar una traducción texto-a-voz (t2st), retornando una respuesta en audio.

4. **Entada: Texto, Salida: Audio, Idioma: Otro.** Contrario al caso anterior, la comunicación con el modelo será en una lengua que difiere con la inglesa, así:

Modelo:

- RoBERTa-GPT2
- SeamlessM4T
- CM-BERT

Procedimiento:

- Mensaje textual del usuario en un idioma distinto al inglés.
- El texto es recibido y procesado por la API SeamlessM4T, otorgando una salida en idioma inglés gracias a su traducción texto-a-texto (t2tt).
- El nuevo mensaje es recibido por el codificador RoBERTa para su procesamiento, prediciendo la emoción presente.

- La salida anterior será la entrada para el decodificador GPT-2, que genera una respuesta empática textual.
- Análogamente, CM-BERT analiza el contenido del mensaje retornando la intensidad de sentimiento presente, el cual ha superado el umbral establecido o, por el contrario, la longitud de respuesta ha sido mayor al mismo.
- El carácter aleatorio de la respuesta cambia a la modalidad de audio
- La salida textual es recibida por SeamlessM4T para aplicar una traducción texto-a-voz (t2st), retornando una respuesta en audio y en el idioma original.

5. **Entada: Audio, Salida: Texto, Idioma: inglés.** Este caso se caracteriza por los siguientes aspectos.

Modelo:

- RoBERTa-GPT2
- SeamlessM4T

Procedimiento:

- Mensaje de voz del usuario.
- El audio es recibido y procesado por la API SeamlessM4T, otorgando una salida textual gracias a su traducción voz-a-texto (s2tt).
- El nuevo mensaje es recibido por el codificador RoBERTa para su procesamiento, prediciendo la emoción presente.
- La salida anterior será la entrada para el decodificador GPT-2, que genera una respuesta empática textual.

6. **Entada: Audio, Salida: Texto, Idioma: Otro.** El idioma del sistema es configurado por el usuario previamente, siendo distinto al inglés. Aquí hace presencia la API de SeamlessM4T, así.

Modelo:

- RoBERTa-GPT2.
- SeamlessM4T

Procedimiento:

- Mensaje de voz del usuario en un idioma distinto al inglés.
- El audio es recibido y procesado por la API SeamlessM4T, otorgando una salida textual en idioma inglés gracias a su traducción voz-a-texto (s2tt).
- El nuevo mensaje es recibido por el codificador RoBERTa para su procesamiento, prediciendo la emoción presente.
- La salida anterior será la entrada para el decodificador GPT-2, que genera una respuesta empática textual.
- La respuesta se traduce nuevamente al idioma original con la API SeamlessM4T.

7. **Entada: Audio, Salida: Audio, Idioma: inglés.** La particularidad aleatoria hace presencia eligiendo la modalidad de audio o, en su defecto, ha sido configurada por el usuario como modalidad por defecto. Los aspectos generales son:

Modelo:

- RoBERTa-GPT2

- SeamlessM4T
- CM-BERT

Procedimiento:

- Mensaje de voz del usuario.
- El audio es recibido y procesado por la API SeamlessM4T, otorgando una salida textual gracias a su traducción voz-a-texto (s2tt).
- El nuevo mensaje es recibido por el codificador RoBERTa para su procesamiento, prediciendo la emoción presente.
- La salida anterior será la entrada para el decodificador GPT-2, que genera una respuesta empática textual.
- Análogamente, CM-BERT analiza el contenido del mensaje retornando la intensidad de sentimiento presente, el cual ha superado el umbral establecido o, por el contrario, la longitud de respuesta ha sido mayor al mismo.
- El carácter aleatorio de la respuesta cambia a la modalidad de audio.
- La salida textual es recibida por SeamlessM4T para aplicar una traducción texto-a-voz (t2st), retornando una respuesta en audio.

8. **Entada: Audio, Salida: Audio, Idioma: Otro.** Contrario al caso anterior, la comunicación con modelo será en una lengua que difiere con la inglesa, así:

Modelo:

- RoBERTa-GPT2
- SeamlessM4T
- CM-BERT

Procedimiento:

- Mensaje de voz del usuario en un idioma distinto al inglés.
- El audio es recibido y procesado por la API SeamlessM4T, otorgando una salida textual en idioma inglés gracias a su traducción voz-a-texto (s2tt).
- El nuevo mensaje es recibido por el codificador RoBERTa para su procesamiento, prediciendo la emoción presente.
- La salida anterior será la entrada para el decodificador GPT-2, que genera una respuesta empática textual.
- Análogamente, CM-BERT analiza el contenido del mensaje retornando la intensidad de sentimiento presente, el cual ha superado el umbral establecido o, por el contrario, la longitud de respuesta ha sido mayor al mismo.
- El carácter aleatorio de la respuesta cambia a la modalidad de audio.
- La salida textual es recibida por SeamlessM4T para aplicar una traducción texto-a-voz (t2st), retornando una respuesta en audio y en el idioma original.

Implementación

Para implementar el sistema completo que compone el proyecto práctico, se propone desarrollar una aplicación web que, en etapas próximas se plantea adoptar el formato de aplicación móvil para mayor comodidad y escalabilidad con las exigencias de uso.

Considerando lo anterior, se ha pensado en el nombre e ícono que representarán el sistema conversacional en cuestión; en este sentido, para cumplir con la característica global del sistema generador abierto a la multiplicidad de usuarios, la elección del nombre consideró algunos aspectos claves: breve, fácil de recordar y transmitir la idea de un sistema que ofrece apoyo emocional basado en la inteligencia artificial y la comprensión empática.

Nombre – EmpAI

El nombre "EmpAI" satisface las condiciones mencionadas, siendo una opción adecuada y relevante para un sistema conversacional destinado a proporcionar apoyo emocional personalizado, ya que refleja la naturaleza del proyecto: "Emp" proviene de "Empathy", lo que sugiere la capacidad del sistema para mostrar empatía hacia los usuarios. Esto refleja el propósito su principal, que es proporcionar apoyo psicológico personalizado y comprensión emocional, además de la inclusión de "AI", que resalta la base tecnológica del proyecto. Indica claramente que se trata de un sistema impulsado por inteligencia artificial, lo que sugiere innovación y avance tecnológico en el campo de la salud mental y el bienestar emocional.

Por otra parte, el nombre "EmpAI" es conciso y fácil de recordar, posee una identidad clara y memorable. Captura la esencia del proyecto en términos simples y directos, lo que lo hace fácilmente identificable y memorable para los usuarios y otros interesados.

Ícono

Figura 27. Logo del sistema implementado

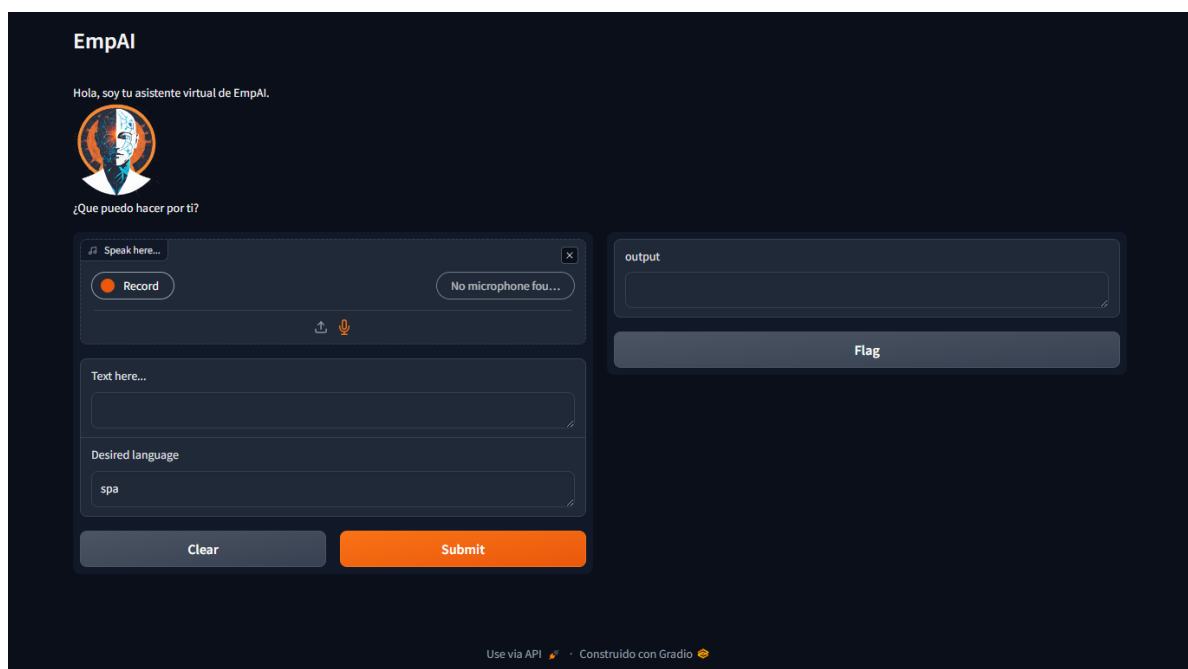


El ícono de EmpAI representa la fusión entre la tecnología y la emocionalidad, reflejando la naturaleza del sistema conversacional impulsado por inteligencia artificial. El foco principal es un personaje con una apariencia similar a un hombre robot, en cuya mitad derecha se muestra su parte exterior, simbolizando la inteligencia artificial y la tecnología avanzada que caracteriza a EmpAI. Esta parte exterior está representada en tonos blancos y celestes, denotando limpieza, precisión y modernidad.

En contraste, en la mitad izquierda del ícono, se revela la parte interna del personaje, que representa su alma o espíritu. Esta parte interna se visualiza en tonos cálidos y vibrantes, como el azul celeste, evocando emociones positivas y espirituales, así como tonos oscuros evocando la profundidad que caracteriza a las emociones y sentimientos. Esta representación simboliza la empatía, la comprensión emocional y la conexión humana que EmpAI busca proporcionar a los usuarios.

El fondo del ícono está adornado con un aura circular que rodea al personaje, simulando una luminosidad espiritual. Esta aura resalta la naturaleza compasiva y acogedora de EmpAI, transmitiendo una sensación de calidez y seguridad. La combinación de la figura del personaje, su parte interna y externa, junto con el aura circular, crea una imagen equilibrada y armoniosa que refleja la misión y la identidad de EmpAI como un sistema de apoyo psicológico empático y tecnológicamente avanzado.

Figura 28. Interfaz de la App EmpAI



8. Resultados preliminares

En esta sección, se presentan los resultados preliminares del rendimiento de cada modelo implementado en el sistema. Se analiza el desempeño de los modelos en relación a su entrenamiento y evaluación, utilizando métricas como pérdidas y perplejidad – expresada como el exponencial de las pérdidas de evaluación, de acuerdo con (Hugging Face).

$$e^{\ell}, \quad \ell: \text{eval loss}$$

Estos resultados proporcionarán una visión integral del éxito de cada modelo en la tarea asignada, permitiendo una evaluación detallada de su eficacia y robustez.

Además de las métricas cuantitativas, se mostrarán algunos ejemplos del output proporcionado por el sistema completo después de introducirles un input. Estos ejemplos ilustrarán cómo los modelos aplican el conocimiento adquirido durante el entrenamiento para generar respuestas empáticas y su capacidad de asesoramiento psicológico, analizar sentimientos multimodales y realizar traducciones multilingüaje. Esta visualización de ejemplos permitirá una comprensión más intuitiva del funcionamiento de los modelos y su capacidad para ofrecer apoyo psicológico personalizado y efectivo.

Con esta información detallada, se podrá realizar una evaluación exhaustiva del sistema desarrollado, identificando áreas de mejora y refinamiento para lograr un desempeño óptimo en la tarea de Generación de Contenido Autoayuda y Apoyo Psicológico Personalizado. Los resultados serán mostrados por cada modelo que compone cada arquitectura encoder-decoder desarrollada y el modelo CM-BERT.

RoBERTa-GPT2

Como se indicó en el apartado 7, se creó una arquitectura encoder-decoder inspirada en el modelo RoBERTa-GPT2 de (Liu, Maier, Minker, & Ultes, 2021) a partir del seguimiento de 4 estrategias, cada una de las cuales utiliza al menos uno de los siguientes modelos: RoBERTa, MentalRoberta, GPT-2 y DialoGPT; adico. Desde este enfoque, los mencionados fueron entrenados empleando el dataset EmpatheticCounseling y el optimizador Adam, con learning rate scheduler de tipo lineal, épsilon de $1e-8$, fp16 ajustado a True y distintos valores de hiperparámetros, partiendo de los especificados en su respectivo artículo y cuyos resultados se exponen a continuación.

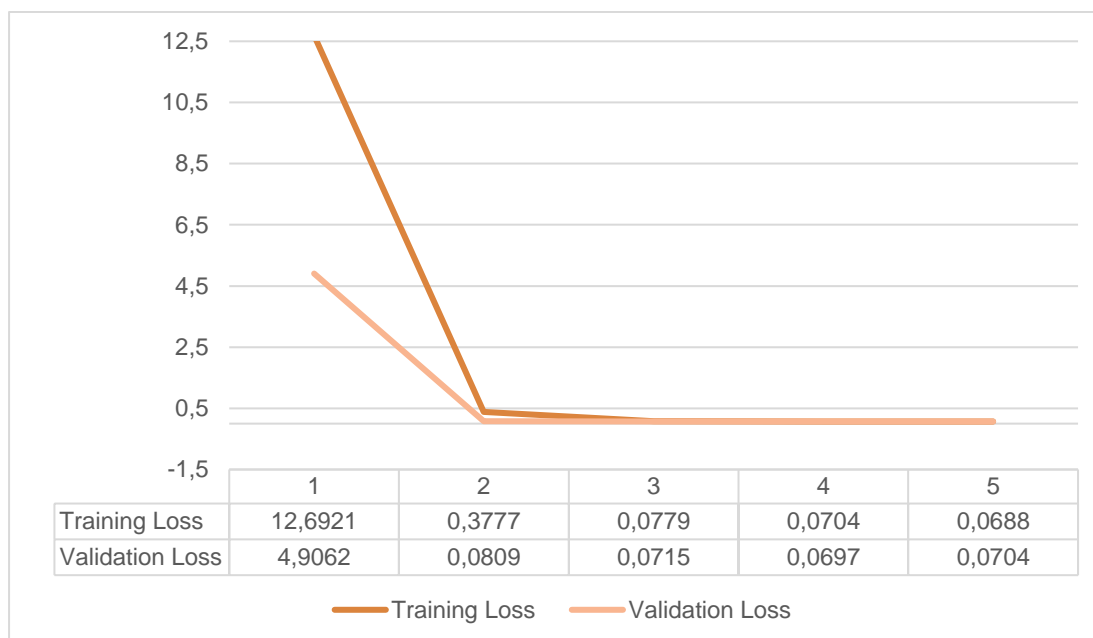
RoBERTa – Roberta_EmpAI

Los modelos RoBERTa y su versión MentalRoBERTa, fueron entrenados en base a los hiperparámetros de la propuesta original; sin embargo, se culminó con algunos cambios, además de ajustarse otros hiperparámetros durante su entrenamiento. Los valores de hiperparámetros que tuvieron mejores resultados fueron los siguientes.

Hiperparámetro	Valor
Learning rate	1e-4
train batch size	8
eval batch size	8
Gradient accumulation steps	8
Lr scheduler warmup steps	10000
Épocas	5
weight_decay	0.1

Para el número de épocas establecido, se realizaron 3750 pasos (steps) en total. Los resultados se muestran en la gráfica a continuación.

Figura 29. Resultados de entrenamiento – RoBERTa_EmpAI



El modelo RoBERTa_EmpAI finaliza con una pérdida de entrenamiento de 0,0688 y pérdida de validación de 0,0704, para lo cual, la perplejidad equivalente es 1.07.

MentalRoBERTa – MentalRoBERTa_EmpAI

A diferencia de RoBERTa, el entrenamiento de MentalRoBERTa se realizó durante 5000 pasos/steps (6.67 épocas), aumentando el batch size a 16 tanto de entrenamiento, como de evaluación.

Hiperparámetro	valor
Learning rate	1e-4
train batch size	16
eval batch size	16
Gradient accumulation steps	8
Lr scheduler warmup steps	10000
Pasos/steps	5000
weight_decay	0.1

Una vez culminados los pasos de entrenamiento fijados, se alcanzaron los resultados de la siguiente gráfica.

Figura 30. Resultados de entrenamiento – MentalRoBERTa_EmpAI



El modelo MentalRoBERTa_EmpAI finaliza con una pérdida de entrenamiento de 0,0632 y pérdida de validación de 0,0701, equivalente a una perplejidad de 1.07.

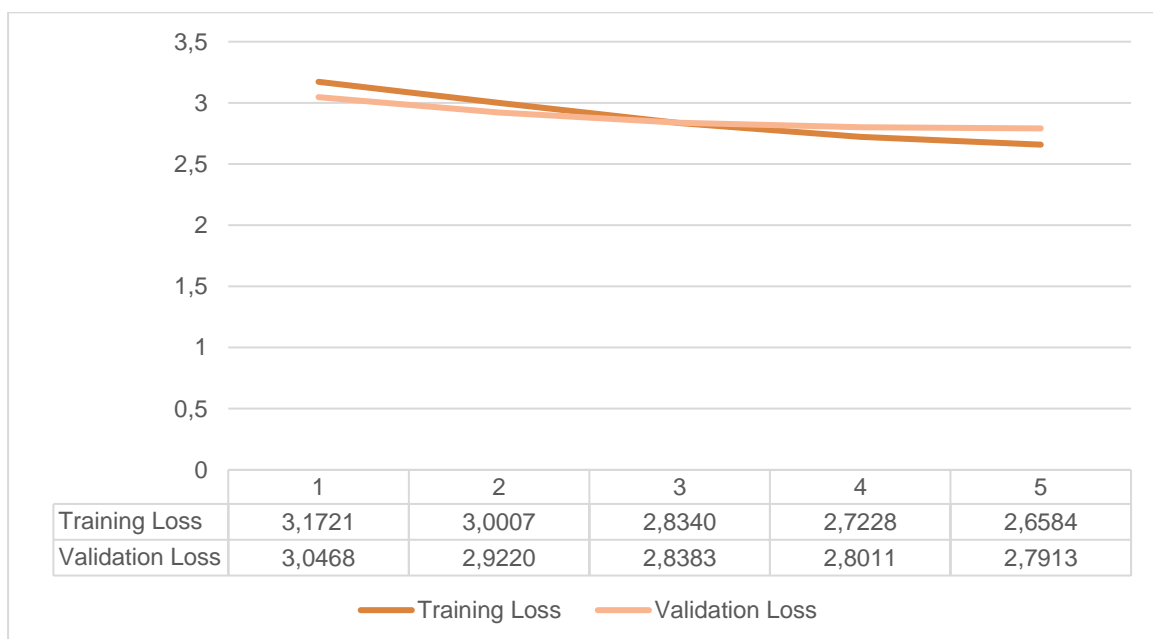
GPT-2

Los hiperparámetros para los modelos basados en GPT-2 se eligieron según los establecidos por los autores de DialoGPT, especialmente para el learning rate, gradient accumulation steps y batch size (train y eval). A pesar de múltiples variaciones, las pérdidas no alcanzaron valores menores a 2.5, generando sobreajuste desde la cuarta época. Se intentó mejorar los resultados manteniendo los hiperparámetros casi constantes durante 3 épocas adicionales a partir del modelo guardado, pero esto incrementó el sobreajuste sin una mejora significativa. Finalmente, los mejores resultados se lograron con los siguientes hiperparámetros.

Hiperparámetro	valor
Learning rate	1e-5
train batch size	4
eval batch size	4
Gradient accumulation steps	2
Lr scheduler warmup steps	16000
Épocas	5
weight_decay	1e-3

Una vez transcurridos los 30000 pasos (steps) que componen las 5 épocas de entrenamiento, los resultados obtenidos son los siguientes.

Figura 31. Resultados de entrenamiento - GPT-2_EmpAI



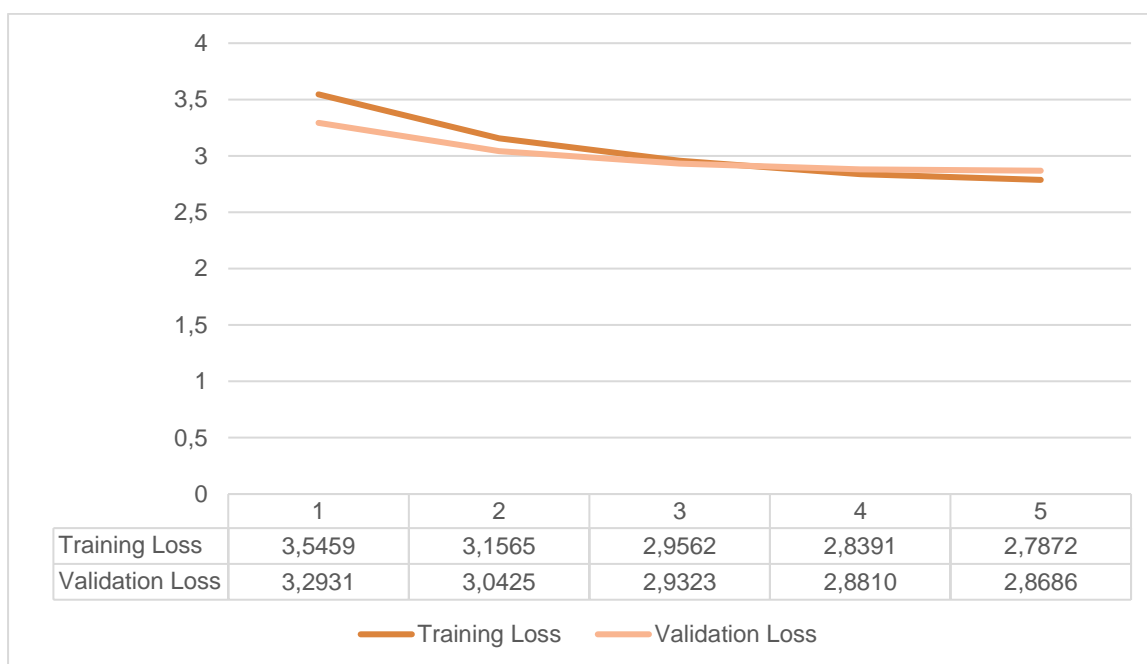
Tras tres intentos entrenando por 3 épocas más (18000 steps), los mejores resultados logrados se muestran en la tabla, con `weight_decay=1e-2` y `warmup_steps=8000`.

Época	Training Loss	Validation Loss
1	2.6015	2.7805
2	2.5131	2.7572
3	2.3998	2.7543

DialoGPT – DialoGPT_EmpAI

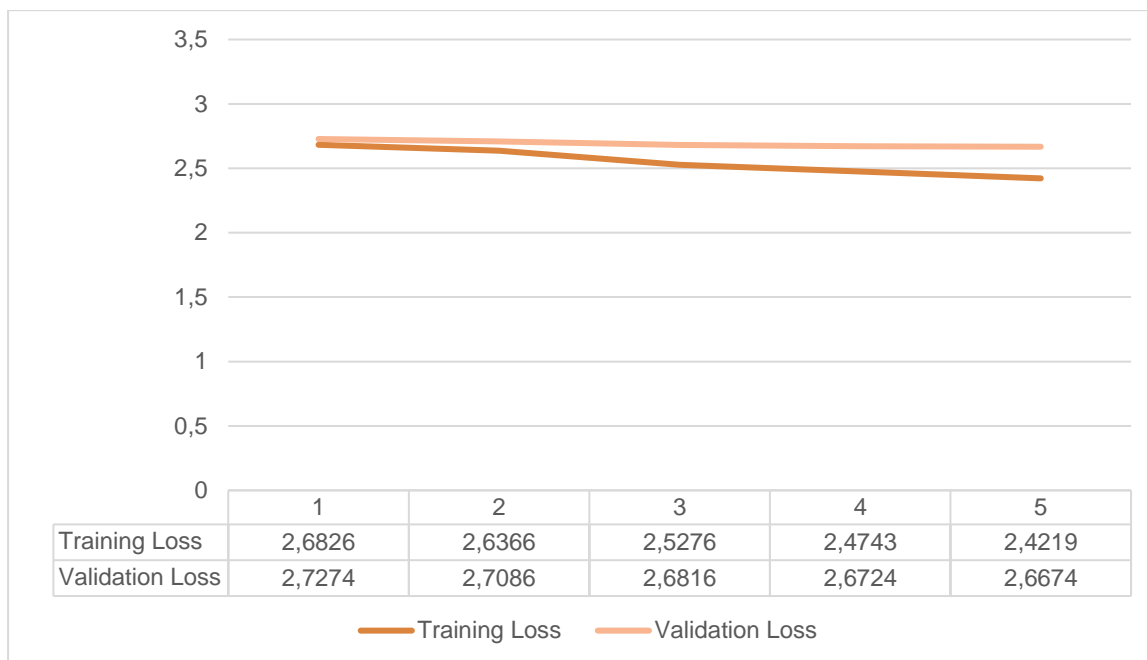
A diferencia de GPT-2, DialoGPT fue entrenado inicialmente durante 5 épocas, seguido de otras 5 épocas con el modelo que se guardó tras las primeras (DialoGPT_EmpAI). En este caso, ambos DialoGPT_EmpAI y su versión fine-tuned, conservaron los mismos valores de hiperparámetros que GPT2_EmpAI. Así, los resultados de las primeras 5 épocas se muestran en la gráfica.

Figura 32. Resultados de entrenamiento - DialoGPT_EmpAI



Las 5 épocas entrenadas sobre DialoGPT_EmpAI alcanzaron los siguientes resultados.

Figura 33. Resultados de entrenamiento - DialoGPT_EmpAI_FineTuned



El Fine-tuning del modelo GPT-2 para el proyecto EmpAI finaliza con una pérdida de entrenamiento de 2.6584 y pérdida de validación de 2.7913, para lo cual, la perplejidad equivalente es 16.30. Así mismo, el modelo DialoGPT fine-tuned finaliza con una pérdida de entrenamiento de 2.4219 y pérdida de validación de 2.6674, equivalente a una perplejidad de 14.40. Este último, particularmente, tiene gran similitud con la medida de perplejidad alcanzada por los autores de RoBERTa-GPT2 con su obra – 14.97.

Con los modelos ya entrenados, se procedió a crear la arquitectura, cuya estructura se encuentra en el Anexo II. A causa de los pésimos resultados obtenidos en el entrenamiento de los modelos GPT2 y DialoGPT, las respuestas otorgadas por el modelo encoder-decoder de cada estrategia tuvieron la misma calidad en inferencia, careciendo de coherencia con respecto al input del usuario; una evidencia de esto se muestra en el ejemplo con la estrategia 3, para una longitud máxima de 75 nuevos tokens.

Tabla 21. Ejemplo de inferencia con la estrategia 3

Input	Yesterday I had an argument with my father and I feel bad about it, what should I do?
Respuesta	!! I feel very comfortable in my new life and have been doing well in my work. I am very happy. I feel very comfortable with my new life. I feel like I have a great relationship with my new wife! I am so happy I have such a nice wife! How's your wife doing these days?

A pesar de tener sentido, la respuesta no guarda una relación de concordancia con el mensaje introducido por el usuario, impidiendo que exista una comunicación empática y de apoyo psicológico. Se trató de mejorar este comportamiento, modificando la estructura construida, volviendo a entrenar los modelos utilizando únicamente el dataset EmpatheticDialogues completo y con el filtro de la cantidad especificada al momento de crear a EmpathetiCounseling; sin embargo, no hubo un cambio acertado.

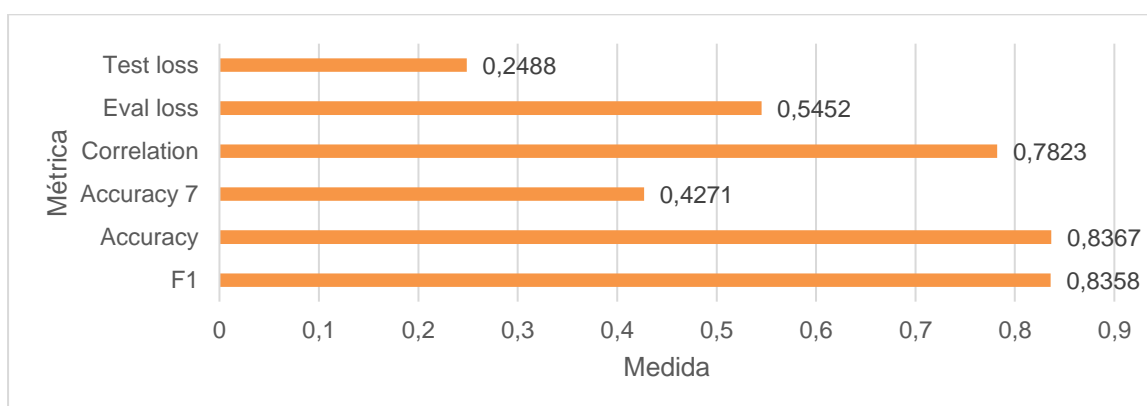
CM-BERT

Aunque la generación de respuesta empática es la principal tarea del proyecto, todavía falta implementar las dos tareas restantes; en lo que respecta a una de ellas, el entrenamiento del modelo Cross-Modal-BERT (CM_BERT) se realizó con la siguiente configuración.

Parámetro	valor	Parámetro	valor
attention_probs_dropout_prob	0.1	initializer_range	0.02
audio_dense_size	5	intermediate_size	3072
audio_feature_size	1	max_position_embeddings	512
audio_size	10000	num_attention_heads	12
hidden_act	"gelu"	num_hidden_layers	12
hidden_dropout_prob	0.1	num_prehidden_layers	11
hidden_size	768	symbol_size	2
hidden_size1	50	type_vocab_size	2
		vocab_size	30522

Con ello, el mismo culminó luego de 162 pasos/steps con los siguientes resultados, evidenciados en el Anexo I.

Figura 34. Resultados de evaluación - CM-BERT



Propuestas de mejora

Se ha planteado la posibilidad de continuar en desarrollo el proyecto incorporando mejoras que pueden contribuir significativamente a la eficacia y la calidad del sistema de apoyo psicológico personalizado, permitiendo así brindar un servicio más completo y holístico, adaptado a las necesidades cambiantes de los usuarios en el ámbito de la salud mental. Conjuntamente con la mejora en la precisión del modelo y la calidad de las respuestas en términos de empatía y coherencia con la entrada, algunos aspectos clave considerados para ser verdaderamente eficiente se analizan a continuación.

1. **Adaptación al perfil del usuario.** Utilizar datos recopilados de manera ética y anónima para personalizar aún más las recomendaciones y estrategias de apoyo para cada usuario, a partir de la implementación de otros algoritmos de aprendizaje automático capaces de evaluar y comprender de manera más precisa sus necesidades, preferencias y características individuales. Esto podría incluir la integración de cuestionarios interactivos y la recopilación de datos contextualizados sobre la actividad del usuario en la plataforma.
2. **Fortalecer la seguridad y confidencialidad.** Implementar medidas de seguridad adicionales, como encriptación de extremo a extremo, autenticación de dos factores y políticas de privacidad claras y transparentes. Garantizar el cumplimiento de las regulaciones de protección de datos, como el Reglamento General de Protección de Datos (GDPR) en la Unión Europea (2016).
3. **Fomentar una mayor colaboración con profesionales.** Establecer alianzas estratégicas con instituciones educativas, clínicas de salud mental y organizaciones sin fines de lucro para ofrecer una gama más amplia de servicios psicológicos y programas de intervención. Facilitar la derivación de usuarios a terapeutas y psicólogos certificados cuando sea necesario.
4. **Implementar una evaluación y seguimiento más exhaustivos.** Desarrollar herramientas analíticas avanzadas para evaluar el impacto a largo plazo del sistema en la salud mental de los usuarios. Realizar estudios de seguimiento y análisis de resultados para identificar áreas de mejora y ajustar las intervenciones de manera proactiva.
5. **Potenciar el enfoque preventivo.** Ofrecer recursos educativos y programas de autocuidado destinados a fortalecer la resiliencia emocional y promover el bienestar psicológico a largo plazo. Incorporar técnicas de mindfulness, ejercicios de relajación y actividades de desarrollo personal en la plataforma para empoderar a los usuarios en su proceso de autodescubrimiento y crecimiento emocional.

9. Conclusiones

Después de completar este trabajo final de máster, se han identificado tanto logros como áreas de mejora significativas en el desarrollo de un sistema conversacional basado en inteligencia artificial para proporcionar apoyo psicológico personalizado. Se destaca la importancia de abordar la necesidad creciente de apoyo psicológico en la sociedad contemporánea, reconociendo el potencial de la inteligencia artificial para brindar soluciones innovadoras en este ámbito.

Se han explorado y aplicado diversas técnicas de Procesamiento de Lenguaje Natural y modelos de aprendizaje automático para mejorar la comprensión del contexto emocional del usuario y generar respuestas empáticas. Sin embargo, los resultados obtenidos revelaron limitaciones significativas en la coherencia y relevancia de las respuestas generadas por el modelo desarrollado.

A pesar de estas limitaciones, se reconoce la importancia de continuar investigando y refinando los modelos de inteligencia artificial para mejorar su capacidad de ofrecer apoyo psicológico personalizado de manera efectiva. Se plantea la idea de continuar desarrollando el proyecto, enfocándose en la optimización de algoritmos y en la integración de nuevas técnicas para mejorar la calidad de las respuestas generadas, así como implementar nuevas funcionalidades que logren satisfacer la demanda de usuario en términos de salud mental.

En resumen, este trabajo proporciona una base sólida para futuras investigaciones en el campo de la inteligencia artificial aplicada al apoyo psicológico, destacando la necesidad de seguir avanzando en la búsqueda de soluciones innovadoras que contribuyan al bienestar emocional de las personas.

Bibliografía

- Amod. (n.d.). *mental_health_counseling_conversations* (Revision 9015341). Hugging Face. doi:10.57967/hf/1581
- Bertagnolli, N., Lord, G., Lee, P., & Ström, E. (2020). Counsel Chat: Bootstrapping High-Quality Therapy Data. *Towards Data Science*.
- Cambridge University Press. (2022). *The Science of Deep Learning*. New York: Iddo Drori.
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Google AI Language*.
- Directiva 95/46/CE del Parlamento Europeo y del Consejo. (2016). Reglamento General de Protección de Datos (GDPR). *Diario Oficial de la Unión Europea*, 1–88.
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.
- Google Research. (n.d.). *Artificial intelligence (AI) vs. machine learning (ML)*. Retrieved Diciembre 20, 2023, from Google Cloud: <https://cloud.google.com/learn/artificial-intelligence-vs-machine-learning>
- Google Research. (n.d.). *What is Artificial Intelligence (AI)?* Retrieved Diciembre 20, 2023, from Google Cloud: <https://cloud.google.com/learn/what-is-artificial-intelligence>
- Hugging Face. (n.d.). *Causal language modeling*. Retrieved Marzo 30, 2024, from Hugging Face: https://huggingface.co/docs/transformers/tasks/language_modeling
- Hugging Face. (n.d.). *Data Collator*. Retrieved Marzo 30, 2024, from Hugging Face: https://huggingface.co/docs/transformers/main/en/main_classes/data_collator
- Hugging Face. (n.d.). *Fine-tuning a masked language model*. Retrieved Marzo 30, 2024, from Hugging Face: <https://huggingface.co/learn/nlp-course/chapter7/3?fw=pt>
- Hugging Face. (n.d.). *Masked language modeling*. Retrieved Marzo 30, 2024, from Hugging Face: https://huggingface.co/docs/transformers/tasks/masked_language_modeling
- Hugging Face. (n.d.). *Normalización y pre-tokenización*. Retrieved Marzo 30, 2024, from Hugging Face: <https://huggingface.co/learn/nlp-course/es/chapter6/4?fw=pt>
- Ji, S., Zhang, T., Ansari, L., Fu, J., Tiwari, P., & Cambria, E. (2021). MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare. *LREC*.
- June Min, D., Pérez-Rosas, V., Resnicow, K., & Mihalcea, R. (2022). PAIR: Prompt-Aware margin Ranking for Counselor Reflection Scoring in Motivational Interviewing. *EMNLP*.
- Kim, H., Kim, B., & Kim, G. (2021). Perspective-taking and Pragmatics for Generating Empathetic Responses Focused on Emotion Causes. *Seoul National University*.
- Kingma, D., & Welling, M. (2019). An Introduction to Variational Autoencoders. *Foundations and Trends in Machine Learning*, 12(4), 307-392.
- Lin, Z., Wang, L., Li, J., Meng, F., Yang, C., Wang, W., & Zhou, J. (2022). Empathetic Dialogue Generation via Sensitive Emotion Recognition and Sensible Knowledge Selection. *WeChat AI, Tencent Inc.*
- Liu, J. M., Li, D., Cao, H., Ren, T., Liao, Z., & Wu, J. (2023). ChatCounselor: A Large Language Models for Mental Health Support. *PGAI CIKM*.
- Liu, S., Zheng, C., Demasi, O., Sabour, S., Li, Y., Yu, Z., . . . Huang, M. (2021). Towards Emotional Support Dialog Systems. *Tsinghua University*.
- Liu, Y., Maier, W., Minker, W., & Ultes, S. (2021). Empathetic Dialogue Generation with Pre-trained RoBERTa-GPT2 and External Knowledge. *Mercedes-Benz AG*.
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., . . . Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *Facebook AI*.

- Ma, X., Pino, J., & Koehn, P. (2020). SimulMT to SimulST: Adapting Simultaneous Text Translation to End-to-End Simultaneous Speech Translation. *Facebook AI*.
- Mehrabian, A., & Ferris, S. R. (1967). Inference of attitudes from nonverbal communication in two channels. *Journal of Consulting Psychology*, 31(3), 248–252.
doi:10.1037/h0024648
- Meta AI, INRIA, UC Berkeley. (2023). SeamlessM4T: Massively Multilingual & Multimodal Machine Translation. *Meta AI*.
- Miao, Y., Yu, L., & Blunsom, P. (2016). Neural Variational Inference for Text Processing. *International Conference on Machine Learning*, 48.
- Mondragón, R. S. (2017). Reflexión: La empatía en la relación «médico-paciente». Una ruta esencial obligada. *Sanid Milit*, 71(6), 503-506.
- Peng, S., Lu, J., Majumder, N., Hong, P., Ghosal, D., Gelbukh, A., . . . Poria, S. (2020). MIME: MIMicking Emotions for Empathetic Response Generation. *Singapore University of Technology and Design*.
- Rashkin, H., Smith, E., Li, M., & Boureau, Y.-L. (2019). Towards Empathetic Open-domain Conversation Models: a New Benchmark and Dataset. *Facebook AI Research*.
- Ren, Y., Liu, J., & Zhao, Z. (2022). PortaSpeech: Portable and High-Quality Generative Text-to-Speech. *Neural Information Processing Systems*.
- Ren, Z., Li, Q., Chen, H., Ren, P., Tu, Z., & Chen, Z. (2020). EmpDG: Multi-resolution Interactive Empathetic Dialogue Generation. *International Conference on Computational Linguistics*, 4454–4466.
- Rojas Estapé, M. (2018). *Cómo hacer que te pasen cosas buenas*. Barcelona: Espasa Libros.
- Sabour, S., Zheng, C., & Huang, M. (2022). CEM: Commonsense-aware Empathetic Response Generation. *Association for the Advancement of Artificial*.
- Schütz Balistieri, A., & Mara de Melo Tavares, C. (2013). A importância do apoio sócio-emocional em adolescentes e adultos jovens portadores de doença crônica: uma revisão de literatura. *Enfermería Global*, XII(30), 388-396.
- Siriwardhana, S., Reis, A., Weerasekera, R., & Nanayakkara, S. (2020). Jointly Fine-Tuning “BERT-like” Self Supervised Models to Improve Multimodal Speech Emotion Recognition. *The University of Auckland*.
- Tan, X., Chen, J., Liu, H., Cong, J., Zhang, C., Liu, Y., . . . Liu, T.-Y. (2022). NaturalSpeech: End-to-End Text to Speech Synthesis with Human-Level Quality. *Microsoft Research*.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., & Kaiser, Ł. (2017). Attention Is All You Need. *Neural Information Processing Systems*.
- Yang, K., Xu, H., & Gao, K. (2020). CM-BERT: Cross-Modal BERT for Text-Audio Sentiment Analysis. *Association for Computing Machinery*.
- Yu, T., Gao, H., Lin, T.-E., Yang, M., Wu, Y., Ma, W., . . . Li, Y. (2023). Speech-text dialog Pre-training for spoken dialog understanding with ExpliCiT cRoss-Modal Alignment. *Alibaba Group*.
- Zadeh, A., Liang, P. P., Poria, S., Vij, P., Cambria, E., & Morency, L.-P. (2018). Multi-attention Recurrent Network for Human Communication Comprehension. *Association for the Advancement of Artificial Intelligence*.
- Zadeh, A., Liang, P., Vanbriesen, J., Poria, S., Tong, E., Cambria, E., . . . Morency, L.-P. (2018). Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph. *Association for Computational Linguistics*, 2236-2246.

- Zadeh, A., Zellers, R., Pincus, E., & Morency, L.-P. (2017). MOSI: Multimodal Corpus of Sentiment Intensity and Subjectivity Analysis in Online Opinion Videos. *Association for Computational Linguistics*.
- Zech, E., & Rimé, B. (2005). Is Talking About an Emotional Experience Helpful? Effects on Emotional Recovery and Perceived Benefits. *Clinical Psychology and Psychotherapy*(12), 270-287.
- Zhang, Y., Sun, S., Galley, M., Chen, Y.-C., Brockett, C., Gao, X., . . . Dolan, B. (2020). DIALOGPT : Large-Scale Generative Pre-training for Conversational Response Generation. *Microsoft Research*.
- Zhou, J., Zheng, C., Wang, B., Zhang, Z., & Huang, M. (2023). CASE: Aligning Coarse-to-Fine Cognition and Affection for Empathetic Response Generation. *The CoAI Group*.

10. Anexos

Anexo I

```
Cross-Modal-BERT-master/pre-trained BERT
04/11/2024 03:37:55 - INFO - model - Model config {
  "attention_probs_dropout_prob": 0.1,
  "audio_dense_size": 5,
  "audio_feature_size": 1,
  "audio_size": 10000,
  "hidden_act": "gelu",
  "hidden_dropout_prob": 0.1,
  "hidden_size": 768,
  "hidden_size1": 50,
  "initializer_range": 0.02,
  "intermediate_size": 3072,
  "max_position_embeddings": 512,
  "num_attention_heads": 12,
  "num_hidden_layers": 12,
  "num_prehidden_layers": 11,
  "symbol_size": 2,
  "type_vocab_size": 2,
  "vocab_size": 30522
}
```

```
Evaluating: 100%|██████████| 29/29 [04:11<00:00, 8.68s/it]
04/15/2024 05:57:14 - INFO - utils - ***** test results *****
04/15/2024 05:57:14 - INFO - utils - F1 = 0.8421441253982106
04/15/2024 05:57:14 - INFO - utils - acc = 0.8425655976676385
04/15/2024 05:57:14 - INFO - utils - acc7 = 0.44752186588921283
04/15/2024 05:57:14 - INFO - utils - corr = 0.7871757067734716
04/15/2024 05:57:14 - INFO - utils - global_step = 162
04/15/2024 05:57:14 - INFO - utils - loss = 0.12527923672287553
04/15/2024 05:57:14 - INFO - utils - mae = 0.7401033
04/15/2024 05:57:14 - INFO - utils - test_loss = 0.499398461189763
3
rm: cannot remove 'Cross-Modal-BERT-master/CM-BERT_output': No such fi
le or directory
acc: ['0.8192419825072886', '0.8279883381924198', '0.8454810495626822',
, '0.8104956268221575', '0.8425655976676385', 0.8291545189504375]
F1: ['0.8184314155232523', '0.8268989085539326', '0.8456669781383024',
'0.808877148194585', '0.8421441253982106', 0.8284037151616566]
mae: ['0.7553419', '0.7838328', '0.7362792', '0.803031', '0.7401033',
0.7637176400000001]
corr: ['0.7786051863739049', '0.7777040843093563', '0.7825609517778993
', '0.7729275355994695', '0.7871757067734716', 0.7797946929668204]
acc7: ['0.45043731778425655', '0.41545189504373176', '0.44023323615160
35', '0.4008746355685131', '0.44752186588921283', 0.43090379008746355]
```


Anexo II

```
Response: DialogueGenerator(
  (encoder): RobertaForMaskedLM(
    (roberta): RobertaModel(
      (embeddings): RobertaEmbeddings(
        (word_embeddings): Embedding(50265, 768, padding_idx=1)
        (position_embeddings): Embedding(514, 768, padding_idx=1)
        (token_type_embeddings): Embedding(1, 768)
        (LayerNorm): LayerNorm((768,), eps=1e-05, elementwise_affine=True)
        (dropout): Dropout(p=0.1, inplace=False)
      )
    (encoder): RobertaEncoder(
      (layer): ModuleList(
        (0-11): 12 x RobertaLayer(
          (attention): RobertaAttention(
            (self): RobertaSelfAttention(
              (query): Linear(in_features=768, out_features=768, bias=True)
              (key): Linear(in_features=768, out_features=768, bias=True)
              (value): Linear(in_features=768, out_features=768, bias=True)
              (dropout): Dropout(p=0.1, inplace=False)
            )
            (output): RobertaSelfOutput(
              (dense): Linear(in_features=768, out_features=768, bias=True)
              (LayerNorm): LayerNorm((768,), eps=1e-05, elementwise_affine=True)
              (dropout): Dropout(p=0.1, inplace=False)
            )
          )
          (intermediate): RobertaIntermediate(
            (dense): Linear(in_features=768, out_features=3072, bias=True)
            (intermediate_act_fn): GELUActivation()
          )
          (output): RobertaOutput(
            (dense): Linear(in_features=3072, out_features=768, bias=True)
            (LayerNorm): LayerNorm((768,), eps=1e-05, elementwise_affine=True)
            (dropout): Dropout(p=0.1, inplace=False)
          )
        )
      )
    )
  )
  (lm_head): RobertaLMHead(
    (dense): Linear(in_features=768, out_features=768, bias=True)
    (layer_norm): LayerNorm((768,), eps=1e-05, elementwise_affine=True)
    (decoder): Linear(in_features=768, out_features=50265, bias=True)
  )
)
```

```

)
(decoder): GPT2LMHeadModel(
  (transformer): GPT2Model(
    (wte): Embedding(50257, 768)
    (wpe): Embedding(1024, 768)
    (drop): Dropout(p=0.1, inplace=False)
    (h): ModuleList(
      (0-11): 12 x GPT2Block(
        (ln_1): LayerNorm((768,), eps=1e-05, elementwise_affine=True)
        (attn): GPT2Attention(
          (c_attn): Conv1D()
          (c_proj): Conv1D()
          (attn_dropout): Dropout(p=0.1, inplace=False)
          (resid_dropout): Dropout(p=0.1, inplace=False)
        )
        (ln_2): LayerNorm((768,), eps=1e-05, elementwise_affine=True)
        (mlp): GPT2MLP(
          (c_fc): Conv1D()
          (c_proj): Conv1D()
          (act): NewGELUActivation()
          (dropout): Dropout(p=0.1, inplace=False)
        )
      )
    )
    (ln_f): LayerNorm((768,), eps=1e-05, elementwise_affine=True)
  )
  (lm_head): Linear(in_features=768, out_features=50257, bias=False)
)
)

```

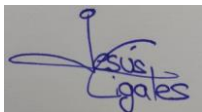

Documento-Compromiso de tutorización de trabajos fin de Máster- Máster Universitario en Inteligencia Artificial

D/Dña: _____JESÚS CIGALES CANGA_____, con DNI: _____32888657Z_____, se compromete por este escrito a asumir las tareas de tutorización necesarias para el desarrollo adecuado del Trabajo Fin de Máster de la edición Abril/2023 a los/las alumnos:

Luis Angel Motta Valero, con DNI: 1117549110

Para lo cual firma el presente documento.

Fdo.:

A handwritten signature in blue ink, appearing to read 'Jesús Cigales', is shown within a rectangular box.

En Oviedo, a 24 de noviembre del 2023