**Chatbot con Comprensión Emocional**



Carlos Pérez Simón

Facultad de Ciencias Estadísticas

Trabajo de Fin de Master en Big Data, Data Science e IA

Septiembre de 2024

Tutores:

**Abstract**

Esta investigación presenta un sistema basado en inteligencia artificial que utiliza Redes Neuronales Convolucionales (CNN) y Procesamiento de Lenguaje Natural (NLP) para detectar e interpretar emociones en tiempo real. Este sistema no solo identifica estados emocionales a partir de entradas multimodales, sino que tambi´en recomienda actividades personalizadas que pueden mejorar el bienestar emocional del usuario, abordando así desafíos sociales significativos como la soledad y la baja productividad en entornos laborales y personales en España.

# Capítulo 1

# Introducción

En el ámbito de la investigación psicológica y el desarrollo tecnológico, la relación entre el bienestar emocional y la productividad ha emergido como un área de enfoque crítica. Estudios recientes subrayan el impacto significativo que los estados emocionales tienen en el rendimiento individual y las interacciones sociales. Este problema es particularmente agudo en España, donde el Observatorio Estatal de la Soledad No Deseada y el Ministerio de Derechos han identificado la soledad no deseada como una preocupación social urgente que requiere soluciones efectivas.

La investigación en aprendizaje automático ha explorado varios modelos para detectar y analizar emociones humanas a través de diversas modalidades como texto, habla y expresiones faciales. Las Redes Neuronales Convolucionales (CNN), por ejemplo, han demostrado capacidades robustas en la clasificación de estados emocionales a partir de datos visuales. Sin embargo, la aplicación directa de estas tecnologías para abordar problemas como la procrastinación y el aislamiento social sigue siendo relativamente inexplorada.

Esta investigación busca cerrar esta brecha utilizando herramientas de código abierto para desarrollar un sistema basado en CNN capaz de detectar e interpretar emociones en tiempo real. El sistema emplea bibliotecas de procesamiento de lenguaje natural como NLTK y TextBlob para el reconocimiento de emociones basadas en texto, y OpenCV para el análisis de expresiones faciales. La integración de estas herramientas tiene como objetivo mejorar la productividad y mitigar la soledad, alineándose con las prioridades del Observatorio y el Ministerio.

La contribución de esta investigación es doble: extiende las metodologías existentes de aprendizaje profundo a la aplicación práctica en la lucha contra la soledad y la mejora de la productividad, y demuestra cómo las tecnologías de código abierto pueden ser efectivamente utilizadas para abordar desafíos sociales y psicológicos significativos. Este enfoque innovador establece una nueva dirección para futuras intervenciones tecnológicas en la salud pública y el bienestar social.

Después del marco teórico, se probarán los mejores modelos para cada modalidad (texto, audio y video). Luego, se unirán las características multimodales y se integrarán en un chatbot. El objetivo es que el chatbot sea capaz de entender las emociones del usuario y proporcionar respuestas adaptativas que ayuden a mejorar la gestión del tiempo y el bienestar emocional.

# Capítulo 2

# Marco Teórico y Trabajos Previos

## 2.1. Teoría de las Emociones

## 2.1.1. Modelo 6DE

El modelo 6DE, del inglés “Six Emotional Dimensions”, recoge diferentes aspectos que caracterizan las emociones humanas:

* **Arousal:** Mide el nivel de energía asociada a una emoción, desde un estado de calma hasta uno de excitación.
* **Valencia:** Indica la naturaleza positiva o negativa de una emoción.
* **Dominancia**: Mide el grado de control del sujeto sobre su emoción. Una dominancia alta sería mantener el control y baja sumisión a las emoción.
* **Agencia:** Mide el grado de voluntad de una emoción, desde involuntaria hasta controlada.
* **Fidelidad:** Mide la congruencia de la emoción con el contexto sucedido, desde desalineación hasta congruencia.
* **Novedad:** Mide la frecuencia de una emoción, desde emoción habitual hasta una nunca experimentada.

Esta caracterización de las emociones resulta muy útil a la hora de comprender cómo siente el sujeto realmente la emoción, facilitando la recomendación personalizada de actividades.

## 2.2. Detección Multimodal de Emociones

Un modelo multimodal contiene diferentes modelos en función del tipo de contenido que se quiera estudiar. Los más frecuentes en la detección de emociones son audio, video, speech, texto, gestos corporales y señales EGG. En este estudio nos centraremos en el contenido audio visual, veamos el trabajo previo en estos campos:

## 2.2.1. Detección de Emociones en Audio

La extracción de features en audio suele implicar la transformación de la señal al dominio de frecuencias mediante la Transformada de Fourier. Sin embargo existen diversas clases de features en esta área:

* Físicas (prosodic): El tono, la potencia de la señal, etc.
* Cualitativas: Calidad del audio, tensión, respiración, etc.
* Espectrales: Todas las del dominio de frecuencias, ya sea el espectrograma Log-Mel, los MFCC, LPC, LFPC, etc.

Normalmente, con idea de capturar la dimensión temporal de los datos de audio, los datos se organizan por ventanas deslizantes de Hamming, las cuales contienen segmentos del audio de una longitud y solapamiento previamente determinados. Además, para detección de emociones se utiliza un rango de frecuencias bajas.

A continuación, se presenta un resumen de diversos enfoques y resultados en el reconocimiento de emociones utilizando diferentes técnicas y modelos:

* **Modelos SVM y GMM:** Sezgin et al. (2012) introducían features físicas en modelos simples como SVM.
* **Redes Neuronales Recurrentes (RNN):** Yoon et al.(2018) presentaban un modelo MDREA multimodal, donde cada modo se procesa en una red recurrente ARE o TRE, y se aplica posteriormente un mecanismo de atención posteriormente.
* **Autoencoders Transformers**: Li-Wei Chen et al.(2023) iestudiaron la implementación de modelos transformers al reconocimiento de emociones, destacando los resultados del modelo Wav2vec 2.0 P-TAPT, el cual alcanzó 74% de accuracy sobre IEMOCAP.
* **Combinación de CNN y LSTM con atención:**

Cabe destacar el uso de DARTS como algoritmo de optimización de hiperparámetros. Debido a la complejidad y variabilidad de los datos de audio, DARTS se adapta a los datos usando gradientes, disminuyendo así el coste computacional. Además, DARTS también optimiza la extracción de features.

## 2.2.2. Detección de Emociones en Video

En la detección de emociones sobre video, lo habitual es que los datos que entran a los modelos sean imágenes extraídas de cada frame del video.

El pre-procesamiento de los vídeos consiste en centrar el modelo en las emociones faciales. Por tanto, en cada frame se aplican algoritmos de detección facial como Dlib o Viola-Jones, y se recorta como una imagen nueva el cuadrado que define la cara del usuario. Además, se transforman todas las imágenes a un formato y tamaño predefinidos, y se aplican técnicas de alineamiento facial, de forma que todas las caras del dataset tengan una estructura y posición similares. Para el ‘Facial Alignment’ existen diferentes técnicas, pero destaca por su simplicidad y eficacia el alineamiento de los ojos mediante la extracción de landmarks faciales en cada frame.

Más allá del uso de los frames, también existen modelos que utilizan el espectrograma Log-Mel y el espectrograma Speech para la extracción de features en cada frame.

Por otro lado, la investigación ha demostrado que el uso de modelos profundos y técnicas de fusión multimodal puede mejorar significativamente la precisión en la detección de emociones. A continuación, se presenta un resumen de diversos enfoques y resultados en el reconocimiento de emociones utilizando diferentes técnicas y modelos:

* **Modelos Convolucionales y Recurrentes (CNN + RNN):** Inicialmente se combinaban modelos CNN, para extraer features espaciales de los frames, con modelos que permitiesen acceder a patrones temporales de los datos, donde destacan las redes recurrentes, RNN, y las redes ‘deep belief’, DBN. Kahou et al. (2015) presentaron un sistema híbrido CNN-RNN para el análisis de expresiones faciales en video, que superó los métodos basados en promedios temporales, logrando un rendimiento destacado en la competencia EmotiW 2015​​. Posteriormente Zhou et al. (2019) exploraron la fusión de características y estrategias de atención para la combinación de características emocionales de audio y video, alcanzando una precisión del 62.48% en el conjunto de prueba de EmotiW 2019​​.
* **Modelos Multimodales y Transformers:** Luna-Jimenez et al. (2021) presentaron un modelo multimodal para audio y video utilizando Transformers, destacando el uso del modelo WavVec 2.0, lo cual empezaba a dar mejores resultados. Estos empezaron a mejorar añadiendo mecanismos de atención y la posibilidad de tratar únicamente un modo de datos por separado.
* **Modelo de Grafos Temporales:** Thi Nguyen et al. (2024) presentan un modelo que combina una red convolucional temporal (RT-GCN) para capturar el contexto local entre modalidades, y un mecanismo de interacción entre los modos (P-CM), que estudia el contexto global a nivel conversación. Este modelo, conocido como CORECT, alcanza una precisión del 69% sobre el dataset IEMOCAP.
* **SISTCM y bi-LSTM:** Jie Wei et al.(2024) emplea un modelo espacio-temporal Convolucional para estudiar el contexto temporal local y, posteriormente, un modelo bidireccional LSTM que relaciona el contexto temporal global. Este modelo resulta ser el más efectivo hasta el momento, entrenándose sobre IEMOCAP y CMU-MOSEI, y alcanzándose una precisión del 98% sobre el dataset CK+.

## 2.2.3. Fusión de Datos Multimodales

Los modelos multimodales consisten en el modelado de diferentes tipos de datos de manera separada y, posteriormente, la fusión de los resultados obtenidos en cada modo. En este estudio se emplearán datos de audio y video.

Existen varias técnicas para fusionar los vectores de características generados por cada modelo, las cuales se distinguen según el momento en que se realiza la fusión:

* **Fusión Intramodelo:** Se destaca el uso de DARTS como algoritmo de optimización de hiperparámetros. DARTS se adapta a los datos mediante gradientes, disminuyendo el coste computacional debido a la complejidad y variabilidad de los datos de audio. Además, DARTS se aplica tanto al modelo como a la extracción de características.
* **Fusión entre Modelos:** En la fusión entre modelos, también conocida como fusión cruzada, se combinan características provenientes de diferentes modalidades (por ejemplo, audio y video). Esta fusión se puede realizar de varias maneras:
  + Fusión Temprana: Las características extraídas de cada modalidad se concatenan o combinan antes de ser ingresadas en el clasificador final. Esto permite que el modelo de clasificación considere las interacciones entre modalidades desde el principio.
  + Fusión Tardía: Las decisiones tomadas por los modelos individuales (por ejemplo, los vectores de probabilidad de cada modalidad) se combinan para tomar la decisión final. Este enfoque permite que cada modalidad se procese de manera independiente hasta la etapa final.
  + Fusión a Nivel de Características: Se utilizan técnicas avanzadas como la factorización de bilinealidad o mecanismos de atención para combinar las características de manera más sofisticada. Por ejemplo, la Factorización de Pooling Bilineal (FBP) puede capturar las interacciones complejas entre las características de diferentes modalidades, mejorando así el rendimiento del modelo final.
* **Mecanismos de Atención:** Los mecanismos de atención pueden dar diferentes pesos a las características provenientes de diferentes modalidades, destacando aquellas que son más relevantes para la tarea de reconocimiento de emociones.

Estas técnicas de fusión permiten aprovechar al máximo la información disponible en múltiples modalidades, mejorando así la precisión y robustez del reconocimiento de emociones en escenarios multimodales.

## 2.3. Integración de Emociones en un Chatbot

## 2.4. Datasets

## 2.5. Resultados Previos

# Capítulo 3

# Material y Métodos

## 3.1. Dataset utilizado

## 3.2. Preprocesamiento de datos

## 3.3. Modelo Multimodal propuesto

## 3.4. Chatbot emocional

Elige un avatar emoji para interactuar con el que se sienta comodo.

# Capítulo 4

# Métodos de Evaluación

# Capítulo 5

# Resultados