Detección de emociones utilizando aprendizaje de máquina multimodal para mejorar la interacción humano-robot de Pepper

Junio 2023

Juan Esteban Padilla Torres
Departamento de Ingeniería Eléctrica y Electrónica
Universidad de los Andes, Bogotá
j.padillat@uniandes.edu.co

Asesor: Juan José García Cárdenas
Departamento de Ingeniería Eléctrica y Electrónica
Universidad de los Andes, Bogotá
jj.garcia10@uniandes.edu.co

Co-asesor: Fredy Enrique Segura Quijano
Departamento de Ingeniería Eléctrica y Electrónica
Universidad de los Andes, Bogotá
fsegura@uniandes.edu.co

***Resumen*—En la actualidad es muy importante para las organizaciones mantener un servicio al cliente de primer nivel. Para esto, es necesario desarrollar sistemas automatizados que detecten las emociones de las personas al responder al acercamiento de un producto o servicio. Después de todo, las decisiones que tomamos son, de manera innata, producto de nuestras emociones; es por esto que, es importante desarrollar un ambiente agradable a la hora de atraer a un cliente, ya sea por la estética del producto, el olor, o la estimulación de otros sentidos. En este reporte se presenta el desarrollo e implementación de una aplicación para el robot Pepper que pueda identificar y extraer del contexto de una conversación las emociones de su interlocutor para ajustar sus respuestas en la siguiente interacción. El espectro de emociones para la clasificación contiene: ira, disgusto, miedo, felicidad, tristeza, sorpresa y neutral. Para el sistema se implementa una rutina de detección de emociones multimodal que extrae las características para la clasificación a partir del vídeo, audio y texto obtenidas mediante una interacción humano-robot.**

Palabras Clave*—*affective computing, aprendizaje de máquina multimodal, detección de emociones, interacción humano-robot, robot Pepper

I. Introducción

El campo de la inteligencia artificial y el *machine learning* está en auge en la actualidad; sin embargo, la aplicación de estas tecnologías en plataformas robóticas aún es un campo por explorar. De esta manera, es necesario avanzar en interfaces robóticas que permitan transmitir los algoritmos que han sido entrenados de manera exhaustiva, para así, generar un impacto en el mundo real y ayudar con tareas que mejoren la interacción entre las máquinas y los humanos. Entre estos algoritmos se encuentran modelos de detección de emociones que incluyen el análisis de datos de usuarios como la voz, el significado del texto, las expresiones faciales, la postura, entre otros; estos datos, se extraen a partir de una interacción física con el usuario. [7] [9]

Las aplicaciones de esta tecnología son muy variadas tanto para mejorar procesos entre los clientes y las empresas como para estudiar diferentes reacciones de las personas hacia estímulos del ambiente. Algunas de estas son: describir los sentimientos de un turista al observar una obra de arte; analizar las percepciones de los clientes a la hora de probar un producto, o calificar la experiencia del usuario en salas de espera. Teniendo en cuenta lo anterior, es relevante realizar una implementación de la detección de emociones en una plataforma robótica estándar en la industria como lo es el robot tipo Pepper. [8]

I-A. Objetivo General

Implementar un sistema de aprendizaje multimodal de detección de emociones en la plataforma robótica estándar tipo Pepper para mejorar su interacción con los seres humanos.

I-B. Objetivos Específicos

- Investigar el estado del arte de los modelos de detección de emociones mediante texto, imágenes y/o vídeo.
- Aplicar de forma remota diversos modelos de detección de emociones mediante texto, imágenes y/o vídeo.
- Integrar los modelos implementados en la plataforma robótica tipo Pepper para una aplicación específica en la identificación de calidad de producto.

II. Marco teórico

El área de la computación asociada a esta temática es conocida como computación afectiva (*affective computing*) la cuál, está dividida en dos ramas principales: análisis de sentimientos (*sentiment analysis*) y detección de emociones (*emotion detection*). [7] La primera se refiere al proceso de analizar automáticamente datos de texto para identificar y extraer información subjetiva como opiniones, actitudes, emociones o sentimientos; usualmente, esta información se clasifica en la práctica en un rango de positivo, neutral o malo. La segunda, se refiere al reconocimiento de estados emocionales a partir de expresiones faciales o patrones del habla para comprender cómo se sienten las personas acerca de ciertos temas o productos; en la práctica, esta información se clasifica en un espectro de emociones definido. El espectro de

emociones que se utiliza en este proyecto fue desarrollado por Paul Ekman, el cuál, clasifica las emociones humanas entre: ira, disgusto, miedo, felicidad, tristeza y sorpresa. [6]

Para la realización de este proyecto, se consideraron las diferentes modalidades que puede incluir la detección de emociones multimodal como lo son: texto, visión y voz. [8] En la primera, se hace uso principalmente de herramientas de Procesamiento de Lenguaje Natural (NLP) que implementan modelos de redes neuronales recurrentes (RNN) y cadenas ocultas de Markov (HMM). [3] [12] En la segunda, se realiza análisis de imagen aplicando modelos de redes neuronales convolucionales (CNN) para extraer las características de las personas. [11] Finalmente, es posible identificar patrones de emociones a partir de los patrones de la voz utilizando la entonación y el estrés al hablar (*prosody*). [8]

Utilizando esta teoría de *machine learning* es posible hacer un estudio de los mejores modelos implementados en la literatura para desarrollar un aplicativo que permita al robot hacer esta detección de emociones de forma precisa y robusta.

III. Metodología

Inicialmente, se consolidó un estado del arte relacionado con la detección de emociones utilizando aprendizaje de máquina multimodal (*multimodal machine learning (MML)*). En esta investigación se encontraron los principales retos del aprendizaje multimodal como se muestra en la figura 1.

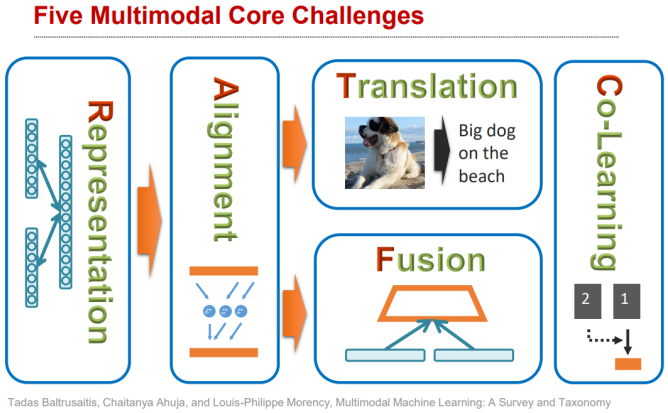

Figura 1. Los principales retos del aprendizaje multimodal [2]

En esta figura se detallan los retos relacionados con: representación, alineación, traducción, fusión y co-learning. La representación implica extraer las características de cada modalidad ya sea texto, video o voz; la alineación busca llevar estas representaciones al mismo nivel dimensional para que tenga sentido realizar alguna comparación, esto, porque cada modo se codifica de forma diferente. Después, el reto se divide en dos categorías, la traducción y la fusión. La primera busca llevar información de un modo a otro; por ejemplo, que el modelo reciba la imagen (modo 1) de un perro en una playa y que logre identificar y producir en un texto (modo 2) la descripción de esta imagen ("Big dog on the beach"). La segunda (fusión) es la de interés para este proyecto dónde, a partir de la información que viene de la alineación, se busca inferir algo más, en este caso, clasificar la emoción de la persona durante una interacción. Finalmente, existe el paradigma del co-learning en dónde se busca que el aprendizaje desarrollado por una modalidad beneficie al aprendizaje de otra, pero esto es un acercamiento a un problema diferente al desarrollado en el proyecto.

Posterior a la identificación del objetivo a seguir (fusión) se establece un modelo simple que ilustra el aprendizaje multimodal que se evidencia en la figura 2. En primer lugar, se extrae la información para entrenar el modelo de internet o las redes sociales, luego se separa esta información en cada modalidad para sacar las características y fusionarlas en un vector multimodal que permita clasificar la interacción dentro del espectro de emociones. Este proceso puede ser utilizado en la práctica para identificar la satisfacción de un cliente hacia un producto o servicio.

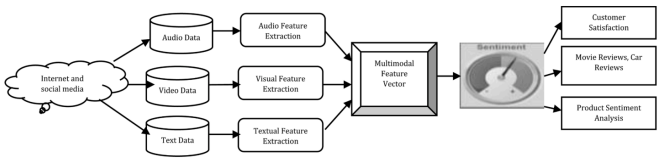

Figura 2. Modelo del proceso de aprendizaje multimodal [7]

En la literatura existen diferentes modelos de fusión. El primero y más sencillo es el denominado *early fusion* en dónde se extraen las características de cada modalidad y posteriormente se representan en un vector de fusión multimodal para realizar la clasificación como se evidencia en la figura 3. El segundo se denomina *late fusion* en dónde por cada modalidad se realiza una clasificación de las emociones y posteriormente se trabaja con esta clasificación individual para realizar una predicción más certera, presentado en la figura 4. Finalmente, existe algo denominado *memory fusion network* en dónde se trabaja con un sistema de recuerdos a corto plazo (*Long-Short Term Memories (LSTMs)*) que actualiza la clasificación a lo largo del tiempo para cada modalidad, como se ilustra en las figuras 5 y 6. Existe una variante de este modelo que utiliza eficacias (*efficacies*) para hacer una comparación con diferentes combinaciones de modalidades que permitan hacer una clasificación más fiable evidenciado en la figura 7, estas combinaciones se comparan sobre un *dataset* de reseñas extraídas de youtube y el resultado se muestra en la figura 8. Estas eficacias se muestran a los lados laterales y, aquellas con un color más rojizo, son las que aportan con mayor peso en la toma de decisión para la clasificación de emociones.

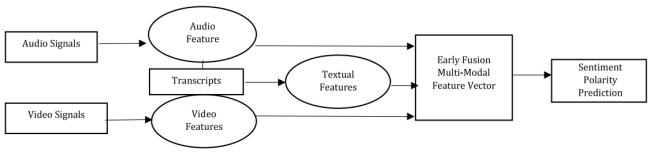

Fig. 3. Early Fusion for Multimodal Sentiment Analysis.

Figura 3. Modelo de fusión temprana (*early fusion*) [7]

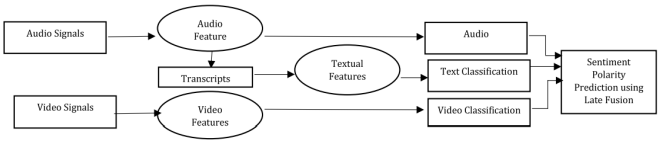

Fig. 4. Late fusion for multimodal sentiment analysis.

Figura 4. Modelo de fusión tardía (*late fusion*) [7]

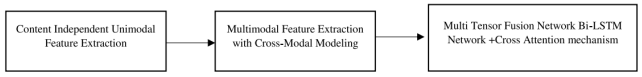

Fig. 5. MTFN architecture for tensor fusion.

Figura 5. Arquitectura MTFN para fusión de tensores [7]

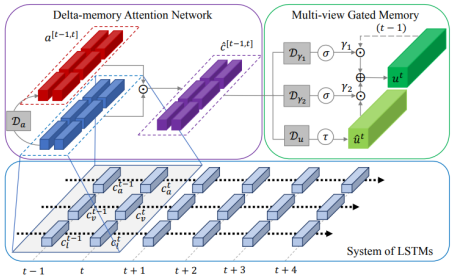

Figure 1: Overview figure of Memory Fusion Network (MFN) pipeline. σ denotes the *sigmoid* activation function, τ the *tanh* activation function, $\odot$ the Hadamard product and $\oplus$ element wise addition. Each LSTM encodes information from one view such as language (l), video (v) or audio (a).

Figura 6. Descripción de las LSTMs en el contexto de MFN [13]

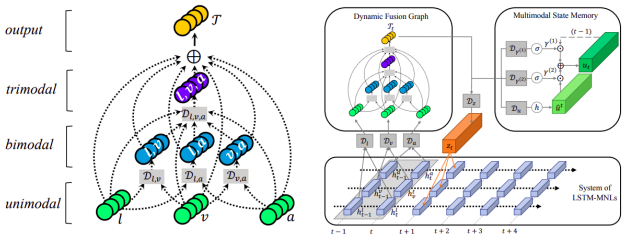

Figura 7. Modelo del Grafo de fusión dinámica (*dynamic fusion graph (DFG)*) [1]

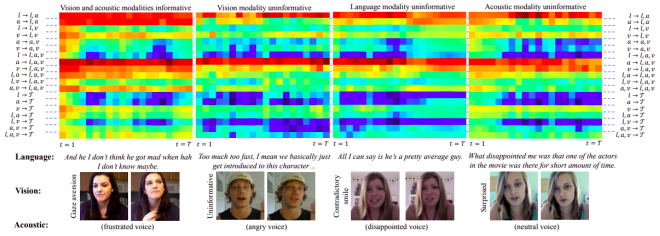

Figura 8. Visualización de las eficacias de DFG en el tiempo [1]

Para la implementación de estos modelos de aprendizaje multimodal es necesaria una gran capacidad de cómputo, sin embargo, los recursos computacionales locales con los que se cuenta durante el desarrollo del proyecto no soportan el entrenamiento conjunto de un modelo de detección de emociones multimodal. Debido a esto, entre los modelos anteriores se escoge la metodología de *late fusion* o fusión tardía la cual extrae características individualmente por cada modo y finalmente las junta para dar una predicción final. Se han realizado pruebas de clasificación unimodal para los tres tipos de datos: videos, texto y audio. Un ejemplo de prueba de detección de emociones en videos se evidencia en la figura 9; el video completo se encuentra disponible en el siguiente link.

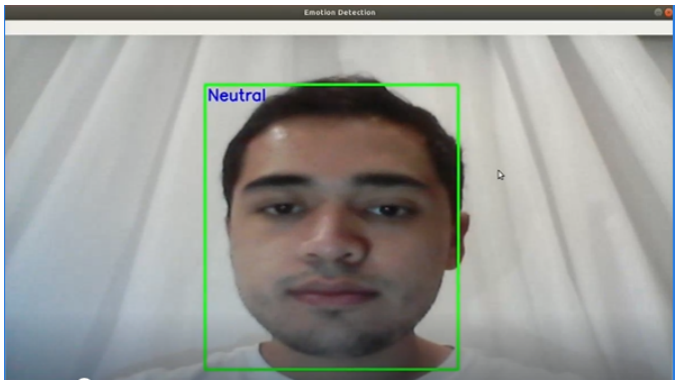

Figura 9. Pruebas para clasificación unimodal (visión) de emociones

Durante el entrenamiento de este modelo se utiliza la configuración de una red neuronal convolucional (CNN) con la estructura mostrada en la figura 10, se requiere de un amplio *dataset* y tiempo de entrenamiento para lograr ajustar esta aproximación de una función a la distribución de los datos. El *dataset* de entrenamiento es FER-2013 compuesto por más de 30000 imágenes de 48x48 píxeles en escala de grises. En un procesador i5 con 12GB de RAM el modelo puede tardar poco más de 5 horas en entrenarse. Para un video el análisis se hace por frame de tal manera que la emoción expresada en mayor cantidad de frames se toma como la resultante de la clasificación dentro de la rutina implementada.

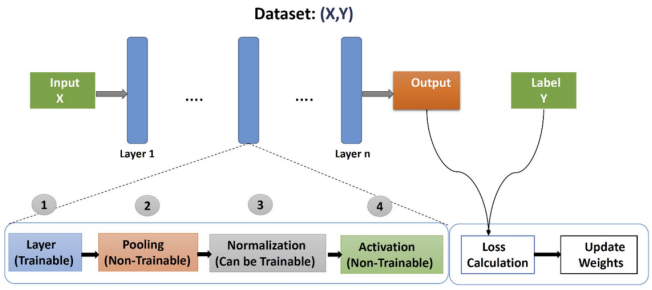

Figura 10. Configuración básica de una CNN simple [5]

En la clasificación del audio se hace uso del *dataset* RAVDESS que contiene 7356 archivos con 12 actores profesionales, 12 hombres, 12 mujeres vocalizando de manera emparejada la entonación para el espectro de emociones. [4] Después se implementa el modelo de clasificación entrenado como un Multi-layer Perceptron Classifier (MLPClassifier) y se extraen las características de los audios utilizando la librería de python *librosa* para luego utilizar estas *features* como datos de entrada de la predicción del modelo. [10] Algunas señales de audio en el tiempo se pueden evidenciar en la figura 11

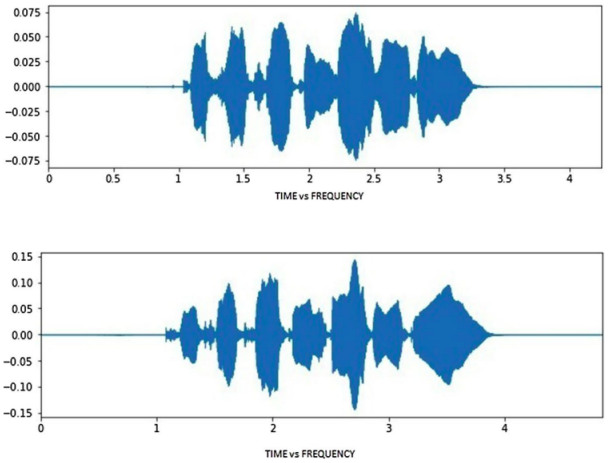

Figura 11. Señales de audio en tiempo para dos emociones: la de arriba es para alegría y la de abajo para sorpresa [4]

Por último se realiza un clasificador de emociones de texto utilizando el modelo de regresión logística para un *dataset* de tuits etiquetado dentro del espectro de emociones. Para esto, en primer lugar se realiza una limpieza de los datos filtrando cada tuit de palabras y símbolos extraños. En segundo lugar, se obtiene la raíz de las palabras utilizadas y se transforma en formato numérico por medio de una codificación tipo *bag-of-words* a partir de la función CountVectorizer de la librería *sklearn* de python.

IV. Resultados

La precisión alcanzada para los modelos realizados de video, audio y texto son 91 %, 72.4 % y 82 % , respectivamente. Sin embargo, la tasa de identificación para emociones poco representadas dentro de los *dataset* como lo son el disgusto y la furia pueden llegar a valores como 45 % o 41 %. [5] Adicionalmente, existen emociones difíciles de diferenciar dentro de la entonación para el reconocimiento de audios, por ejemplo el miedo dado que tiene valores negativos en características relevantes para la clasificación como el placer o la dominancia. [4]

Para la aplicación utilizando el robot tipo Pepper se desarrolla un paquete de ROS que se comunica con el robot y es capaz de grabar tanto audio como video de manera sincrónica durante una interacción con un humano, este paquete se encuentra disponible en el siguiente link. Posterior a esta grabación se realiza la extracción de características y se pasa como entrada a cada modelo para obtener la clasificación, la transcripción del audio se lleva a cabo utilizando el aplicativo *whisper* de OpenAI. Al obtener la clasificación de los 3 modelos, inicialmente, se plantea utilizar un método común de fusión multimodal como lo es el promedio ponderado sobre probabilidades. Sin embargo, las probabilidades deben estar asociadas a la misma distribución para que la respuesta sea consistente. No obstante, como cada modelo fue entrenado sobre *datasets* diferentes, se opta por no realizar esta aproximación dado que puede que la información brindada no corresponda de manera precisa al resultado que mostraría este método de fusión tardía.

En la validación de la solución se seleccionan dos lineas etiquetadas con la emoción real que reflejan y se realiza la prueba con el algoritmo desarrollado, los resultados se muestran en el cuadro I para pruebas tanto de manera remota en el PC como con el robot Pepper.

Cuadro I
Respuesta del sistema de detección de emociones para cada modalidad

Line	True Emotion	Text/PC	Video/PC	Audio/PC	Text/PEPPER	Video/PEPPER	Audio/PEPPER
Good! Let's go now.	Happy	Happy	Surprised	Disgust	Happy	Happy	Anger
I am fucking awesome	Happy	Happy	Happy	Happy	Happy	Anger	Disgust
Her voice shook, her face was white with anguish	Sadness	Sadness	Surprised	Disgust	Sadness	Surprised	Anger
For the owner, it can be heartbreak at the loss of a family pet	Sadness	Sadness	Fearful	Disgust	Sadness	Sadness	Anger
I was at home and we were sleeping, then we heard some robbers breaking into our house.	Fear	Fear	Fearful	Disgust	Happy	Fearful	Anger
I thought that I had failed the examination	Fear	Sadness	Surprised	Disgust	Sadness	Fearful	Anger
Pizza hut just delivered my pizza 20 minutes early and it is one of the top five best pizzas I've ever had from there	Surprise	Surprise	Fearful	Disgust	Surprise	Surprised	Anger
The awkward moment when your on a date and you look down and realize your pants aren't zipped	Surprise	Surprise	Surprised	Disgust	Sadness	Surprised	Anger
He looked crazy, a bitter anger in his eyes	Anger	Anger	Fearful	Disgust	Sadness	Happy	Anger
I hate people with bad taste in music! Seriously shut up!!	Anger	Anger	Fearful	Happy	Anger	Sadness	Anger
Never eating beef, chicken or anything involving some sort of meat!	Disgust	Disgust	Fearful	Disgust	Disgust	Sadness	Anger
Ew! I hate potato Salad	Disgust	Disgust	Fearful	Disgust	Sadness	Happy	Anger

Utilizando el computador para la modalidad de texto se tiene una tasa de reconocimiento del 91.6 %; para video del 25 % y para audio del 25 %. Mientras que para el robot Pepper se tiene una tasa de reconocimiento del 58 % para la modalidad de texto, 50 % para video y 16 % para audio. En estos resultados se evidencia que el modo más significativo para la detección de emociones es el texto, esto es porque los datos que se le dan de entrada a esta modalidad es exactamente el mismo tipo de datos que representa la distribución sobre la cuál fue entrenada, es decir, un *string* o cadena de caracteres. Por otro lado, para los modelos de video y de audio, la configuración de la cámara y micrófono, así como las condiciones del ambiente en el que se encuentra al momento de la grabación hacen que sea muy variable a la hora de compararlos con los *datasets* sobre los

cuáles fueron entrenados. Lo anterior porque se encuentran construidos dentro de condiciones controladas, mientras que en la práctica las condiciones reales hacen que la variabilidad en la predicción sea muy grande. Un ejemplo de la realización de estas pruebas se encuentra aquí.

V. Conclusiones

Es claro que la toma de decisiones en las personas se encuentra dirigida en gran medida por sus emociones. Es por esto que existe una necesidad latente por un sistema que detecte y extraiga a partir de una interacción esta información. En la investigación del estado del arte se evidencian modelos bastante robustos que permiten la identificación de las emociones utilizando aprendizaje multimodal como los algoritmos de MFN y DFG. Sin embargo, la capacidad de computación necesaria para entrenar estos modelos es bastante grande y se requiere un clúster de computación o (HPC).

A partir de las herramientas disponibles se escoge el modelo de fusión tardía (*late fusion*) que permite extraer las características individualmente por cada modo. Si bien los resultados de precisión durante el entrenamiento son bastante altos con probabilidades mayor al 70 %, las tasas de reconocimiento para los modos de video y audio difieren considerablemente al exponer los modelos a condiciones reales. Esto sucede debido a que las condiciones y extracción de características en diferentes medios como el robot Pepper cambian con respecto a las condiciones de entrenamiento de los modelos individuales.

No obstante, se observa una mejoría en la detección de emociones con la modalidad de videos utilizando el robot Pepper, lo que puede ser consecuencia del formato de la imagen que se obtiene al ser un tipo de imagen más comprimida. Adicionalmente, se observa un importante *bias* de parte de la modalidad de audio hacia una emoción en específico la cual varía en función del dispositivo de entrada; esto se explica porque los datos de entrenamiento pueden tener un *baseline* de frecuencias diferentes ajustadas a los parámetros de audio específicos del modelo. Una forma de mitigar esta situación es mediante la alineación (ver figura 1) en donde se busca llevar a cada modo al mismo nivel dimensional y realizar un entrenamiento conjunto en el futuro.

VI. Referencias

[1] A. Bagher Zadeh, P. P. Liang, S. Poria, E. Cambria y L.-P. Morency, «Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph,» en *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia: Association for Computational Linguistics, jul. de 2018, págs. 2236-2246. DOI: 10.18653/v1/P18-1208. [Internet]. Disponible en: https://aclanthology.org/P18-1208.

[2] T. Baltrusaitis, C. Ahuja y L. Morency, «Multimodal Machine Learning: A Survey and Taxonomy,» *CoRR*, vol. abs/1705.09406, 2017. arXiv: 1705.09406. [Internet]. Disponible en: http://arxiv.org/abs/1705.09406.

[3] K. Cho, B. van Merrienboer, C. Gulcehre *et al.*, «Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation,» jun. de 2014. [Internet]. Disponible en: http://arxiv.org/abs/1406.1078.

[4] A. Christy, S. Vaithyasubramanian, A. Jesudoss y M. D. Praveena, «Multimodal speech emotion recognition and classification using convolutional neural network techniques,» *International Journal of Speech Technology*, vol. 23, págs. 381-388, 2 jun. de 2020, ISSN: 15728110. DOI: 10.1007/s10772-020-09713-y.

[5] T. Debnath, M. M. Reza, A. Rahman, A. Beheshti, S. S. Band y H. Alinejad-Rokny, «Four-layer ConvNet to facial emotion recognition with minimal epochs and the significance of data diversity,» 2022. DOI: 10.1038/s41598-022-11173-0. [Internet]. Disponible en: www.nature.com/scientificreports/.

[6] P. Ekman, *Handbook of cognition and emotion*, T. Dalgleish y M. J. Power, eds. 1999, cap. 3, pág. 843, ISBN: 0471978361. [Internet]. Disponible en: https://onlinelibrary.wiley.com/doi/10.1002/0470013494.ch3.

[7] A. Gandhi, K. Adhvaryu, S. Poria, E. Cambria y A. Hussain, «Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions,» *Information Fusion*, vol. 91, págs. 424-444, mar. de 2023, ISSN: 1566-2535. DOI: 10.1016/J.INFFUS.2022.09.025.

[8] W. Graterol, J. Diaz-Amado, Y. Cardinale, I. Dongo, E. Lopes-Silva y C. Santos-Libarino, «Emotion Detection for Social Robots Based on NLP Transformers and an Emotion Ontology,» *Sensors*, vol. 21, n.º 4, 2021, ISSN: 1424-8220. DOI: 10.3390/s21041322. [Internet]. Disponible en: https://www.mdpi.com/1424-8220/21/4/1322.

[9] K. Kim y S. Park, «AOBERT: All-modalities-in-One BERT for multimodal sentiment analysis,» *Information Fusion*, vol. 92, págs. 37-45, abr. de 2023, ISSN: 15662535. DOI: 10.1016/j.inffus.2022.11.022.

[10] R. F. Livingstone SR, «The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English,» *PLoS ONE 13(5)*, 2018. DOI: 10.1371/journal.pone.0196391.

[11] G. Meena, K. K. Mohbey, A. Indian y S. Kumar, «Sentiment Analysis from Images using VGG19 based Transfer Learning Approach,» *Procedia Computer Science*, vol. 204, págs. 411-418, 2022, ISSN: 18770509. DOI: 10.1016/j.procs.2022.08.050.

[12] Z. Yao, Z. Wang, W. Liu, Y. Liu y J. Pan, «Speech emotion recognition using fusion of three multi-task learning-based classifiers: HSF-DNN, MS-CNN and LLD-RNN,» *Speech Communication*, vol. 120, págs. 11-19, jun. de 2020, ISSN: 01676393. DOI: 10.1016/j.specom.2020.03.005.

[13] A. Zadeh, P. P. Liang, N. Mazumder, S. Poria, E. Cambria y L.-P. Morency, «Memory Fusion Network

for Multi-view Sequential Learning,» feb. de 2018. [Internet]. Disponible en: http://arxiv.org/abs/1802.00927.