ISSN 1870-4069

Reconocimiento multimodal de emociones orientadas al aprendizaje

Ramón Zatarain Cabada, María Lucia Barrón Estrada,
Héctor Manuel Cárdenas López

Instituto Tecnológico de Culiacán,
Culiacán, Sinaloa, México
{rzatarain, lbarron, hector_cardenas}@itculiacan.edu.mx

Resumen. Las emociones centradas en el aprendizaje tienen un rol significativo en el proceso pedagógico del estudiante. Por esta razón, es relevante en los ambientes de aprendizaje virtual tomar en consideración no solo los aspectos cognitivos del estudiante si no también los aspectos afectivos. Algunos métodos como el auto reporte, observacional y las imágenes de actividad neuronal, han sido usados extensivamente para medir emociones. Sin embargo, hoy en día otros métodos de análisis de sentimientos y reconocimiento facial usando inteligencia artificial han demostrado ser una mejor alternativa para el reconocimiento automatizado de afecto. Los resultados son superiores cuando se combinan diferentes métodos para combinar el reconocimiento de expresiones faciales, expresiones textuales, aplicadas en ambientes de aprendizaje. Para la implementación de reconocedores, utilizamos nuestro propio corpus para la clasificación que emplea técnicas de aprendizaje profundo. Evaluamos la eficiencia de tres métodos de fusión comparándolas contra métodos unimodales de reconocimiento de emociones. La mejora de uno de los métodos (Representación basada en embebidos) de fusión fue de 4% (precisión de 86.20% y perdida 1.35%). Los resultados son muy prometedores y pensamos que serán mejor con tres o más reconocedores participando en el sistema multimodal.

Palabras clave: reconocimiento de emociones, reconocimiento multimodal, método de fusión.

Multimodal Recognition for Learning Centered Emotions

Abstract. Learning-centered emotions have a significant role in the pedagogical process of a student. For this reason, it is relevant that virtual learning environments take into account not only the cognitive aspects of the student but also the affective ones. Some approaches such as self-report, observational, and neuroimaging have been used extensively to measure emotions. However, today other methods of artificial intelligence such as the recognition of facial expressions and sentimental analysis have proven to be a better alternative in the automatic recognition of affect, and the results are superior when we combine and integrate several modes of recognition. In this work, we present three different methods to combine the recognition of facial expressions and textual expressions, applied to virtual learning environments. For the implementation of recognizers, we use our own corpora with classifiers that employ deep learning techniques. We evaluate the efficiency of the three fusion methods by comparing them against unimodal methods of emotion recognition. The improvement of one

of the methods (Embedding-based representation) of fusion was 4% (precision of 86.20% and loss of 1.35%). The results are very encouraging, and we think they will be better with three or more recognizers participating in the multimodal system.

Keywords: emotion recognition, multimodal recognition, fusion method.

1. Introducción

El mundo que nos rodea contiene múltiples modalidades. Vemos objetos, escuchamos sonidos y olemos olores. De esta manera el ser humano relaciona cada una de estas modalidades con un método de obtención de datos, sea la vista, el oído y el olfato respectivamente. De la misma manera nuestro cerebro relaciona a algunos olores con percepción de sentidos diferentes como la vista, creando relaciones importantes entre el olor de una tarta de fresa y la figura de dicha tarta. De la misma manera la inteligencia artificial busca el uso de estas diferentes modalidades para entender el mundo alrededor de nosotros y crear relaciones entre imágenes, voz, texto, etc.

Por otro lado, las emociones tienen un rol fundamental en la mayoría de las actividades de los seres humanos, ya sea en el proceso cognitivo y en entornos laborales [1], así como también se ha mostrado que las emociones están fuertemente ligadas con el proceso de aprendizaje y los niveles de concentración de las personas [2,3]. De la misma manera las emociones están ligadas a los procesos socioeconómicos [4], en los cuales se ha mostrado que existe una correlación entre el movimiento de la bolsa de valores y los movimientos generales económicos con el sentimiento generalizado de la población.

En la última década ha habido un gran interés en el desarrollo de tecnologías y estudios que permitan realizar un reconocimiento de emociones de manera automatizada, utilizando técnicas de aprendizaje máquina (ML) y aprendizaje profundo (DL). El reconocimiento de emociones se logra a través de la extracción de características en diferentes modalidades, como la voz, expresión facial, expresiones corporales, análisis de sentimientos (SA) y minería de opiniones (OM). Dentro del área de aplicación de tecnologías para el aprendizaje y más específicamente en el uso de ambientes de aprendizaje (AA) el reconocimiento de emociones es una herramienta de gran valor que permite mejorar el proceso de enseñanza-aprendizaje ya que una retroalimentación de un estudiante a través de un sistema de aprendizaje ayuda a comprender de mejor manera el proceso de aprendizaje del alumno. Esta retroalimentación contiene opiniones abiertas que permiten al AA realizar cambios en la programación de los materiales, estrategias y otros aspectos del proceso de enseñanza.

Uno de los principales problemas del uso de sistemas de clasificación se encuentra en la robustez que presentan estos sistemas para ser utilizados en ambientes no controlados que comúnmente generan muestras con ruido. Dentro de las técnicas de ML y DL se ha descubierto que el uso de diferentes modalidades para la clasificación genera modelos que presentan una mayor precisión y robustez. Debido a esto nosotros proponemos la implementación de un sistema multimodal para el reconocimiento de emociones orientadas al aprendizaje para su aplicación en AA.

En este artículo, presentamos tres metodologías creadas para la fusión de datos de emociones orientadas al aprendizaje para el reconocimiento multimodal (rostro y texto) utilizando DL. En la primera metodología utilizamos una fusión de representación de datos basada en imágenes para el entrenamiento de una red neuronal convolucionada (CNN). En la segunda metodología aplicamos una fusión de representación de datos basada en embebidos N-dimensionales para el entrenamiento de una red neuronal convolucionada combinada con memorias largas de corto plazo (CNN-LSTM). Por último, se utilizó un sistema hibrido de capas convolucionadas (CC) y con memorias largas de corto plazo (CLSTM) para la extracción de características y una capa de fusión de vectores de características para entrenar una red densamente conectada (FC).

Para la implementación de la fusión de datos se utilizaron tres diferentes corpus; uno de ellos basado en imágenes de rostros de personas recopilados durante el proceso de aprendizaje de programación java y dos corpus recopilados de opiniones en texto relacionadas también con el proceso de aprendizaje.

Nuestra principal contribución es el uso de metodologías multimodales para el reconocimiento de emociones orientadas al aprendizaje para ambientes no controlados, realizando una comparativa entre el desempeño de diferentes metodologías unimodales y nuestras metodologías multimodales centradas en la fusión de datos.

2. Trabajos relacionados

En esta sección se describen algunos trabajos de investigación relacionados con la aplicación de técnicas de DL para el reconocimiento de emociones, el uso de sistemas multimodales para la creación de modelos de DL y algunos trabajos de reconocimiento multimodal de emociones. Estos trabajos contienen similitud con nuestro trabajo y algunas técnicas de ellos fueron consideradas y abordadas para el desarrollo de este proyecto.

La relación entre las emociones básicas y las expresiones faciales fue estudiada extensivamente por Paul Ekman [5]. Su trabajo ha servido como una guía para el desarrollo de sistemas computacionales que sean capaces de reconocer emociones a través de las diferentes modalidades de entrada, como el rostro, la voz, expresiones faciales, etc.

Hoy en día existen muchos trabajos de investigación enfocados al área de reconocimiento de emociones utilizando modalidades como la voz y el rostro para la extracción de características que han logrado crear modelos de alta precisión utilizando técnicas de ML [6] y DL [7] con el uso de redes de creencia profunda (BLN). Además, se ha observado un gran avance en la minería de opiniones y la clasificación de emociones en texto utilizando técnicas de DL [8].

Las técnicas de reconocimiento de emociones también han sido utilizadas en conjunto con ambientes de aprendizaje utilizando diferentes clasificadores de DL. En [9] se utilizó minería de opiniones para obtener datos relacionados con el proceso de aprendizaje de Twitter para entrenar una CNN-LSTM para la clasificación de emociones. De la misma manera se ha utilizado el reconocimiento de emociones en el rostro para la clasificación de emociones utilizando diferentes técnicas de extracción de características y reconocimiento de expresiones faciales [10].

Sin embargo, el reconocimiento de emociones es un desafío, especialmente si los datos utilizados para clasificar las emociones son ruidosos. Esta es una de las razones por las cuales muchos sistemas de predicción o reconocimiento de emociones no reconocen correctamente en ambientes no controlados (fuera de laboratorios). En busca de resolver este problema se han creado concursos con corpus en ambientes no controlados y DL ha mostrado ser una de las técnicas más efectivas para este problema [11].

Existen diferentes trabajos que utilizan sistemas multimodales que demuestran un mejor desempeño comparados con sus contrapartes unimodales para diferentes propósitos, como el reconocimiento y auto etiquetado de contenido multimedia para búsquedas en línea [12], la clasificación de géneros de música [13], reconocimiento de contexto y actividades [14] y reconocimiento de objetos más robustos utilizando sistemas multimodales con DL [15]. Se ha realizado mucho trabajo en el campo de reconocedores de emociones multimodales. Un ejemplo es el uso de Tumblr para el análisis de sentimientos y reconocimiento de emociones [16] donde se realizó una comparación de desempeños entre dos sistemas unimodales y un sistema multimodal para el etiquetado de imágenes según su relación con alguna emoción o sentimiento.

Otro trabajo importante ha sido el reconocimiento multimodal de emociones utilizando arquitecturas de DL con datos fisiológicos, de voz y rostro [17], en donde se utilizó una red convolucional de creencia profunda para clasificar emociones sutiles o de baja intensidad.

De la misma forma, se ha realizado investigación acerca de modelos de reconocimiento de emociones multimodales utilizando datos duros de sensores obtenidos de dispositivos móviles y accesorios [18] en donde se utilizó un modelo multimodal CNN-LSTM y unimodal CNN para reconocimiento de emociones en humanos, en el cual se prueba que los modelos multimodales con técnicas de DL tienen un mejor desempeño que los modelos unimodales.

3. Reconocimiento multimodal

El reconocimiento es un proceso para predecir una etiqueta o una clase dada una cierta cantidad de información. Los reconocedores multimodales son una aproximación de la inteligencia artificial (IA) para darle la capacidad a un modelo de procesar y relacionar información de múltiples modalidades para discriminar y clasificar información.

Realizar este tipo de reconocedor requiere de crear sistemas que permitan procesar la información de cada una de estas modalidades para obtener una representación generalizada de cada una de ellas, ya sea a través de la fusión de datos en su representación, a la creación de sistemas de reconocimiento independientes, fusionando las clasificaciones directamente o a la creación de sistemas que permitan extraer características directamente de los datos y fusionar estas características para entrenar modelos de clasificación.

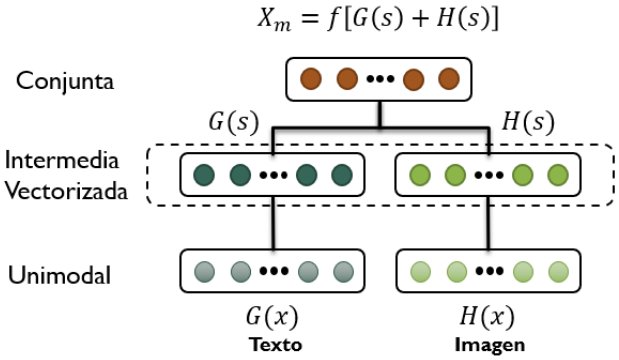
3.1. Representación multimodal de datos

La representación de datos se refiere a la creación de estructuras numéricas que representen de manera adecuada alguna entrada de valor de datos. Representar datos

duros en un formato que un modelo computacional pueda trabajar es uno de los más grandes retos del área de ML y DL multimodal. La representación de datos está fuertemente ligada al modelo de clasificación que se pretende implementar. Las redes neuronales convolucionadas por ejemplo utilizan imágenes como su medio de entrada que son pasadas por filtros para la extracción de características.

Para la representación multimodal de datos utilizamos una representación conjunta la cual consiste en la creación de representaciones intermedias basadas en modelos de interpretación de datos para ambas modalidades, de imagen y texto para posteriormente realizar una operación matemática creando así una nueva representación conjunta usada para entrenar modelos de DL. Un ejemplo de representación conjunta se puede observar en la figura 1.

Fig. 1. Representación conjunta de datos.



Para la representación de los datos de la modalidad de imagen se utilizaron dos modelos de color; el primero fue el modelo de color RGB con el cual se representaron los datos individuales de cada píxel a través de su valor hexadecimal creando una matriz representativa de cada uno de los colores contenidos en la imagen. El segundo fue el modelo de color a escala de grises, con el cual se determinaron números enteros representativos de cada uno de los valores de intensidad del color blanco dentro de la escala de grises, creando una representación matricial de la imagen a través de los valores individuales de cada píxel contenido en ella. Ambos modelos se observan en la figura 2.

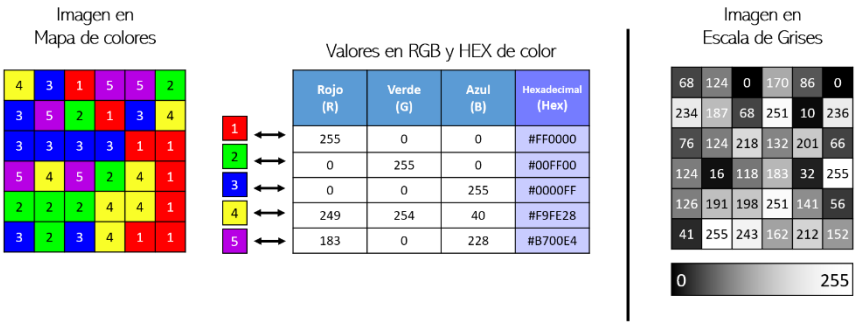


Fig. 2. Modelos de color para representación de imágenes.

Para la representación de los datos en la modalidad de texto se crearon diccionarios que utilizaban los índices para crear representaciones numéricas de cada una de las palabras contenidas dentro del corpus, creando un nuevo índice por cada palabra nueva encontrada dentro del corpus cuando este fue analizado. Un ejemplo de este diccionario se muestra en la figura 3.

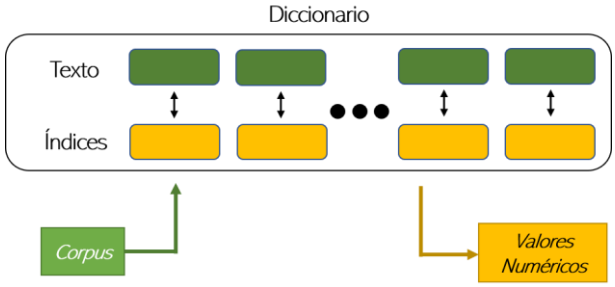


Fig. 3. Representación de diccionario para la traducción.

Una vez obtenidas las representaciones unimodales de datos se crearon corpus con representaciones conjuntas a través de sistemas de representación multimodal, tomando en cuenta los modelos a utilizar para la clasificación y el reconocimiento de emociones. En el caso de la CNN se utilizó una representación conjunta basada en imágenes debido a los métodos de extracción de características de este tipo de modelos y para el entrenamiento de una CNN-LSTM se utilizó una representación con embebidos de información. Para esto se realizó una representación intermedia en forma de vectores tanto de imágenes como de texto y se utilizó una capa de creación de embebidos previa al entrenamiento de este modelo.

3.2. Descripción de los corpus

El corpus es de suma importancia para la creación de cualquier modelo de clasificación o reconocimiento. Para este experimento utilizamos tres diferentes corpus unimodales, dos de ellos de texto y uno de imagen. Los corpus fueron creados en diferentes tiempos con técnicas diferentes. A continuación, se describen los tres corpus de manera breve, así como una pequeña explicación de su metodología de creación. Una descripción más detallada de la metodología de creación de estos corpus de imagen se puede observar en [9] y de texto en [10].

- **Corpus Sentitext (CST):** Consiste en un corpus recopilado de Twitter con 23,173 mensajes enfocados en el área del aprendizaje, etiquetados de manera manual buscando polaridad positiva o negativa en el texto.
- **Corpus Edusere (CES):** Consiste en un corpus recopilado de Twitter con 9,963 mensajes con enfoque en el área del aprendizaje, etiquetado de manera manual utilizando 5 de las emociones orientadas al aprendizaje (aburrido, enganchado, emocionado, concentrado, interesado)
- **Corpus Insight (CEI):** Este corpus de emociones en imágenes de rostro fue recopilado usando una diadema EMOTIV Insight con 5 canales de electroencefalograma a través del uso de un programa computacional para enseñar

JAVA utilizado en 38 estudiantes, 28 hombres y 10 mujeres con el cual se obtuvo un total de 5560 imágenes etiquetadas en 5 emociones orientadas al aprendizaje (aburrido, enganchado, emocionado, concentrado, interesado)

Estos corpus fueron utilizados para la creación de nuevos corpus multimodales utilizando las técnicas de representación conjunta de datos. Para esto se utilizó el CST en conjunto con el CEI para la creación de un corpus con una representación basada en imágenes (CSTEIBI). Además de esto se creó un segundo corpus con una representación conjunta utilizando los corpus CST y CEI creando así un nuevo corpus con representación basada en embebidos (CSTEIBE).

3.3. Sistemas de fusión de representación

Para llevar a cabo la fusión de representaciones unimodales en una representación multimodal, se debe convertir una modalidad en función de otra. Para ello se llevan a cabo tres pasos. Primero se hace un preprocesamiento de la información contenida en los corpus unimodales, donde se da a cada entidad del corpus una representación intermedia, ya sea a través de números, vectores o datos duros. Después se define el medio de representación final de los datos a la cual se convertirá esta representación intermedia a través de modelos de conversión o diccionarios. Por último, se realiza la creación de una metodología de conversión para los datos en su representación intermedia a su representación final. Este proceso se aprecia en la figura 4.

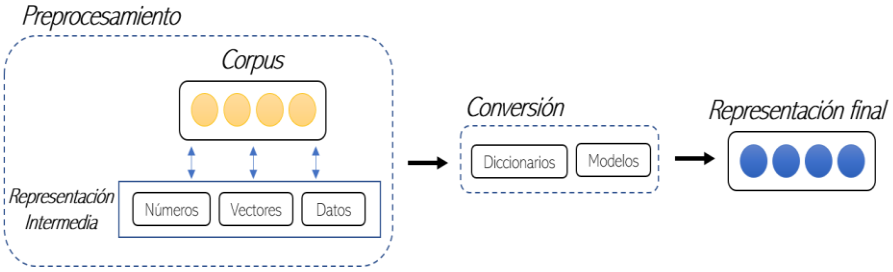


Fig. 4. Proceso de conversión para la fusión de modalidades.

Ambos sistemas de fusión con representación de datos siguen este patrón y sus implementaciones se muestran a continuación de manera detallada.

Fusión con representación basada en imágenes (FRBI). En la fusión con representación basada en imágenes se utilizaron dos corpus; el CEI para la modalidad de imagen con cinco etiquetas de emociones orientadas al aprendizaje y el CST para la modalidad de texto con dos etiquetas de polaridad en texto. Los corpus fueron fusionados para crear el nuevo CSTEIBI con diez etiquetas de polaridad en conjunto con emoción orientada al aprendizaje.

El preprocesamiento se realizó a través de un diccionario para la conversión del CST a una representación intermedia vectorizada. Primero creamos tokens de los documentos del corpus, después creamos un vector para contener y ordenar esos tokens.

Se indexaron los tokens no repetidos para asignarles índices en el diccionario, utilizando el índice como la representación de la palabra en número entero.

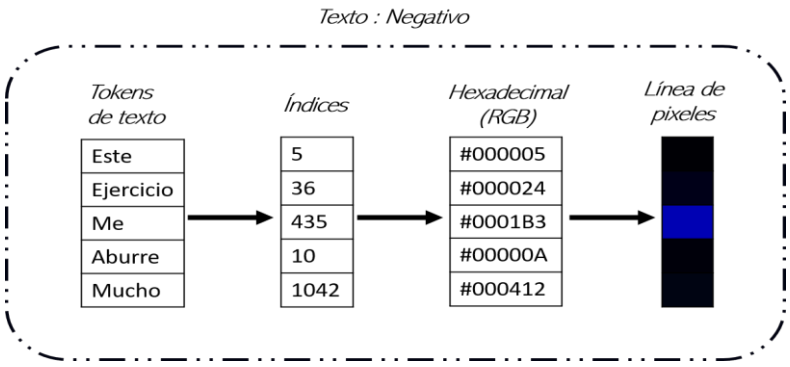


Fig. 5. Proceso de conversión de texto a imagen.

Dado que la representación final se encuentra en formato de imagen se utilizó el valor entero del índice de los tokens convertido a valor hexadecimal para crear un nuevo vector de 150 de tamaño. Utilizando el formato RGB se creó una representación en imagen del vector de texto rellenando los espacios vacíos con ceros. Esto se convirtió a su representación de una imagen de 1 x 150 pixeles de tamaño. Este proceso se muestra en la figura 5.

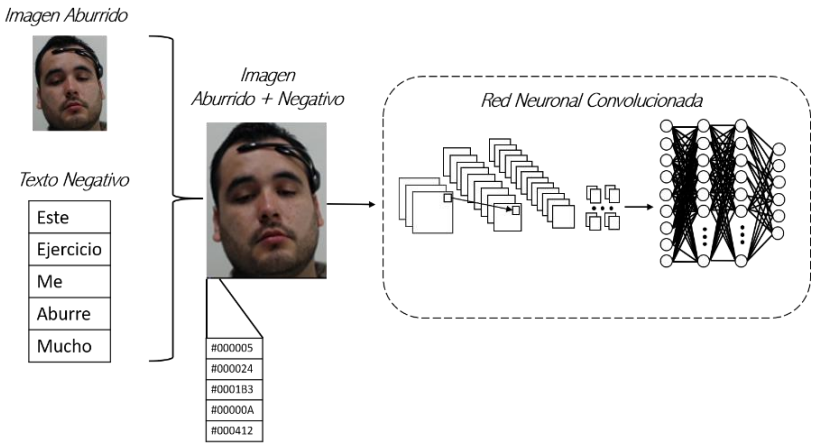


Fig. 6. Sistema de fusión de representación basada en imágenes.

Después se obtuvo la imagen del CEI y se preproceso para su transformación en una imagen de 150x150 pixeles de tamaño con 3 dimensiones en formato RGB. Una vez que ambas representaciones se encuentran en su formato de imagen se concatenan formando una nueva imagen de 150x151 pixeles de tamaño que representan la imagen del rostro y el texto. Se concatenaron las etiquetas del texto con la imagen para crear nuevas etiquetas que representaran ambas modalidades. Esta nueva imagen

representativa de ambas modalidades se anexo al nuevo CSTEIBI. Por cada imagen se obtuvieron polaridades positivas y negativas para crear representaciones conjuntas en forma de imagen.

Por último, se creó un modelo de CNN en el cual se utilizó el CSTEIBI para el entrenamiento y la validación. Esto se puede observar en la figura 6.

Fusión con representación basada en embebidos (FRBE). Para la fusión con representación basada en embebidos se utilizaron dos corpus; el CEI para la modalidad de imagen con cinco etiquetas de emociones orientadas al aprendizaje, y el CST para la modalidad de texto con dos etiquetas de polaridad, los cuales fueron fusionados para crear el nuevo CSTEIBE con diez etiquetas con pares de polaridad y emoción orientada al aprendizaje.

Para el preprocesamiento se utilizó un diccionario con los primeros 255 índices reservados para la conversión del CST y el CEI. Se obtuvieron las imágenes del CEI de las cuales se obtuvo un subconjunto de imagen centrada en el rostro. Se redimensionó la imagen a un tamaño de 150x150 pixeles y se convirtió a escala de grises, creando una matriz con la cual representa la intensidad del color blanco en la escala de grises con un máximo de 255 de valor. Por último, se procedió a convertir la matriz a vector, concatenando el último número de una fila con el primero de la siguiente. Este proceso se observa en la figura 7.

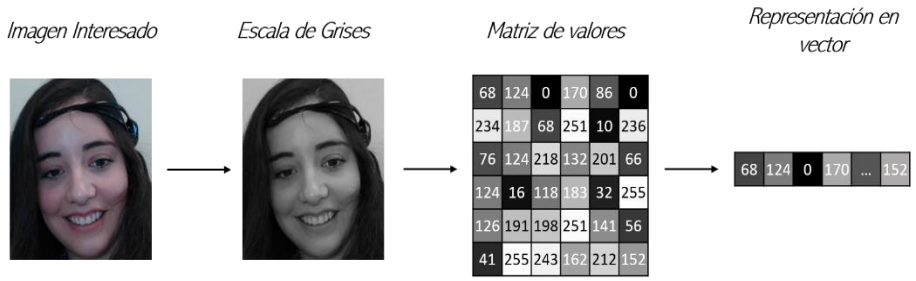


Fig. 7. Representación de imagen a vector.

De la misma manera preprocesamos el CST utilizando la misma técnica anteriormente mencionada a una representación intermedia vectorizada. Primero creamos tokens de los documentos del corpus, después creamos un vector para contener y ordenar esos tokens. Se indexaron los tokens no repetidos a partir del índice 256 para asignarles índice en el diccionario, utilizando el índice como su representación de la palabra en número entero. Se creó un vector para contener la información de cada uno de los documentos y se creó la representación vectorizada utilizando el diccionario.

Dado que la representación final para el entrenamiento de la red era un espacio de embebidos, se procedió a concatenar los vectores de imagen y texto para crear un nuevo vector conjunto, y de la misma forma se concatenaron las etiquetas para formar el nuevo CSTEIBE, el cual contenía la información en vector del texto y la imagen. Se utilizaron dos polaridades de texto por imagen para la creación de este nuevo corpus.

Finalmente, se creó un modelo de CNN-LSTM en el cual se colocó una capa de embebidos inicial, se utilizó el siguiente número superior al índice total del diccionario

para la creación de las dimensiones de entrada de la capa de embebidos, y se utilizó un valor arbitrario de 128 para las dimensiones del espacio vectorial para la creación de los embebidos. Se realizó el entrenamiento y se validó el desempeño del modelo.

3.4. Sistema de fusión de características (SFC)

Comúnmente se utilizan los términos característica y representación de manera intercambiable, ambos refiriéndose a una representación de tensor o vector de una entidad. Sin embargo, para la definición de este experimento, la palabra característica se define como el conjunto numérico vectorizado que representa una porción de dicha entidad, mientras que la representación es el medio en el cual se ve reflejada dicha entidad, pudiendo ser esto, una imagen, un texto, etc.

Para el sistema de fusión de características se realizó un proceso de preentrenamiento de un modelo CNN para la detección de cinco emociones orientadas al aprendizaje en imágenes utilizando el corpus CEI, y un modelo de CNN-LSTM con una capa de embebidos para la detección de cinco emociones en texto utilizando el corpus CES. Una vez que ambos modelos fueron entrenados y validados se guardó la configuración del grafo de cada modelo y sus pesos.

Se creó un nuevo modelo hibrido con las capas de extracción de características del modelo CNN y las capas de extracción de características del modelo CNN-LSTM con la capa de embebidos. Se anexó a este nuevo modelo una capa de concatenación de características con la cual se combinan las características extraídas de ambos modelos unimodales y se anexó al modelo una red neuronal densamente conectada para la clasificación. Esta arquitectura se muestra en la figura 8.

Finalmente, se inicializaron los pesos utilizando los valores obtenidos de las capas de extracción de características de los dos modelos unimodales de texto y de imagen, se entrenó y evaluó este nuevo modelo.

Para el entrenamiento se preprocesaron las imágenes del CEI obteniendo una imagen de la cara de 150 x 150 pixeles de tamaño con tres dimensiones en formato RGB. De la misma manera se preprocesó el texto utilizando un proceso de creación de tokens de texto, creación de un diccionario y creación de un vector representativo de los documentos del corpus.

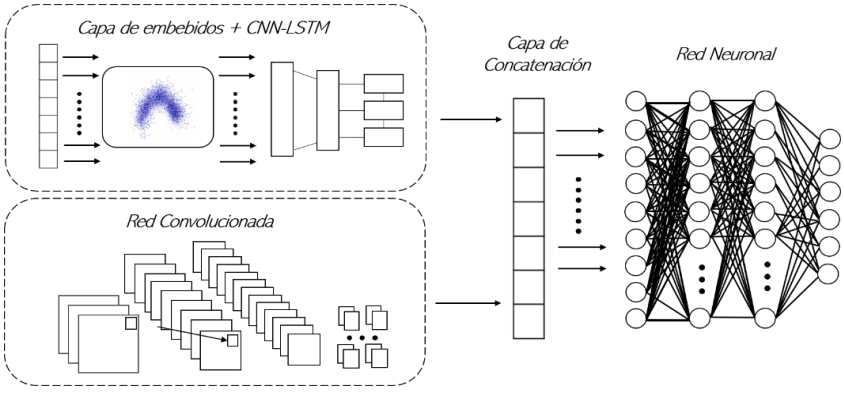


Fig. 8. Modelo de fusión de características.

Debido a que los corpus provenían de diferentes medios de información, se buscó que el entrenamiento se realizara con etiquetas idénticas en los datos utilizados.

4. Pruebas y discusión

Para los experimentos se utilizó el corpus CEI para el entrenamiento de un modelo de CNN y modelos de CNN-LSTM con una capa de embebidos para el CST y el CES. Los experimentos fueron realizados con validación cruzada de k iteraciones con k = 5 debido a que el entrenamiento de este tipo de modelos tomaba entre 1 y 2 días, especialmente con los modelos que utilizan una capa de embebidos. De la misma manera se tomó en consideración que la distribución de clases de los corpus dentro de las iteraciones fuera uniforme. En cada uno de los modelos entrenados la distribución de los corpus fue 80% para el entrenamiento y 20% para la validación del modelo.

Se utilizaron diferentes versiones de los dos corpus de texto (CST 2018, CST 2019a, CST 2019b, etc.). Las diferentes versiones que fueron creadas se diferencian en algunas expresiones que fueron agregadas y corregidas en cada uno de ellos.

Las métricas utilizadas para la evaluación de estos modelos y sistemas fueron basadas en los siguientes parámetros: la precisión en la clasificación, la perdida de clasificación, el tamaño total de los corpus y el número de clases. La Tabla 1 muestra los resultados de las pruebas.

El modelo multimodal FRBI presentó los peores resultados (62.13 %) comparado con los otros modelos unimodales y multimodales. El modelo multimodal FRBE presentó resultados prometedores con 86.20 % de precisión en la etapa de validación. Los modelos unimodales que utilizaron el mismo corpus obtuvieron 91% de precisión en clasificación de polaridad (solamente dos clases) en texto usando el CST2019b y 74.16% en clasificación de emociones para expresiones faciales (cinco clases) utilizando el CEI. Como podemos ver en la tabla, el desempeño del FRBE fue mejor en la clasificación de diez clases comparado con los modelos unimodales.

Tabla 1. Comparación de los diferentes modelos y corpus del experimento.

Corpus	Modelo	Precisión	Perdida	Tamaño	Clases
CST 2018	CNN+LSTM	79.08%	1.72%	7492	2
CST 2019a	CNN+LSTM	87.28%	1.25%	164990	2
CST 2019b	CNN+LSTM	91.80%	1.10%	22554	2
CES 2018	CNN+LSTM	69.16%	2.76%	3017	5
CES 2019a	CNN+LSTM	69.47%	3.36%	4504	5
CES 2019b	CNN+LSTM	69.47%	3.45%	4504	5
CEI	CNN	74.16%	2.41%	5056	5
(CEI+CST2019b)	FRBI	62.13%	4.20%	50164	10
(CEI +CST2019b)	FRBE	86.20%	1.35%	50164	10
(CEI+CES2019b)	SFC	81.16%	1.57%	9560	5

El modelo SFC presentó buenos resultados con una precisión del 81.16 % comparado con los modelos unimodales para el reconocimiento de emociones en imágenes utilizando el CEI donde el modelo de CNN presentó 74.16 % de precisión. Por el otro lado los modelos unimodales de reconocimiento de texto usando diferentes versiones del corpus CES presentaron resultados menores al 70% de predicción durante su validación.

5. Conclusiones y trabajo futuro

En este artículo propusimos tres diferentes aproximaciones al reconocimiento de emociones multimodal, utilizando técnicas de representación conjunta con fusión de datos y un modelo de fusión de características. El objetivo de esta investigación fue el buscar una mejoría en el reconocimiento de emociones utilizando técnicas multimodales en vez de crecer los corpus de manera directa.

Usando técnicas multimodales logramos mejorar la precisión de modelos de reconocimiento utilizando corpus unimodales a través de representación conjunta de datos para el entrenamiento. Se ha mostrado también que esta mejoría solamente es posible cuando la representación de datos es apropiada. En la FRBI concluimos que la representación de los datos de texto (a través de agregar texto como una línea de pixeles) no era adecuada para el entrenamiento de un modelo de CNN por lo cual entregó resultados de baja precisión. Sin embargo, en la FRBE de ambos imagen y texto mostró resultados prometedores con el uso de un modelo de CNN-LSTM. De la misma manera utilizando un SFC ha demostrado ser valioso en cuanto al desempeño en precisión de modelos para detección de emociones.

También a través de los resultados mostrados en este estudio podemos concluir que se pueden utilizar diferentes corpus unimodales para el entrenamiento de sistemas multimodales mientras se mantenga una relación semántica similar como en el caso del SFC donde las clases utilizadas para la clasificación son iguales o se haga una representación conjunta de los corpus.

Existen varias limitaciones con el estudio existente que necesitan ser observadas en trabajos futuros. Una limitación es el tamaño del corpus CEI que resulta pequeño para una tarea de reconocimiento de emociones. Otra limitante es la cantidad de métodos de representación conjunta utilizados dentro de este experimento, la cual es baja, además de que otros métodos de representación conjunta pueden ser utilizados, como por ejemplo el uso de una representación basada en filtros del texto aplicado al corpus de imágenes. Otra limitante de este estudio es que las pruebas de desempeño se realizaron en ambientes de laboratorio y solamente con datos de los corpus de entrenamiento originales. Además de esto también se vio limitada la cantidad de K en las técnicas de validación cruzada, ya que se puede utilizar una K más grande u otras técnicas de validación para corroborar el desempeño de estos modelos.

Como trabajo futuro proponemos realizar evaluaciones de desempeño de los modelos de este artículo utilizando ambientes de aprendizaje en aplicaciones con ambientes no controlados. También proponemos ampliar el corpus CEI y realizar una mayor cantidad de balanceo de clases para obtener aún mejores resultados. También proponemos continuar con la investigación de diferentes aproximaciones para mejorar el desempeño de modelos de reconocimientos de emociones como la transferencia de aprendizaje, algoritmos genéticos para selección de hiperparámetros en modelos multimodales para obtener mejores resultados.

Referencias

1. Fisher, C.D., Ashkanasy, N.M.: The emerging role of emotions in work life: an introduction, J. Organ. Behav. 21(2), 123–129 (2000)

2. Bollen, J., Mao, H., Zeng, X.-J.: Twitter mood predicts the stock market. (2010)
3. Tyng, C.M., Amin, H.U., Saad, M.N.M., Malik, A.S.: The Influences of Emotion on Learning and Memory. Front. Psychol., vol. 8, p. 1454 (2017)
4. Linnenbrink-Garcia, L., Pekrun, R.: Students' emotions and academic engagement: Introduction to the special issue, Contemp. Educ. Psychol. 36(1), 1–3 (2011)
5. Ekman, P.: An argument for basic emotions. Cogn. Emot. 6(3–4), 169–200 (1992)
6. Crangle, C.E., Wang, R., Perreau-Guimaraes, M., Nguyen, M.U., Nguyen, D.T., Suppes, P.: Machine learning for the recognition of emotion in the speech of couples in psychotherapy using the Stanford Suppes Brain Lab Psychotherapy Dataset (2019)
7. Hassan, M.M., Alam, M.G.R., Uddin, M.Z., Huda, S., Almogren, A., Fortino, G.: Human emotion recognition using deep belief network architecture. Inf. Fusion, vol. 51, pp. 10–18 (2019)
8. Chatterjee, A., Gupta, U., Chinnakotla, M.K., Srikanth, R., Galley, M., Agrawal, P.: Understanding Emotions in Text Using Deep Learning and Big Data. Comput. Human Behav., vol. 93, pp. 309–317 (2019)
9. González-Hernández, F., Zatarain-Cabada, R., Barrón-Estrada, M.L., Rodríguez-Rangel, H.: Recognition of learning-centered emotions using a convolutional neural network. J. Intell. Fuzzy Syst. 34(5), 3325–3336 (2018)
10. Oramas Bustillos, R., Zatarain Cabada, R., Barrón Estrada, M.L., Hernández Pérez, Y.: Opinion mining and emotion recognition in an intelligent learning environment. Comput. Appl. Eng. Educ. (2018)
11. Levi, G., Hassner, T.: Emotion Recognition in the Wild via Convolutional Neural Networks and Mapped Binary Patterns. In :Proceedings of the 2015 ACM on International Conference on Multimodal Interaction - ICMI '15, pp. 503–510 (2015)
12. Yuhas, B.P., Goldstein, M.H., Sejnowski, T.J.: Integration of acoustic and visual speech signals using neural networks. IEEE Commun. Mag. 27(11), 65–71 (1989)
13. Bartolini, I., Ciaccia, P.: Scenique. In: Proceedings of the working conference on Advanced visual interfaces - AVI '08, p. 476 (2008)
14. Oramas, S., Barbieri, F., Nieto, O., Serra, X.: Multimodal Deep Learning for Music Genre Classification. Trans. Int. Soc. Music Inf. Retr. 1(1), 4–21 (2018)
15. Radu, V., Bhattacharya, S., Lane, N.D., Marina, M.K., Tong, C., Mascolo, C., Kawsar, F.: Multimodal Deep Learning for Activity and Context Recognition, vol. 1, no. 4, p. 27 (2017)
16. Jost, A.E., Springenberg, T., Spinello, L., Riedmiller, M., Burgard, W.: Multimodal Deep Learning for Robust RGB-D Object Recognition.
17. Hu, A., Flaxman, S.: Multimodal Sentiment Analysis to Explore the Structure of Emotions, (2018)
18. Ranganathan, H., Chakraborty, S., Panchanathan, S.: Multimodal emotion recognition using deep learning architectures. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1–9 (2016)