SEP TNM

INSTITUTO TECNOLÓGICO DE CULIACÁN

Reconocimiento multimodal de emociones orientadas al aprendizaje para tutores inteligentes en ambientes Android

TESIS

PRESENTADA ANTE EL DEPARTAMENTO ACADÉMICO DE ESTUDIOS DE POSGRADO DEL INSTITUTO TECNOLÓGICO DE CULIACÁN EN CUMPLIMIENTO PARCIAL DE LOS REQUISITOS PARA OBTENER EL GRADO DE

MAESTRO EN CIENCIAS DE LA COMPUTACIÓN

POR:

HÉCTOR MANUEL CÁRDENAS LÓPEZ
INGENIERO EN MECATRONICA

DIRECTOR DE TESIS:

DR. RAMON ZATARAIN CABADA

CULIACÁN, SINALOA Agosto 2019

"2019, Año del Caudillo del Sur, Emiliano Zapata"

Culiacán, Sin., 5 de Agosto del 2019

DIVISIÓN DE ESTUDIOS DE POSGRADO E INVESTIGACIÓN
OFICIO: DEPI: 329/VIII/2019

ASUNTO: **Autorización Impresión**

ING. HÉCTOR MANUEL CÁRDENAS LÓPEZ
ESTUDIANTE DE LA MAESTRÍA EN CIENCIAS DE LA COMPUTACIÓN
PRESENTE.

Por medio de la presente y en virtud de que ha completado los requisitos para el examen de grado de la **Maestría en Ciencias de la Computación**, se concede autorización para la impresión de la tesis titulada: **"RECONOCIMIENTO MULTIMODAL DE EMOCIONES ORIENTADAS AL APRENDIZAJE PARA TUTORES INTELIGENTES EN AMBIENTES ANDROID"** bajo la dirección del(a) **Dr. Ramón Zatarain Cabada**

Sin otro particular reciba un cordial saludo.

A T E N T A M E N T E
Excelencia en Educación Tecnológica®

SEP TecNM
Instituto Tecnológico de Culiacán
División de Estudios de Posgrado e Investigación

M.C. MARÍA ARACELY MARTÍNEZ AMAYA
JEFE(A) DE LA DIVISIÓN DE ESTUDIOS DE POSGRADO E INVESTIGACIÓN

C.c.p. archivo

MAMA/lucy *

"RECONOCIMIENTO MULTIMODAL DE EMOCIONES ORIENTADAS AL APRENDIZAJE PARA TUTORES INTELIGENTES EN AMBIENTES ANDROID"

Tesis presentada por:

ING. HÉCTOR MANUEL CÁRDENAS LÓPEZ

Aprobada en contenido y estilo por:

Dr. Ramón Zatarain Cabada
Director de Tesis

Dra. María Lucía Barrón Estrada
Secretario

Dr. Héctor Rodríguez Rangel
Vocal -1

M.C. Gloria Ekaterine Peralta Peñuñuri
Vocal -2

M.C. María Aracely Martínez Amaya
Jefe(a) de la División de Estudios de Posgrado e Investigación

Dedicatoria

Esta tesis está dedicada a todas las personas que han tenido un impacto importante en mi vida, en el ámbito de formación académica, a mis maestros, a los maestros Roberto León Piña, José Ángel Alcaraz Vega y al Doctor Raúl Santiesteban Cos por su esfuerzo y dedicación en darme una formación profesional, de integridad, solidaridad y respeto, además de enseñarme a enorgullecerme y respetar mi trabajo profesional, al igual que a los doctores Ramon Zatarain Cabada, Lucia Barrón Estrada, Ricardo Rafael Quintero Meza y Hector Rodríguez Rangel y que me enseñaron la importancia del pensamiento crítico, la documentación y que me mostraron que lo mas importante es nunca rendirse con la tarea que tenemos enfrente. De la misma manera le dedico esta tesis a todos aquellos que han estado involucrados en mi desarrollo personal, a mis padres Hector Manuel Cardenas Cota y María Adelaida Lopez Rochin, por enseñarme los valores básicos de la convivencia y por nunca rendirse con buscar educarme en todos los aspectos posibles, a mi padre en lo académico y a mi madre en lo social, a mis amigos que me han acompañado siempre y han sido como una segunda familia, causándome mucha felicidad de saber que siempre cuento con ellos, los Ases con los que refine mis habilidades de convivencia, la Familia con los que mejore mi autoestima y mi autovaloración y los del cubículo quienes me han ayudado a refinar mi debate y la defensa de mis ideas.

Por ultimo y más importante, a mi compañera de vida Tzitzi Guerrero Gallardo con quien he tenido el gusto de compartir los últimos 8 años de mi vida, en lo personal y académico, muchísimas gracias por acompañarme hoy y siempre, nada de esto seria posible sin tu cariño, compañía, ayuda profesional y tu entendimiento, en resumen, eres la mejor.

Agradecimientos

Agradezco a los trabajadores del Tecnológico Nacional de México, campus Culiacán por su impecable tarea de mantener las certificaciones e instalaciones, a los docentes por su ardua labor de enseñanza y su dedicación.

Agradezco también al CONACYT por la oportunidad de brindar una beca de apoyo monetario que permitió ayudarme a salir adelante en los procesos económicos de mi formación académica.

Agradezco también al CIMAT zacatecas por la oportunidad de realizar una estancia de investigación y por permitirme tener contacto con profesionistas y académicos del área, que ayudaron a esclarecer algunas dudas durante las etapas experimentales de esta tesis.

Declaración de autenticidad

Por la presente declaro que, salvo cuando se haga referencia específica al trabajo de otras personas, el contenido de esta tesis es original y no se ha presentado total o parcialmente para su consideración para cualquier otro título o grado en esta o cualquier otra Universidad. Esta tesis es resultado de mi propio trabajo y no incluye nada que sea resultado de algún trabajo realizado en colaboración, salvo que se indique específicamente en el texto.

Héctor Manuel Cárdenas López.

Culiacán, Sinaloa, México, 2019

Resumen

En este trabajo se presenta una investigación de diferentes técnicas de fusión temprana de datos para el reconocimiento multimodal de emociones orientadas al aprendizaje con las modalidades de texto e imagen. Se utilizaron los modelos de dinámica cognitiva de D'Mello y el modelo de emociones de Russell para el alineamiento de las representaciones de tres conjuntos diferentes de datos. Los tres conjuntos de datos (corpus) fueron creados en trabajos de investigación anteriores, el primero de ellos fue realizado utilizando rostros de estudiantes y fue llamado EmotivInsight debido a la diadema electroencefalográfica utilizada para la etiquetación de las emociones de los usuarios. Este corpus presenta seis emociones discretas significativas para el proceso de aprendizaje. El segundo conjunto fue llamado EduSere por el sistema utilizado para la recolección de la información y presenta cinco emociones discretas para el proceso de aprendizaje. El último titulado SentiText presenta solamente polaridad de sentimiento en texto.

En este trabajo de investigación, se logró fusionar estos conjuntos de datos a través de las técnicas de fusión de representaciones y utilizarlos para llevar a cabo el reconocimiento multimodal de emociones orientadas al aprendizaje. Los reconocedores multimodales desarrollados en esta investigación fueron comparados con reconocedores unimodales utilizando los mismos conjuntos de datos, y se demostró que, dadas las condiciones experimentales utilizadas en este documento, la clasificación de emociones orientadas al aprendizaje de manera multimodal puede superar hasta en un 8% a los mejores reconocedores unimodales utilizados.

Palabras clave

AC: *Affective Computing*, Computación Afectiva

AI: *Artificial Intelligence*, Inteligencia Artificial

ANN: *Artificial Neural Network*, Red Neuronal Artificial

CNN: *Convolutional Neural Network*, Red Neuronal Convolucionada

DL: *Deep Learning*, Aprendizaje Profundo

ER: *Emotion Recognition,* Reconocimiento de Emociones

FM: *Fusion Methodology,* Metodología de Fusión.

ILE: *Intelligent Learning Environment*, Ambiente Inteligente de Aprendizaje

ITS: *Intelligent Tutoring System,* Sistema Tutor Inteligente

LSTM: *Long Short-Term Memory*, Redes de Gran Memoria de Corto Plazo

ML: *Machine Learning*, Aprendizaje Automático

MML: *Multimodal Machine Learning,* Aprendizaje Máquina Multimodal

NLP: *Natural Language Processing*, Procesamiento de Lenguaje Natural

SA: *Sentiment Analysis*, Análisis de Sentimiento

SERE: Sistema de Evaluación de Recursos Educativos

Índice general

Índice de Figuras

Índice de tablas

Capítulo 1

1.Introducción

En este capítulo se expone el planteamiento del problema, los objetivos generales y específicos del experimento de tesis, además de esto se plantea la hipótesis y se establece la estructura general de la tesis.

1.1. Planteamiento del problema

Las emociones están fuertemente ligadas a todas las actividades que realizan los seres humanos. Una actividad muy importante desde el inicio de los tiempos para el ser humano ha sido el proceso de aprendizaje, gracias a éste nuestros antepasados fueron capaces de adaptarse y sobrevivir a entornos hostiles a través de la transferencia de conocimiento.

Hoy en día se cuenta con una gran cantidad de herramientas que permiten automatizar el proceso de enseñanza y aprendizaje. Existe una diversa gama de aplicaciones que permiten a un estudiante obtener conocimiento acerca de un tema específico sin la necesidad de la interacción directa con un profesor. A estos sistemas se les conoce como Sistemas Tutores Inteligentes (STI).

Debido a que, dentro del proceso de aprendizaje, la capacidad cognitiva del alumno se ve directamente relacionada con el proceso cognitivo de la información, el uso de reconocedores de emociones automático dentro de un STI presenta una oportunidad importante para el desarrollo de sistemas que consideren la parte cognitiva y la parte afectiva del usuario.

Muchos trabajos han sido realizados para la detección de emociones en el aprendizaje en las últimas décadas. Algunos trabajos como los de (De Silva, Miyasato, and Nakatsu 1999) para detectar emociones en el rostro y de (Binali, Wu, and Potdar 2010) para detectar emociones en texto, han presentado muy buenos resultados en la detección de emociones en ambientes controlados, sin embargo debido a la gran cantidad de ruido que presentan algunas de las señales de ambientes no controlados, se ha demostrado que estos sistemas aún no se encuentran listos para su uso en el mundo real.

Muchas propuestas de solución se han realizado para resolver el problema de la falta de robustez de los clasificadores en el mundo real, una de ellas es el aprendizaje máquina multimodal, que en los últimos años ha encontrado una gran cantidad de momento en el ámbito académico y de investigación. Sin embargo, estas metodologías presentan grandes problemas a resolver como lo son el alineamiento y fusión de los datos o características en varios reconocedores, esto presenta un problema fundamental en la aplicación de modelos multimodales que permitan detectar las emociones de los usuarios.

1.2. Objetivos

1.2.1. Objetivo General

Crear un reconocedor multimodal de emociones orientadas al aprendizaje para su aplicación en Sistemas Tutores Inteligentes utilizando las modalidades de imagen y texto mitigando el problema de clasificaciones erróneas en una sola modalidad.

1.2.2. Objetivos Específicos

1) Diseñar y desarrollar metodologías para la aproximación semántica o alineamiento de los conjuntos de datos de imagen y texto.
2) Diseñar, desarrollar y evaluar diferentes arquitecturas de aprendizaje profundo para el reconocimiento de emociones multimodal.
3) Diseñar las metodologías para la implementación de reconocedores multimodales en un ambiente de aprendizaje.
4) Integrar el sistema multimodal de reconocimiento de emociones a un STI.
5) Implementar el ILE con el reconocedor dentro de una plataforma Android.

1.3. Hipótesis

El uso de técnicas de fusión temprana de datos para el reconocimiento multimodal de emociones producirá una mejora significativa en la precisión de los modelos de reconocimiento de emociones orientadas al aprendizaje.

1.4. Estructura de la tesis

Esta tesis está estructurada de la siguiente manera: en el capítulo 2 se expone un estudio del marco teórico, con los temas más relevantes para la investigación realizada, presentando temas como aprendizaje máquina, computación afectiva y ambientes inteligentes de aprendizaje.

En el capítulo 3 se presenta un estudio del estado del arte en materia de reconocedores multimodales de emociones y ambientes de aprendizaje con reconocimiento multimodal de emociones.

El capítulo 4 muestra el desarrollo del proyecto, a través de una descripción de los conjuntos de datos, técnicas de representación, sistemas multimodales y la metodología de aplicación de sistemas multimodales en ambientes de aprendizaje.

El capítulo 5 describe los experimentos realizados, las pruebas a los modelos multimodales y un caso de estudio de un reconocedor multimodal de emociones aplicado a un ambiente de aprendizaje.

El capítulo 6 presenta las conclusiones del trabajo, sus aportaciones y limitaciones y por último su trabajo a futuro.

Por último, se incluye una lista de trabajos consultados en la sección de Referencias.

Capítulo 2

2.Marco teórico

En esta sección se presenta un análisis de la literatura de los diferentes temas que comprende la tesis, los cuales son: El aprendizaje máquina, la computación afectiva, el reconocimiento multimodal de emociones y los ambientes inteligentes de aprendizaje. En esta sección se analiza a mayor profundidad el tema de reconocimiento multimodal de emociones por ser el tema de mayor impacto de la tesis.

2.1. Aprendizaje Máquina

La diferencia fundamental entre las computadoras y los seres humanos ha sido por un largo tiempo que los seres humanos son capaces de pensar, aprender de sus errores y mejorar en la realización de una tarea o en la solución de un problema. Los algoritmos computacionales tradicionales no tienen la capacidad de observar sus resultados y por lo tanto no pueden aprender de sus errores y mejorar su comportamiento. El campo del aprendizaje máquina busca asemejar el comportamiento de los seres humanos en máquinas, a través de la creación de algoritmos computacionales capaces de observar sus resultados y de mejorar su comportamiento aprendiendo a través de sus errores y de una mayor cantidad de entrada de datos, sin embargo, estos algoritmos están aún lejos de asemejar completamente el aprendizaje humano, pero son una excelente aproximación a la inteligencia artificial

Las técnicas de aprendizaje máquina crean modelos matemáticos a través del uso de datos de muestra normalmente mencionados como “datos de entrenamiento” con los cuales realizan predicciones y valoraciones sin la necesidad de ser explícitamente programados para realizar esta tarea (Bishop 2006).

Aprendizaje máquina es esencialmente una forma de estadísticas aplicadas con alto énfasis en el uso de computadoras para estimar funciones estadísticamente complicadas con bajo énfasis en probar intervalos de confianza entre estas funciones (Ian Goodfellow, Yoshua Bengio 2016).

El aprendizaje máquina hoy en día se utiliza de manera cotidiana, sin el conocimiento de muchas personas, en diferentes áreas como economía, ciencia, política, etc. Un ejemplo claro de esto es el algoritmo de aprendizaje de tendencias del usuario que presenta el servicio de transmisión de video Netflix, el cual puede predecir con gran exactitud el tipo de series o películas prefiere un usuario (Gomez-Uribe and Hunt 2015).

2.1.1. Aprendizaje máquina tradicional

Dentro del aprendizaje máquina tradicional existen diferentes técnicas y modelos utilizados para la creación de algoritmos de solución de problemas, a continuación, se muestran los más comunes en esta área.

Árboles de decisión

Los árboles de decisión son una técnica de clasificación y regresión que se encarga de dar un mecanismo de representación simple a ideas. Un árbol de decisión se construye iterativamente separando el conjunto de datos en diferentes clases o posibles decisiones existentes hasta que un criterio de paro es alcanzado. La forma de representación de los árboles de decisión permite observar los datos separados sistemáticamente debido a que su formato de árbol resulta fácil de comprender para el ser humano.

Uno de los primeros algoritmos de entrenamiento de árboles de decisión fue el dicotomizador iterativo 3 (ID3), y su sucesor el algoritmo C4.5 ambos desarrollados por Ross Quinlan en 1986 y 1993 (Quinlan 1986; Himani Sharma 2016).

En los árboles de decisión los nodos pueden ser separados en nodos y ramas. Los nodos pueden ser divididos en nodos raíz, nodos internos, y nodos finales también llamados hojas. El nodo raíz representa la entrada al árbol de decisión y no tiene ramas de entrada. Los nodos internos tienen exactamente una rama de entrada y dos ramas de salida. Estos nodos contienen las pruebas a realizar basada en los datos que se tienen. Por ejemplo, dichas pruebas podrían preguntar "¿Qué día en el atributo temperatura es mayor que 35°C?". Los nodos hoja consisten en una respuesta al problema de decisión, que puede representar un valor numérico en regresión o una clase concreta en clasificación. Por ejemplo, un problema de decisión podría ser la pregunta de si vas a salir a caminar hoy o no, siendo las clases si o no. Los nodos hoja no tienen ramas de salida y contienen únicamente una rama de entrada. Las ramas representan las decisiones tomadas a partir de los nodos anteriores.

Dado un nodo cualquiera n, todos los nodos siguientes que son separados a n por exactamente una rama son llamados hijos de n, mientras n es llamado padre de estos nodos. En la Figura 2-1 se presenta un ejemplo de árbol de decisión.

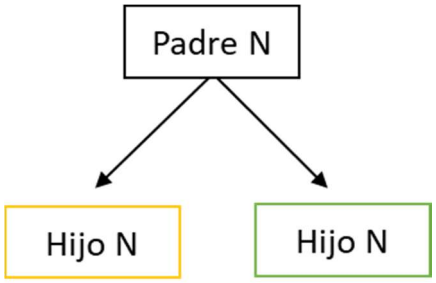


Figura 2-1. Ejemplo de un árbol de decisión.

Entrenar un árbol de decisión tiene el objetivo de buscar patrones en un conjunto de datos de entrenamiento para predecir el valor de un atributo objetivo, utilizando el resto de los atributos de los datos de entrada para realizar la predicción.

El conjunto de datos de entrenamiento se muestra en la Ecuación 1:

$$(\vec{x}, Y) = (x_1, x_2, \cdots, x_n, \boldsymbol{Y}) \quad (1)$$

Donde Y representa al atributo objetivo y $\vec{x}$ representa al vector que contiene n atributos donde n es el número de atributos del conjunto de datos.

Para poder entrenar un árbol de decisiones, se requiere que el conjunto de datos de entrenamiento contenga el atributo objetivo, el resto de los atributos de entrada, un criterio de separación y un criterio de paro. En un nodo dado, el criterio de separación calcula el valor de todos los atributos. Este valor representa una medida de la cantidad de información que se gana al separar el nodo utilizando este atributo. Después, el mejor valor de todos los atributos se toma y el nodo se divide en los posibles resultados de ese atributo. En este punto, el proceso de encontrar las mejores divisiones entre los atributos se aplica de manera recursiva para generar todos los subárboles hasta alcanzar el criterio de paro.

Algunos criterios de paro comunes son:

- Se alcanzó la altura máxima del árbol.
- El número de operaciones en el nodo es menor al mínimo requerido.

- El mejor criterio de división no sobrepasa un cierto límite en términos de información obtenida.

Entrenar árboles de decisión con este proceso automatizado puede resultar en árboles de decisión muy largos, con secciones con muy poca relevancia para la clasificación. Además de esto los árboles tienden a presentar sobreajuste (*overfitting* en inglés), lo que significa que están demasiado adaptados a los datos de entrenamiento. Esto resulta en un mal desempeño cuando son aplicados a datos diferentes a los utilizados para el entrenamiento. Por lo tanto, se desarrolló una técnica llamada poda (*prunning* en inglés). La cual tiene como objetivo remover las partes menos productivas de un árbol de decisión, las partes basadas en el uso de datos ruidosos o partes que presentan *overfitting*. Este proceso es importante debido a que los datos del mundo real contienen errores o ruido.

Máquinas de soporte vectorial

Las Máquinas de Soporte Vectorial (SVM por sus siglas en inglés) pertenecen al área del aprendizaje supervisado y por lo tanto requieren de etiquetas para clasificar nuevos datos no utilizados para el entrenamiento. Las SVM son modelos que se crean a partir de algoritmos de aprendizaje que se utilizan para la clasificación o regresión de datos analizados. Dado un conjunto de datos para el entrenamiento, todos ellos con una de dos posibles etiquetas, un algoritmo de entrenamiento de SVM construye un modelo que asigna a nuevos datos una categoría u otra. Un modelo de SVM es una representación de ejemplos como puntos en un espacio mapeado para que los ejemplos de las diferentes categorías estén separados a través de una línea en una distancia tan grande como se pueda unos de otros. Los nuevos datos son entonces mapeados en el mismo espacio y predichos entre una categoría u otra (Kecman 2001).

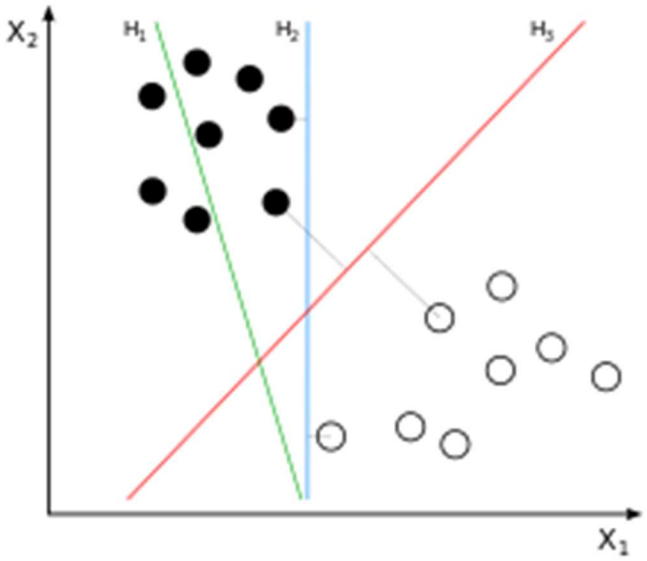


Figura 2-2. Ejemplo de máquina de soporte vectorial separando dos diferentes clases.

Redes bayesianas (BAYES)

Las redes bayesianas son grafos dirigidos (nodos y conexiones dirigidas entre esos nodos que simbolizan dependencia entre ellos). Estos son modelos gráficos acíclicos probabilísticamente dirigidos. Cada modelo representa un atributo de interés dada una tarea. La red bayesiana más básica es nombrada *Naive Bayes* y la razón por la cual es nombrada *Naive* es que esta red asume que no existe dependencia entre los atributos. Esto es raramente el caso dentro de una aplicación de clasificación o regresión real, por lo tanto, estos algoritmos tienden a obtener los peores resultados cuando son comparados con algoritmos más sofisticados. Las redes bayesianas normales utilizan los datos conocidos para estimar la dependencia entre atributos y clases, utilizando esta información para calcular la probabilidad de los posibles diferentes resultados de eventos futuros o clases. Estas automáticamente aplican el teorema de Bayes para problemas complejos y por lo tanto son capaces de obtener conocimiento acerca del estado de los atributos y sus relaciones (Dodge 2008). El teorema de Bayes se presenta en la Ecuación 2.

$$P(B|A) = \frac{P(B|A) * P(A)}{P(B)} \quad (2)$$

Donde

- A y B son eventos

- *P(X)* es la probabilidad de que el evento X ocurra
- *P(X|Y)* es la probabilidad condicional de que el evento X ocurra si el evento Y se conoce como verdadero

Cada nodo en el grafo esta etiquetado con una distribución de probabilidad que define el efecto del nodo padre a los nodos hijos. Un ejemplo de una red bayesiana se muestra en la Figura 2-2.

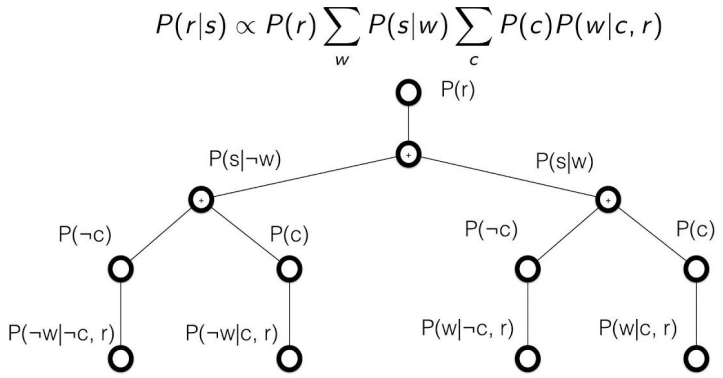


Figura 2-3. Ejemplo de una red Bayesiana.

Aprendizaje basado en instancias (kNN)

El aprendizaje basado en instancias describe el proceso de la solución de problemas basado en soluciones similares de problemas conocidos, también se le conoce como aprendizaje de vecinos más cercanos. Cada sistema de aprendizaje basado en instancias requiere de cumplir ciertos parámetros.

- Una función de distancia que mida la similitud entre los problemas o entradas de datos. Esto es necesario para medir la proximidad que existe entre los datos cercanos de un nuevo problema.
- Un número de vecinos que considerar cuando se busque resolver un nuevo problema o se inserte un nuevo dato.
- Una función de peso que permite cuantificar aún más los vecinos encontrados lo cual permite mejorar la predicción y la calidad del aprendizaje.
- Un método de evaluación que describe una función de cómo usar los vecinos encontrados para resolver dicho problema.

El aprendizaje basado en instancias o vecinos más cercanos es parte de los métodos de aprendizaje perezosos, lo que significa que no existe una necesidad de procesar los datos antes de dárselos al sistema. Estos métodos existen en contraste con los ansiosos de aprender como los árboles de decisión que requieren de un preprocesamiento de estructuración de los datos para su entrenamiento (Sammut and Webb 2010). En la Figura 2-4 se puede observar un ejemplo de un sistema de clasificación de vecinos más cercanos, que utiliza dos diferentes clases para determinar un nuevo ejemplo para clasificar, utilizando k como el número de vecinos para compararlo.

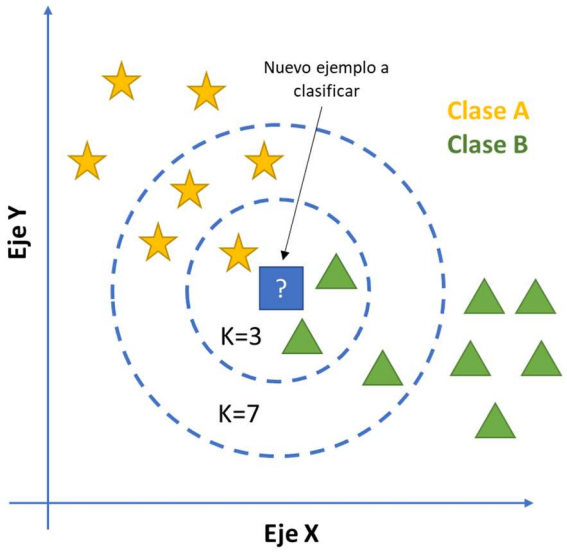


Figura 2-4. Ejemplo de vecinos más cercanos aplicados a dos clases.

Redes neuronales artificiales

Las Redes Neuronales Artificiales (ANN por sus siglas en inglés) son herramientas de modelado computacional que pueden ser definidas como estructuras comprimidas densamente interconectadas, de elementos adaptativos de procesamiento simple llamados (neuronas o nodos), que son capaces de realizar operaciones de computación paralela masivas para el procesamiento y representación de conocimiento (Hecht-Nielsen 1988)(Schalkoff 1997). Las ANN están basadas en la idea de replicar el comportamiento del cerebro biológico, sin embargo, no logran replicar la operación de este completamente, pero hacen uso de lo que se conoce como la funcionalidad de las redes neuronales biológicas para la resolución de problemas complejos.

Una ANN está compuesta de neuronas o nodos, conexiones entre estos nodos con pesos los cuales pueden ser adaptados durante la etapa de aprendizaje de la red y una función de activación en cada neurona que define el valor de salida dependiendo de los valores de entrada en esa neurona. Cada red neuronal se compone de tres tipos de capa, la primera es la capa de entrada donde se ingresan la información de medios externos a través de entrada de datos, la capa de salida que presenta la clasificación o valor esperado de la red neuronal y una o más capas escondidas que conectan la capa de entrada con la capa de salida. El valor de entrada de cada neurona es calculado como la sumatoria de todas las neuronas anteriores multiplicadas por el peso respectivo de la interconexión entre dichas neuronas.

Los valores de salida de cada nodo son calculados utilizando todos los valores de entrada a para calcular una función predefinida llamada función de activación que normalmente es la misma para todas las neuronas de una red. En la Figura 2-5, se muestran los tipos más comunes de función de activación para redes neuronales.

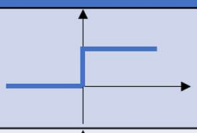
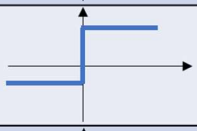
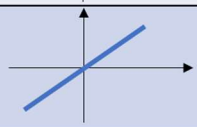
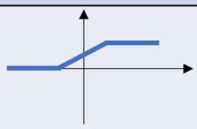
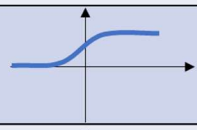
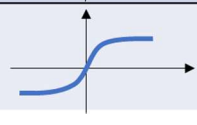
Función de Activación	Ecuación	Ejemplo	Gráfica 2D
Paso Unitario	$\Phi(z) = \begin{cases} 0, & z < 0 \\ 0.5, & z = 0 \\ 1, & z > 0 \end{cases}$	Variable Perceptron	
Signo	$\Phi(z) = \begin{cases} -1, & z < 0 \\ 0, & z = 0 \\ 1, & z > 0 \end{cases}$	Variable Perceptron	
Lineal	$\Phi(z) = z$	Adaline, Regresión linear	
Dependiente de piezas	$\Phi(z) = \begin{cases} 1, & z \geq \frac{1}{2} \\ z + \frac{1}{2}, & -\frac{1}{2} < z < \frac{1}{2} \\ 0, & z \leq -\frac{1}{2} \end{cases}$	Máquina de soporte de vectores	
Logística	$\Phi(z) = \frac{1}{1 + e^{-z}}$	Regresión logística, Multicapa NN	
Tangente Hiperbólica	$\Phi(z) = \frac{e^{z} - e^{-z}}{e^{z} + e^{-z}}$	Multicapa NN	

Figura 2-5. Tipos más comunes de funciones de activación en redes neuronales artificiales (Jain, Jianchang Mao, and Mohiuddin 1996).

Las ANN pueden ser divididas en dos tipos dependiendo de su método de entrenamiento.

- **Redes neuronales prealimentadas (por sus siglas en ingles DFN):** estas redes neuronales son en las que las conexiones entre las neuronas no forman un ciclo, es decir no tienen una retroalimentación. Esto significa que los datos fluyen en un solo sentido, desde los nodos de entrada pasando por los nodos de las 0 a n capas ocultas y a los nodos de la capa de salida. No hay información regresada para readaptar el sistema.
- **Redes neuronales recurrentes (por sus siglas en ingles RNN):** estas redes neuronales si forman un ciclo entre las neuronas, obteniendo así una retroalimentación. Estas redes neuronales son capaces de reutilizar los datos de entrada para readaptarse en las últimas etapas del proceso de aprendizaje.

La Figura 2-6 muestra un ejemplo de una ANN con tres nodos en la capa de entrada cuatro nodos en la capa escondida y tres nodos en la capa de salida.

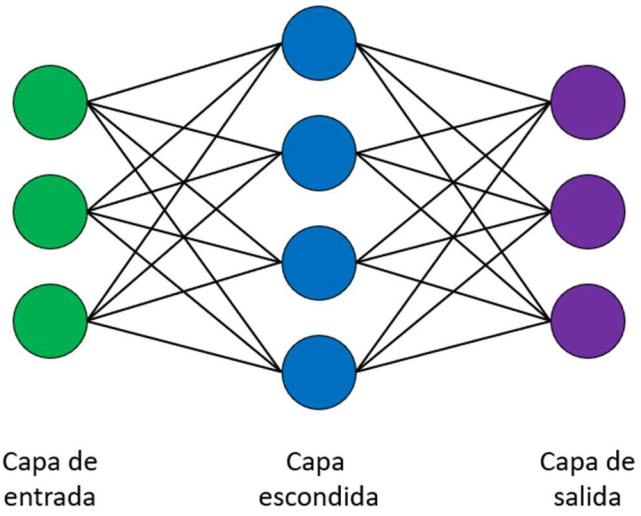


Figura 2-6. Ejemplo de una red neuronal artificial.

En un escenario de aprendizaje supervisado, las ANN pueden ser entrenadas utilizando un algoritmo de propagación hacia atrás (*back propagation*) el cual se encarga de reajustar los pesos de las conexiones en la red neuronal basada en los porcentajes de error local.

La propagación hacia atrás en ANN describe el proceso de utilizar el error local de la red para reajustar los pesos de las conexiones de las neuronas hacia atrás a través de la red neuronal. Esto significa que después de realizar la predicción de un conjunto de datos, el

valor obtenido se compara con el valor esperado para calcular el error de la predicción. Este error después es utilizado para reajustar los pesos de las conexiones de la red comenzando por la capa de salida en dirección hacia la capa de entrada.

Para poder entrenar una ANN es importante entender primero los parámetros principales que pueden ser optimizados durante el proceso de aprendizaje (Erb 1993).

- La **tasa de aprendizaje** (*learning rate*) especifica que tan rápido se realiza el proceso de aprendizaje. El parámetro comúnmente está en valores entre el 0 y 1, y es multiplicado por el error para cada valor de salida. Por lo tanto, una taza de aprendizaje de 0 no permite que exista adaptación alguna al error en la red, mientras que el valor de 1 puede llegar a permitir que los saltos sean tan grandes entre las adaptaciones que nunca se encuentre un valor óptimo para la predicción debido a la oscilación entre los valores del error. Sin embargo, si la tasa de aprendizaje es muy pequeña los saltos adaptativos de la red pueden ser tan pequeños que tome demasiado tiempo a la red y se puede quedar atorada en una máxima local. Para poder encontrar la configuración correcta de la tasa de aprendizaje se debe de agregar un valor de decadencia. Este parámetro asegura que durante los primeros ciclos de aprendizaje la red neuronal aprenda rápidamente evitando la máxima local y conforme se aproxima a su configuración óptima la tasa de aprendizaje disminuya evitando así las oscilaciones.
- El **criterio de parada**, este criterio funciona de manera similar a los árboles de decisión, es un límite del error, que una vez que se obtiene pasa el proceso de aprendizaje termina.

El proceso de entrenamiento de una ANN normalmente se compone de cuatro diferentes etapas (Negnevitsky 2002):

- **Inicialización:** Donde todos los pesos y los niveles límites son definidos dentro de la red de manera aleatoria distribuidos dentro de un rango pequeño.
- **Activación:** Se activa la propagación de la red neuronal a través de aplicar una entrada de datos de entrada y los datos de salida esperados obteniendo los valores de salida reales de la red y el error.

- **Entrenamiento de pesos:** Se actualizan los pesos en función de la propagación hacia atrás de los errores asociados con la salida de la ANN.
- **Iteración:** Se realiza una iteración nueva y se repite el proceso hasta que el criterio de parada es satisfecho.

2.1.2. Aprendizaje profundo

Los métodos convencionales de aprendizaje máquina están limitados en su habilidad para procesar datos naturales, en su forma cruda, es decir, sin pre procesar. Por décadas construir un reconocedor de patrones o un sistema de aprendizaje máquina requería de ingeniería meticulosa y considerable dominio experto del tema, para diseñar un extractor de características que pre procesara la información cruda. Por ejemplo, los pixeles dentro de una imagen, en representaciones internas adecuadas o vectores de características con los cuales los sistemas de aprendizaje, comúnmente clasificadores, pudieran detectar y clasificar estas características de entrada.

El aprendizaje de representación es un conjunto de métodos que permite el alimentar a un algoritmo de aprendizaje máquina con datos crudos y automáticamente definir representaciones que se necesitan para la clasificación. Los métodos de aprendizaje profundo son métodos basados en el principio de aprendizaje de representación con múltiples niveles de representación, obtenidos de componer módulos no lineales simples tal que cada uno transforme la representación un nivel, desde los datos crudos hasta su ultimo nivel de abstracción (LeCun, Bengio, and Hinton 2015).

Por lo tanto, el aprendizaje profundo moderno, provee un *framework* muy poderoso para el aprendizaje supervisado. Al agregar una mayor cantidad de capas con más unidades en cada capa, una red profunda puede representar funciones de alta complejidad. La mayoría de las tareas que consisten en mapear un vector de entrada a un vector de salida, pueden ser logradas por una red profunda, dado un modelo suficientemente grande y un conjunto de datos para el entrenamiento lo suficientemente grande con un buen etiquetado. Por otro lado, las tareas que resultan complejas, que requieren de una toma de decisiones más compleja para una posible solución, aun no son abordables a través del aprendizaje profundo (Ian Goodfellow, Yoshua Bengio 2016).

Redes Neuronales Convolucionadas

Las Redes Neuronales Convolucionadas (CNN por sus siglas en inglés) (Le Cun 1989) son un tipo especializado de redes neuronales para el procesamiento de datos que es conocida por tener una topología parecida a cuadriculas. Este tipo de redes ha tenido excelentes resultados en la solución de problemas de clasificación y categorización de información simple en aplicaciones del mundo real. El nombre red neuronal convolucional indica que este tipo de redes aplican una técnica matemática conocida como convolución.

Las CNN son muy similares a las ANN con la diferencia de que acomodan sus neuronas en 3 dimensiones, alto, ancho y largo. Una neurona dentro de una capa está conectada a una región de la capa anterior, llamado campo de percepción, y no está completamente conectada como una ANN.

La arquitectura de las CNN consiste en diferentes tipos de capas secuenciales, algunas de ellas pueden estar repetidas. A continuación, se describen algunos de los tipos de capas de una CNN.

Capas de convolución

Estas son las capas fundamentales de las redes neuronales convolucionales, son en las que se lleva a cabo la operación de convolución. En su forma más básica, la convolución es una operación sobre dos funciones diferentes a un valor dado (Ian Goodfellow, Yoshua Bengio 2016).

Las capas de convolución toman ventaja del hecho que sus entradas (ej. Imágenes, mapas de características) presentan muchas relaciones espaciales. Estas relaciones espaciales permiten el uso de filtros para la extracción de características de las imágenes o mapas de características de entrada. Por lo tanto, las capas convolucionadas aprenden un conjunto N_k de filtros $F = f_1, f_2..., fNk$, que son convolucionados espacialmente con una entrada x, para producir un conjunto N_k *2D* de mapas de características z, el cálculo de los mapas de características z esta dado por la Ecuación 3.

$$z_k = f_k * x \quad (3)$$

Donde $*$ representa la operación de convolución. Cuando un filtro presenta una relación directa con una región de la imagen de entrada, los mapas de características resultantes de las operaciones de convolución presentan una fuerte relación (ej. Una imagen girada presenta un giro similar en la respuesta dentro del mapa de características). A diferencia de las capas lineales convencionales, los pesos de las capas de convolución son compartidos por toda una imagen, reduciendo el número de parámetros necesarios para realizar cálculos por imagen (Le Cun 1989).

Una representación de la convolución sobre una imagen se muestra en la Figura 2-7.

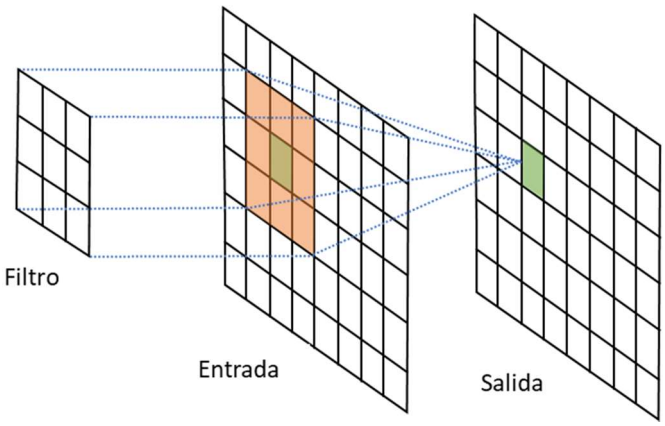


Figura 2-7. Una convolución sobre una imagen.

Los filtros de la capa de convolución son operaciones que se realizan espacialmente sobre la imagen o el mapa de características. Estas operaciones pueden ser vistas como multiplicaciones por matrices. En dos dimensiones, una matriz circulante de doble bloque corresponde a una convolución, la cual corresponde usualmente a matrices muy dispersas (matrices cuyo contenido equivale mayormente a cero). Esto se debe a que el filtro es usualmente más pequeño que la imagen de entrada. Cualquier algoritmo de redes neuronales que trabaje con multiplicación de matrices y que no dependa de propiedades especifica de estas matrices puede trabajar con convolución, sin requerir que el algoritmo cambie su estructura fundamental (LeCun, Bengio, and Hinton 2015).

Capas de agrupación (*pooling*)

Esta capa de las CNN consiste en tres etapas. En su primera etapa, la capa realiza varias convoluciones en paralelo para producir un conjunto de activaciones lineales. En su segunda etapa, cada activación lineal se evalúa con funciones de activación no lineales. Esta etapa es

comúnmente denominada la etapa de detección. En la tercera etapa se utiliza una función de agrupación para modificar la salida a la siguiente capa (Ian Goodfellow, Yoshua Bengio 2016).

Una función de agrupación remplaza a la salida de la red en una cierta locación por una suma estadística de las salidas cercanas. Por ejemplo, la operación de agrupación de máximos (*Max pooling*) reporta la salida en una vecindad rectangular. Otras funciones de agrupación populares incluyen el promedio de una vecindad rectangular, la L^2 de una vecindad rectangular, o un promedio basado en la distancia con el pixel central (Zhou and Chellappa 1988).

El uso de la capa de agrupación ayuda a convertir la representación en algo aproximadamente invariante a pequeñas transiciones de la entrada, reduciendo las dimensiones del mapa de características con la Ecuación 4.

$$p_R = P_{i \in R}(z_i) \quad (4)$$

Donde P es una función de agrupación, en una región de pixeles R, de un mapa de características Z. La agrupación de máximos es la técnica más comúnmente utilizada debido a esta etapa no cancela elementos negativos, y evita la ofuscación de la activación y gradientes a través de la propagación de la red.

Esta capa está definida por su función de agrupación, sus dimensiones de altura y anchura del área donde se aplica, y las propiedades de su convolución (ej. *Padding, stride*), un ejemplo de este tipo de capa se muestra en la Figura 2-8.

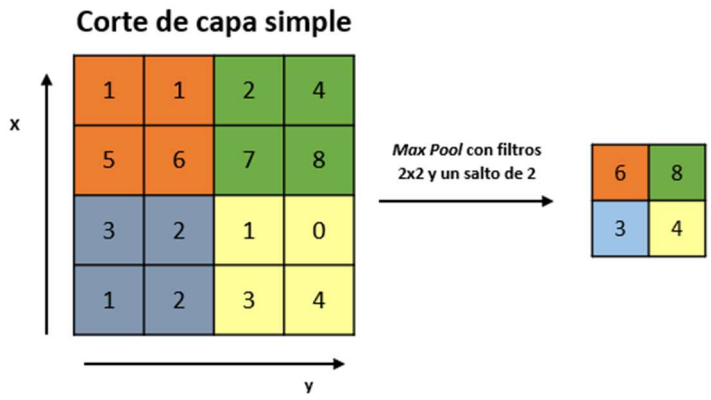


Figura 2-8. Ejemplo de una capa de *pooling*.

Capa de normalización de lotes (*Batch Normalization*)

Esta capa es una capa que aplica un algoritmo computacional de normalización de los datos (cambiando las entradas a promedio cero y varianza unitaria) para lograr que todas las entradas de cada capa tenga características comparables. Esta capa asegura una alta velocidad de aprendizaje (Ioffe and Szegedy 2015).

Capa de *dropout*

Esta capa es una capa utilizada para mejorar el problema de (*overfitting*), su función es eliminar neuronas durante el entrenamiento de la red neuronal, con el fin de mejorar la generalización, forzando a diferentes neuronas durante el proceso de aprendizaje a aprender el mismo concepto.

Redes de gran memoria de corto plazo (por sus siglas en ingles LSTM)

Las redes neuronales de gran memoria de corto plazo son un tipo de arquitectura de redes neuronales recurrentes que utilizan la idea de tener neuronas con auto bucles, para producir caminos por los cuales el gradiente de aprendizaje puede fluir por largas duraciones de tiempo (Hochreiter and Urgen Schmidhuber 1997). Estos auto bucles cuentan con compuertas (neuronas adicionales) que permiten la condicionarlos dependiendo del contexto, permitiendo que la escala de tiempo de integración pueda ser cambiada dinámicamente (Gers, Schmidhuber, and Cummins 2000). Esto significa que aun para una LSTM, con parámetros fijos, la escala de tiempo en la integración puede cambiar basada en la secuencia de entrada,

debido a que las constantes de tiempo son salidas del mismo modelo. Este tipo de redes ha encontrado un gran auge en el reconocimiento de texto, habla, escritura, traducciones, etc.

El diagrama de una celda de una LSTM se muestra en la Figura 2-9.

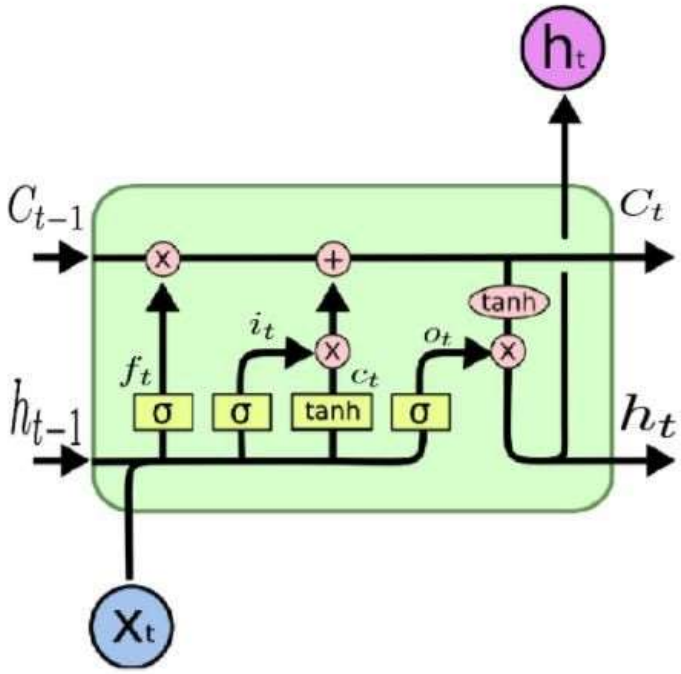


Figura 2-9. Imagen de una celda de LSTM (Hochreiter and Urgen Schmidhuber 1997).

2.1.3. Evaluación del aprendizaje máquina

Una parte fundamental del aprendizaje máquina, es el problema de cómo un programa computacional evalúa si los resultados que obtiene son apropiados y cuales contienen errores. Existen ejemplos de algoritmos computacionales que no presentan este problema, como un algoritmo computacional que trata de predecir si una persona visualizará un video o no. Los datos de entrada de este usuario serán grabados después de que haya visto un video o no, y esto definirá el desempeño de dicho algoritmo. El problema principal existe en escenarios de investigación más complicados, donde no existe un acceso directo a los datos del mundo real, un ejemplo de esto es el reconocimiento de emociones. Esto requiere de esfuerzo adicional humano para evaluar el nivel de reconocimiento de emociones definido a través de clases, que pueden ser comparadas con los resultados del algoritmo. Normalmente la evaluación de estos algoritmos se realiza a través de la separación de datos en datos en un conjunto de

"RECONOCIMIENTO MULTIMODAL DE EMOCIONES ORIENTADAS AL APRENDIZAJE PARA TUTORES INTELIGENTES EN AMBIENTES ANDROID"

Tesis presentada por:

ING. HÉCTOR MANUEL CÁRDENAS LÓPEZ

Aprobada en contenido y estilo por:

Dr. Ramón Zatarain Cabada
Director de Tesis

Dra. María Lucía Barrón Estrada
Secretario

Dr. Héctor Rodríguez Rangel
Vocal -1

M.C. Gloria Ekaterine Peralta Peñuñuri
Vocal -2

M.C. María Aracely Martínez Amaya
Jefe(a) de la División de Estudios de Posgrado e Investigación

entrenamiento y un conjunto de evaluación. El modelo es entrenado con el primero y posteriormente se utiliza el segundo para medir los indicadores de desempeño para medir la calidad del modelo.

Uno de los problemas más comunes en el aprendizaje máquina se encuentra en el acceso a datos de entrenamiento y de prueba limitados. Por lo tanto, el problema de sobreajuste (*overfitting*) puede ser muy común en estos modelos. Para evitar este problema, uno de los métodos más utilizados es la validación cruzada. La validación cruzada describe el proceso de separar el conjunto completo de datos en *X* partes utilizando cada una de ellas secuencialmente como el conjunto de validación, mientras se combinan los otros para formar un conjunto de entrenamiento. Después, todos los indicadores de desempeño son promediados.

No existe un método perfecto de evaluación de un modelo, ya que cada uno tiene sus ventajas y desventajas, sin embargo, existen factores importantes para evaluar el desempeño de un programa de aprendizaje máquina como se muestran a continuación: (Powers and W 2011)

- **Tasa de clasificación errónea (*Missclassification rate*)** describe la cantidad relativa de información falsamente clasificada en un conjunto de datos. Si y'_i es la predicción a un dato *i*, y y_i es la etiqueta real, la tasa de clasificación errónea puede ser definida como se muestra en la ecuación 5:

$$misc_n = \frac{1}{n} * \sum_{i}^{n} (y_i \neq y'_i) \qquad (5)$$

  El principal problema con la tasa de clasificación errónea es el hecho de que depende altamente de la cantidad de etiquetas que se tengan, y la distribución de los datos dados esas etiquetas. Por ejemplo, lograr tasa de clasificación errónea del 0.01% puede parecer muy prometedor sin contexto, pero un ejemplo en donde el 99% del conjunto de datos esta dado con una clase A y el 1% con una clase B esto no es muy difícil de lograr.

  De la misma manera, una tasa de clasificación errónea del 20% representa un modelo mejor para la clasificación de 3 clases que un modelo similar con la misma tasa de error en la clasificación de 2 clases.

- **Benchmarking** describe el proceso de comparar el valor de un indicador a un valor de referencia estandarizado, para darle más fuerza a un argumento o frase.
- **Valor de precisión** describe la cantidad de clasificaciones correctas que presenta el modelo, a través de todas las clasificaciones realizadas.
- **Valor de pérdida** describe el resultado de una función que describe que tan bien un algoritmo computacional crea un modelo que describe un conjunto de datos. Las funciones más comunes de medición de perdida son: *MSE*, *quadratic loss* y *L2 loss*.

  Matriz de confusión Es una de las mejores aproximaciones para ilustrar el desempeño de un algoritmo de aprendizaje máquina. También se le conoce como tabla de contingencia, que distingue entre verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos en una predicción. Un ejemplo de una matriz de confusión se puede observar en la Figura 2-10. La principal desventaja de una matriz de confusión es que requiere de interpretación humana.

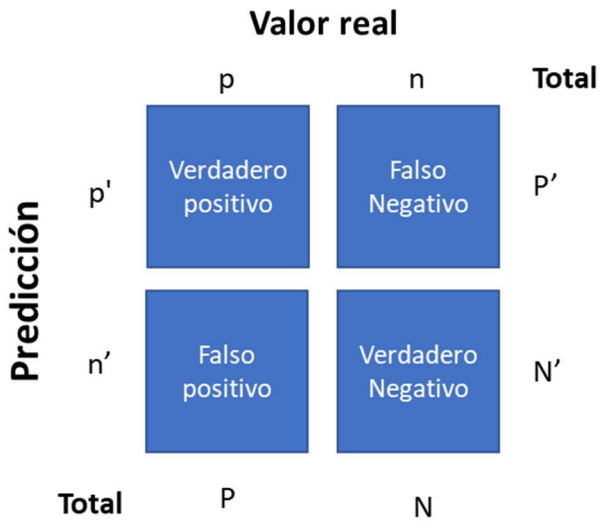


Figura 2-10. Ejemplo de matriz de confusión.

2.2. Computación afectiva

La computación afectiva es una rama de la inteligencia artificial que trata el diseño de sistemas y dispositivos que sean capaces de reconocer, interpretar, procesar y simular emociones humanas (Picard 1997).

Dentro de la computación afectiva, las emociones pueden ser divididas en dos categorías:

- **Emociones primarias:** Emociones instintivas innatas del ser humano. Estas comparten la mayor cantidad de similitudes independientemente del contexto cultural.
- **Emociones secundarias:** Estas emociones se obtienen a través de las diferentes interacciones sociales y los procesos cognitivos. Debido a lo anterior, cambian de acuerdo con el contexto cultural.

Dado que los seres humanos nos comunicamos con otros usando la comunicación verbal y la no verbal, es posible realizar la clasificación de emociones a través de los cambios fisiológicos en la comunicación humana, es decir, podemos observar emociones a través de las diferentes modalidades como el rostro, el texto escrito, la entonación de la voz, posición corporal, a través de los pequeños cambios en alguna de estas modalidades. El uso de modalidades permite separar la computación afectiva a través del medio de comunicación o modalidad clasificar emociones.

Entre los medios más famosos de reconocimiento de emociones, se encuentra el reconocimiento de emociones en rostro, en voz y texto.

2.2.1. Reconocimiento de emociones en rostro

El rostro presenta un rol importante en la comunicación de los seres humanos. Las expresiones faciales y los gestos presentan información no verbal que contribuye a la comunicación del ser humano. El rostro a su vez permite entender el estado emocional de una persona a través de sus diferentes expresiones volviéndolo así una herramienta fundamental para la comunicación de emociones y sentimientos de una persona en un dialogo o interacción.

La hipótesis universal declara que la percepción y expresión de emociones faciales son idénticas independientemente del origen de las personas o su linaje cultural. El trabajo original que estudio las expresiones faciales y su consecuencia fue escrito por Charles Darwin (Darwin, M.A., and F.R.S. 1872). Darwin decía que las expresiones faciales son innatas, y por lo tanto no pueden ser aprendidas, ya que estas tienen una función evolutiva para la sobrevivencia. Paul Ekman (Ekman 1992) realizó un estudio observacional para encontrar si

las expresiones del rostro de las emociones eran independientes del contexto cultural, y encontró que efectivamente como Darwin decía, existe una universalidad relativa de las emociones básicas del ser humano. Este estudio fue realizado con una tribu aislada de nueva guinea donde se observaron los mismos signos de expresiones faciales comparados con personas civilizadas.

En base a estas teorías se acuño la idea de que una máquina, a través del reconocimiento de patrones era capaz de reconocer emociones en seres humanos.

En 1997 Paul Ekman (Ekman and Rosenberg 1997) desarrolló un sistema para describir las emociones a través de la actividad muscular en el rostro utilizando una serie de unidades de acción (AU) al cual llamó Facial Acting Coding System (FACS).

El FACS permitió a los primeros modelos de aprendizaje máquina analizar el rostro de personas y detectar emociones.

Las técnicas de extracción de características del rostro para detección de emociones en el rostro pueden ser divididas en dos vertientes: características basadas en apariencia y características basadas en geometría. Las características basadas en apariencia describen la textura del rostro causada por la expresión, como lo son arrugas y hoyuelos. Las características basadas en geometría describen las formas de la cara a través de componentes como lo son el rostro, la boca y las cejas (Valstar et al. 2012).

Debido a la capacidad de extraer características de los modelos de aprendizaje profundo, la extracción de características manual ha comenzado a ser reemplazada por técnicas automáticas de extracción de características del rostro, sin embargo, se han realizado estudios que relacionan las AU con las características extraídas de modelos de aprendizaje profundo (Khorrami, Paine, and Huang 2015).

2.2.2. Reconocimiento de emociones en voz

El medio clásico de la comunicación entre los seres humanos es el uso de la voz. Desde que el lenguaje fue creado, ha buscado comunicar ideas complejas que han permitido tener una herramienta poderosa de organización en las diferentes etapas evolutivas del ser humano (Darwin, M.A., and F.R.S. 1872).

El reconocer el movimiento de gestos en el sonido es una manera de obtener información en las interacciones humanas. En la comunicación verbal existen varios componentes, la comunicación no es únicamente las palabras que decimos, sino también la serie de propiedades acústicas que le brindamos a la voz que transmiten connotaciones emotivas.

Las emociones están presentes en cada etapa de la comunicación verbal del ser humano, y debido a esto pueden ser clasificadas. Una herramienta muy útil para la clasificación de emociones en voz son los micrófonos, ya que estos traducen las propiedades acústicas en valores de voltaje que permiten darle características físicas a una frase que la representen. Estos voltajes se utilizan para crear series de tiempo que permiten observar el comportamiento de la voz durante la comunicación verbal de una idea, y gracias a esto se han creado diversos sistemas de inteligencia artificial que analizan, al igual que en las emociones en el rostro, las diferentes tonalidades y volúmenes de voz, como imágenes representativas de estas series de tiempo (Popova, Rassadin, and Ponomarenko 2018).

2.2.3. Reconocimiento de emociones en texto

La detección de emociones en lingüística computacional es el proceso de identificar emociones discretas en un texto. Determinar emociones en el texto es muy complejo y altamente dependiente del contexto, esto es por varias razones, por ejemplo, la sensibilidad a múltiples circunstancias personales y contextuales en la escritura del texto, además de las expresiones humanas contienen más de una emoción y por lo tanto es difícil discernirlas unas de otras en un solo texto.

Por lo tanto, para la detección de emociones en el texto se toma una gran consideración en el contexto de la escritura (Seyeditabari, Tabari, and Zadrozny 2018). La mayoría de los estudios de reconocimiento de emociones en el procesamiento natural del lenguaje utilizan tres diferentes técnicas:

- **Texto etiquetado** El texto etiquetado como su nombre lo implica, utiliza grandes bases de datos con texto representativo de las clases utilizando etiquetas en los documentos completos para realizar el entrenamiento.
- **Lexicón de emociones** Los lexicones de emociones a diferencia del texto etiquetado, utilizan los valores de las emociones representadas en cada palabra utilizada, de

manera que se calcula la emoción en un documento a través de medir las diferentes emociones dentro de cada palabra que compone el documento.

- **Embebido de palabras** Esta es una nueva técnica basada en la distribución semántica de la información. Basada en la idea de que la información que aparece en un corpus es semánticamente similar. En estos métodos cada palabra es representada como un vector en un espacio N-dimensional, llamado espacio vectorial y de alguna manera la distancia entre estos vectores representa las diferencias semánticas entre estas palabras.

2.2.4. Reconocimiento multimodal de emociones

El mundo que nos rodea envuelve múltiples modalidades, nosotros podemos observar objetos, oír sonidos, sentir texturas, oler olores, etc. En términos generales una modalidad se refiere a la manera en la cual algo sucede o se experimenta.

Las personas experimentamos estas modalidades a través de las modalidades sensoriales, que presentan nuestros principales canales de comunicación y sensación. Cuando se habla de multimodalidad en un proyecto de aprendizaje máquina, hablamos comúnmente de un conjunto de datos que incluye más de una modalidad.

Una metodología utilizada para que la inteligencia artificial progrese en entender el mundo que tiene alrededor, es el uso de múltiples modalidades para crear un razonamiento e interpretación más completos de la información que la rodea. Todas las actividades humanas contienen diferentes modalidades, un ejemplo claro es cuando una persona practica un deporte como el futbol, no únicamente observa la bola, también escucha los golpes, los pasos, huele el campo, siente el viento, ve la iluminación. De la misma manera la aproximación multimodal pretende brindar de contexto a una actividad de reconocimiento para sistemas de inteligencia artificial.

El reconocimiento multimodal de emociones es un tema que se ha estudiado extensamente en la última década, con el fin de buscar mejorar la precisión de los modelos de reconocimiento y clasificación.

Como ya mencionamos en este capítulo, las emociones humanas tienen diferentes modalidades, o medios en los cuales pueden ser comunicadas y descritas. La idea de utilizar

sistemas multimodales para la clasificación de emociones surge rápidamente como una de las soluciones para desambiguar falsos positivos y buscar de agregar contexto al reconocimiento de emociones, con el fin de mejorar la precisión y tener modelos más robustos.

Todos los sistemas multimodales presentan 5 tipos de desafíos en su desarrollo (Baltrušaitis, Ahuja, and Morency 2016), y gracias a esos desafíos pueden ser catalogados, dependiendo de las estrategias utilizadas para superarlos:

- **Representación:** Es el principal desafío del aprendizaje en general, como representar y resumir la información multimodal, en un medio o forma que aproveche la complementariedad y redundancia de múltiples modalidades. La heterogeneidad de los datos multimodales complica el construir representaciones apropiadas. Por ejemplo, el texto es normalmente simbólico, mientras que los datos audiovisuales son normalmente manejados como imágenes.
- **Traducción:** El segundo desafío es el problema de como traducir (o mapear) los datos de una modalidad a otra. No solamente en datos heterogéneos, si no la relación que existe entre las modalidades, este es normalmente un tema subjetivo. Por ejemplo, las maneras existentes correctas de describir una misma imagen, cada uno tomará una manera diferente de describirla, y una definición perfecta puede nunca llegar a existir.
- **Alineamiento:** El tercer desafío es identificar la relación directa entre (sub)elementos de dos o más diferentes modalidades.
- **Fusión:** El cuarto desafío es el desarrollar una metodología para juntar la información, de dos o más o modalidades, para realizar una predicción. Por ejemplo, para reconocimiento audiovisual del habla, se pueden fusionar las características de movimiento de la boca con la señal de sonido para predecir palabras habladas. La información que proviene de diferentes modalidades puede tener un valor de predicción variable y topología de ruido diferente, con la posibilidad de pérdida de datos en por lo menos una modalidad.
- **Co-aprendizaje:** Un quinto desafío es la transferencia de conocimiento entre modalidades. Esto se ejemplifica en algoritmos de co-aprendizaje, aterrizaje de

conceptos, y aprendizaje de disparo a cero. El co-aprendizaje explora como el aprender el conocimiento de una modalidad, puede ayudar a un modelo computacional a entrenar en una diferente. Este desafío es particularmente relevante, cuando una modalidad tiene recursos limitados.

2.3. Ambientes Inteligentes de aprendizaje

Un ambiente de aprendizaje es un ambiente donde tradicionalmente los estudiantes aprenden. También existen otros tipos de ambientes de aprendizaje, como lo son las herramientas de aprendizaje inteligente, donde el objetivo principal es facilitar diferentes actividades de aprendizaje a estudiantes (Coen 1998).

Un ambiente inteligente de aprendizaje tiene como objetivo el crear software de enseñanza que se adapte al ritmo de aprendizaje del estudiante, tomando en cuenta diferentes valores cognitivos de los mismos. El objetivo es crear escenarios que permitan mantener enganchado al estudiante para mejorar el ritmo de aprendizaje del estudiante y aprender de él. Esto normalmente se logra a través del uso de sistemas de reconocimiento de patrones de los usuarios, ya sean emociones o comportamiento dentro del ambiente de aprendizaje y adaptando la metodología de enseñanza utilizada para el proceso de aprendizaje de los alumnos.

En esta área de investigación relacionada con los ambientes inteligentes de aprendizaje, se combinan diversos campos tales como la pedagogía, psicología, ciencias cognitivas, IA, entre otras, dónde cada uno aporta su visión al desarrollo de esta disciplina.

Capítulo 3

3.Estado del arte

En este capítulo se definen los trabajos más actuales (4 años) en el área de reconocimiento multimodal de emociones, con sus técnicas, corpus o bases de datos utilizadas para el reconocimiento, así como las modalidades utilizadas y su precisión.

3.1. Reconocedores multimodales de emociones

Los avances recientes en las interacciones humano-máquina van más allá de la búsqueda de un medio natural para la transferencia de datos entre el humano y la máquina. Una importante aportación, y posible fuente de retroalimentación, viene de reconocer las emociones expresadas o el afecto de los usuarios al sistema. Las principales modalidades utilizadas para este reconocimiento son el audio y el video o imagen. Sin embargo, estas modalidades, aunque han tenido grandes avances en el reconocimiento de emociones de los usuarios, tienen problemas que todavía no han sido solucionados, la mayoría de ellos debido a la falta de escalabilidad de los modelos de reconocimiento. Lo anterior se debe a consecuencia de que la mayoría de las bases de datos utilizadas para el reconocimiento de emociones son desarrolladas en ambientes controlados, lo cual presenta un problema en su implementación real debido a que existen muchos medios de ruido en la adquisición de datos en ambientes no controlados.

Debido a esto en última década el reconocimiento multimodal de emociones ha encontrado una gran cantidad de atención de parte de la comunidad científica, esto gracias a la capacidad de los modelos multimodales de complementar información de unas modalidades con otras, volviendo más fiable el reconocimiento de las emociones.

Algunos trabajos han sido desarrollados compensando modelos unimodales utilizando métodos multimodales para mejorar el reconocimiento de emociones. En el trabajo de (Huang et al. 2019) utilizan un modelo unimodal de reconocimiento de emociones en rostro

utilizando CNN y lo complementan con lecturas electroencefalográficas utilizando SVM con una fusión tardía utilizando valencia y excitación. En el trabajo de (Liang et al. 2019) utilizan dos modelos unimodales de reconocimiento de emociones con contexto cultural, uno de rostro y uno de voz, con los cuales realiza aprendizaje adversario para confundir la clasificación de un modelo con la clasificación de otro buscando confundirlo para mejorar su precisión, esto lo logra a través de embebidos de semántica de la información escrita y hablada.

Otros trabajos han sido dedicados únicamente a metodologías de fusión temprana. En (Yang et al. 2019) se definieron técnicas de aprendizaje simétrico esparzo utilizando información privilegiada para la fusión de información en modalidades de audio y video para la detección de emociones básicas en video. En (Zhou et al. 2019) utilizan técnicas de codificación con CNN para la fusión de características y un decodificador con CNN para el reconocimiento de emociones básicas utilizando electro encefalogramas y señales fisiológicas externas. Otro trabajo que utiliza fusión temprana es el de (Kim and Shin 2019) en donde utilizan técnicas de estadística para la extracción de características en voz y técnicas de representación por diccionarios para la extracción de características en texto con una fusión a través de concatenación directa y una red densamente conectada para el reconocimiento de emociones básicas. Además en (Gogate, Adeel, and Hussain 2017) utilizan varios modelos de reconocimiento de emociones básicas con tres modalidades, texto, audio y video, donde presentan una metodología novedosa para la fusión temprana a través de concatenación de características y metodologías de comparación para la fusión tardía o de clasificaciones. También se han desarrollado sistemas para la fusión temprana para la detección de emociones en texto e imagen a través de la concatenación de características similar a la presentada por Gogate. En (Hu and Flaxman 2018) crearon un sistema de minería de datos de Tumblr para clasificar diferentes emociones con el fin de analizar la estructura de las emociones, utilizando LSTM con embebidos n-dimensionales para la extracción de características en texto y CNN para la extracción de características en imagen.

A su vez, también existen trabajos de reconocimiento multimodal con fusión tardía, en trabajos de (Kanjo, Younis, and Ang 2019) utilizaron agregación de clasificadores con fiabilidad para el reconocimiento de emociones básicas utilizando clasificadores de emociones en el rostro y en el lenguaje corporal. En (Wei, Jia, and Feng 2017) Utilizaron

técnicas de clasificación multimodal de emociones básicas utilizando electro encefalogramas, expresiones faciales y electrocardiografía, creando métodos de fusión temprana diferente y un método de fusión tardía utilizando pesos según la clasificación realizada por diferentes modelos de clasificación. En (Miao et al. 2018) utilizaron varios sistemas de fusión tardía para el reconocimiento multimodal de emociones básicas en voz y video, utilizando dos conjuntos acústicos con DBN y SVM, CNN y RNN utilizando un sistema de fusión de decisión promedio. En el trabajo de (Patwardhan 2017) utilizaron diferentes SVM para clasificar emociones mixtas y con un método estadístico de suma por pesos para su fusión, utilizando medios unimodales de voz, rostro y posición del cuerpo medido con Kinect, para el reconocimiento multimodal de emociones básicas.

En la tabla 3-1 se muestra una comparación entre los sistemas previamente mencionados, su tipo de fusión, las modalidades que utilizan, sus arquitecturas, sus conjuntos de datos y el nivel máximo de precisión obtenido con sus diversas técnicas multimodales.

Tabla 3-1. Comparativa de sistemas de reconocimiento multimodal de emociones.

Trabajo	Tipo de fusión	Modalidades	Arquitectura	Conjunto de datos	Precisión Máxima
Huang et. al 2019	Fusión tardía	EEG Expresión facial	CNN SVM	DEAP MAHNOB	70%
Liang et. al 2019	Fusión temprana	Expresión facial Voz	ANN Adversaria	CHEAVD AFEW	70.5%
Yang et. al 2019	Fusión temprana	Audio Video	S-ELM-LUPI	EOH LDN LBP	86.4%
Zhou et. al 2019	Fusión temprana	EEG Señales EP	CNN	Elaborado AD HOC	92%
Kim and Shin 2019	Fusión temprana	Audio Video	DNN	IEMOCAP	75.5%
Gogate, Adeel, and Hussain 2017	Fusión temprana	Texto Audio Video	CNN	IEMOCAP	78%
Hu and Flaxman 2018	Fusión temprana	Imagen Texto	CNN-LSTM	TUMBLR	72%
Kanjo, Younis, and Ang 2019	Fusión temprana	EEG Señales EP	MLP CNN CNN-LSTM	ENVBODYSENS	94.7%
Wei, Jia, and Feng 2017)	Fusión temprana y tardía	EEG Expresión facial ECG	Matriz con Pesos	MAHNOB-HCI	70%
Miao et. al 2018	Fusión temprana y tardía	Expresión facial Voz	SVM MLP DBN REPTree RF y CNN	CHEAVD	40.89%
Patwardhan 2017	Fusión temprana y tardía	Expresión facial Posición corporal Voz Gestos corporales	SVM	Ad-Hoc	96.6%

3.2. Ambientes de aprendizaje con reconocimiento multimodal

En los últimos años se han desarrollado una gran cantidad de ambientes de aprendizaje con reconocimiento multimodal. La mayoría de estos ambientes trabajan con reconocimiento de las emociones básicas descritas por (Ekman 1992) y tienen una gran variación en los sensores utilizados para captar las señales que permitan el reconocimiento de emociones.

La tabla 3-2 muestra los ambientes de aprendizaje dentro del estado del arte, las emociones que manejan y los diferentes tipos de datos que utiliza.

La mayoría de los trabajos anteriormente mencionados utilizan técnicas de ML, algunos de ellos tienen una cantidad remarcable de metodologías de extracción de información como por ejemplo VALERIE que presenta extracción de datos de cámara, micrófono, mouse y sensores fisiológicos utilizando metodologías de aprendizaje máquina tradicional. Mientras que por otro lado existen trabajos como INES, Ing-ITS y Crystal Island que realizan la detección de emociones utilizando únicamente variables cognitivas de la actividad de los usuarios.

Sin embargo, todos ellos concuerdan en que la evaluación de emociones en el proceso cognitivo es de suma importancia para la toma de decisiones del sistema de aprendizaje, y una de las emociones más repetida es el aburrimiento.

Tabla 3-2. Comparativa de ambientes de aprendizaje afectivos multimodales.

Ambiente	Sensores	Detección de datos emocionales	Emociones Reconocidas
Auto tutor	Video cámara Presión Silla sensible	Extracción de patrones en postura y ojos, análisis de bitácoras. Clasificadores: Naive Bayes, ANN, Regresión logística, KNN, Árboles de decisión C4.5	Flow, confundido, aburrido, frustrado, eureka, neutral.
Cognitive Tutor Algebra	No usado	Análisis de bitácoras que graban características relacionadas al comportamiento de los estudiantes, histórico de eventos y actividades en el proceso de aprendizaje. Clasificadores: Árboles de decisión J48, Algoritmos K*, Regresión por pasos, JRip, Naive Bayes, Arboles-REP	Aburrido, concentrado, frustrado, confundido.
Crystal Island	No usado	Análisis de encuestas, entrevistas y análisis de bitácoras. Emociones y modelado usando redes bayesianas.	Ansioso, aburrido, confundido, curioso, entusiasta, concentrado, frustrado.
Easy with Eve	Video cámara	Extracción de características faciales. Clasificador: máquina de soporte vectorial (SVM)	Sonriendo, riendo, sorprendido, enojado, asustado, triste, disgustado y neutral.
EER- Tutor	Video cámara	Características del rostro (ojos, cejas, labios) seguimiento y extracción. Clasificador: Geometría del rostro a través de FACS	Feliz, sonriente, enojado, frustrado y neutral.
FERMAT	Video cámara	Extracción de características y puntos de interés del rostro. Clasificador: ANN, lógica difusa	Enojado, disgustado, asustado, feliz, sorprendido y neutral.
Guru Tutor	Eye Tracker Video cámara	Extracción de características en el seguimiento de ojos y mirada, análisis de archivos bitácora y análisis de atención del usuario (tiempo puesto en pantalla)	Desinterés, aburrido
Inq-ITS	No usado	Análisis de archivos de bitácoras. Clasificador: Árbol de decisión J48 regresión de pasos. JRip	Aburrido, confundido, frustrado y concentrado.
INES	No usado	Análisis del nivel de actividad del estudiante, dificultad de la tarea, progreso previo, numero de errores, severidad del error. Emociones predichas a través de reglas de apreciación	Preocupado, confiado, deprimido, entusiasmado.
ITSPOKE	Micrófono	Extracción de acústica-prosódica, características léxicas (intensidad del habla, energía, volumen, duración y pausas) y características de dialogo (ej. la exactitud de la respuesta). Análisis semántico usado para la evaluación de la precisión de respuesta y regresión lineal para evaluar la confianza de la evaluación.	Negativo, positivo y neutral.
MathSpring	No usado	Análisis de archivos de bitácora, reportes de autoevaluación, patrones de comportamiento. Clasificador: regresión lineal	Confiado, preocupado, excitado, inactivo satisfecho, frustrado, interesado, aburrido.
MetaTutor	Eye Tracker	Extracción de datos de características de la mirada y características de áreas de interés en la interfaz del sistema.	Aburrido, curioso, interesado.
PAT2Math	Video Cámara	Análisis de archivos de bitácora y extracción de características de puntos faciales. Emociones basadas en el uso de FACS y modelado psicológico de emociones (modelo OCC)	Satisfecho, decepcionado, feliz, triste, agradecido, enojado, avergonzado.
Prime Climb	Sensores Fisiológicos	Determinación de la conductividad de la piel, velocidad del corazón, actividad muscular y análisis de datos de bitácora.	Feliz, triste (por el juego), admiración, criticismo (por PA), orgullo, tristeza (por sí mismo)
VALERIE	Video Cámara Micrófono Mouse Sensores Fisiológicos	Determinación de la conductividad de la piel, velocidad del corazón, extracción de características faciales y del habla, análisis de actividad del mouse. Clasificadores: KNN, función de análisis discriminativo, Algoritmo Marquat de propagación hacia atrás.	Tristeza, enojo, sorpresa, asustado, frustración, divertido.
WaLLis	No usado	Análisis de archivos de bitácora. Clasificador: Algoritmo de árbol de decisión J48	Frustrado, confundido, aburrido, confiado, feliz, entusiasmado.

Capítulo 4

4.Desarrollo del proyecto

En este capítulo se muestran las diferentes etapas del desarrollo del proyecto, empezando en la descripción de los corpus utilizados para el entrenamiento de los modelos de DL, seguido por las diferentes técnicas utilizadas para la creación de los modelos multimodales, donde se proponen soluciones para los desafíos de representación, traducción, alineamiento y fusión, que existen en el desarrollo de reconocedores multimodales. Además, se describe el funcionamiento de los sistemas multimodales, su arquitectura y por último los modelos de implementación de estos sistemas en ambientes de aprendizaje.

4.1. Conjuntos de datos (*Corpus*)

En un sistema de reconocimiento, el conjunto de datos es una parte crucial en el proceso de creación de los modelos de reconocimiento. Esto se debe a que este conjunto de datos sirve como ejemplo para el reconocedor para el entrenamiento.

Conocer el origen de los datos es un proceso tan importante como el diseño y creación del modelo de reconocimiento, por esto que en este subcapítulo primero se abordan los diferentes conjuntos de datos empleados para el entrenamiento y su proceso de obtención.

A continuación, se muestran los tres conjuntos de datos utilizados en el desarrollo del modelo multimodal y una explicación de su metodología de obtención.

4.1.1. Emotiv Insight (CEI)

La obtención de este conjunto de datos se llevó a cabo en 2016 en el Instituto Tecnológico de Culiacán (Leal Hernandez 2017). Este conjunto de datos se obtuvo a través del uso de una diadema electroencefalográfica de 5 canales (Emotiv Insight) en conjunto con una cámara web (Logitech HD Pro C920) para el muestreo de imágenes de estudiantes durante el uso de

una plataforma web de aprendizaje de java llamada *Java Gaming ILE* y la observación de videos.

El experimento se realizó con alumnos de licenciatura del Instituto Tecnológico de Culiacán en el laboratorio de Maestría en Ciencias de la Computación, en donde se explicó el motivo y los objetivos del corpus. La captura de datos se realizó en dos sesiones. En la primera participaron 18 hombres y 7 mujeres, en la segunda participaron 10 hombres y 3 mujeres, todos los alumnos con un rango de edad de 18 a 47 años.

Se dividieron las muestras en dos grupos, el grupo A fue inducido a las emociones comprometido, interesado, emocionado y enfocado utilizando el *Java Gaming ILE*. El segundo grupo fue inducido a las emociones de aburrido y relajado.

Se utilizaron las señales del Emotiv Insight para relacionar a través de estampas de tiempo el rostro de los alumnos con las emociones inducidas, a través de analizar y comparar el tipo de señales producidas por los canales de la diadema de acuerdo con dos diferentes algoritmos para la clasificación y etiquetado de las imágenes tomadas por la cámara web.

El proceso generalizado de etiquetado de las imágenes con el uso de la diadema se muestra en la Figura 4-1 (Zatarain Cabada, Ramón Lucia et al. 2018).

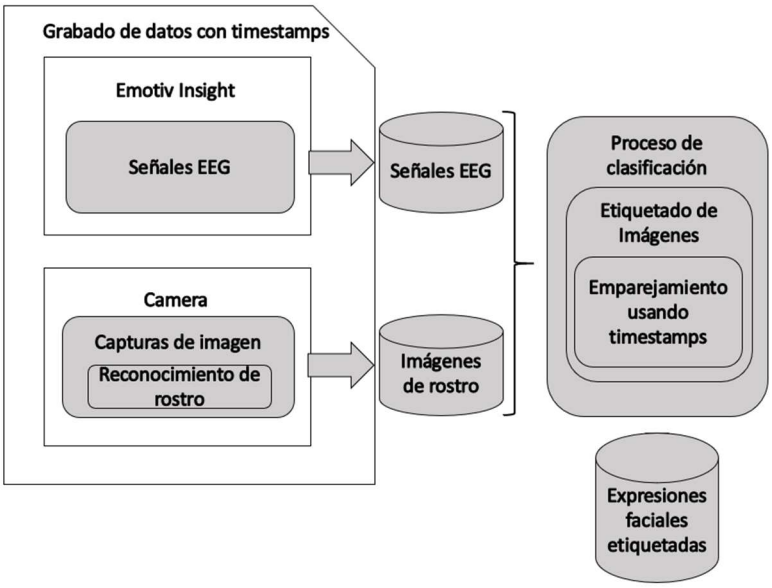


Figura 4-1. Proceso general para el etiquetado de las imágenes con la diadema Emotiv.

De este experimento se crearon 2 conjuntos de datos de imágenes clasificadas de dos algoritmos, sin embargo, debido al nivel de precisión en la clasificación de estos solamente se tomó en cuenta el algoritmo 1. Ambos resultados se muestran en la tabla 4-1.

Tabla 4-1. Resultados de clasificación de imágenes EmotivInsight

Numero de imágenes clasificadas		
Emoción	Algoritmo 1	Algoritmo 2
Aburrido	122	123
Enganchado	1995	461
Emocionado	1661	2953
Enfocado	222	356
Interesado	150	45
Relajado	28	333
Total	4178	4271

4.1.2. SentiText (CST) y EduSere (CES)

Estos conjuntos de datos fueron elaborados en el Instituto Tecnológico de Culiacán, entre los años 2017-2019. Para la construcción de una de las partes del corpus de texto, se desarrolló una aplicación para extraer las opiniones de Twitter con una delimitación geográfica al noroeste de México con las palabras clave: ‘programador’, ‘desarrollador’, ‘maestro’,

‘maestra’ ‘profesor’, ‘profesora’, ‘estudiante’, ‘Java’ y ‘Python’, (Oramas-Bustillos et al. 2018).

Posteriormente, se realizó una recolección manual de opiniones sobre cursos relacionados con lenguajes de programación en diversas plataformas de enseñanza y plataformas sociales como Udemy, YouTube y Platzi.

Para la obtención de frases centradas en el aprendizaje (frustrado, aburrido, neutral, emocionado y comprometido) en el dominio específico de lenguajes de programación se desarrolló un sistema de evaluación de recursos educativos (SERE, por sus siglas) (Barrón-Estrada, Zatarain-Cabada, Oramas-Bustillos, & Ramírez-Ávila, 2017). SERE fue diseñado para interactuar con los estudiantes con el objetivo de permitir la expresión de opiniones y comentarios libremente acerca de los recursos educativos u objetos de aprendizaje de temas en un curso de programación. La Figura 4-2, muestra la interfaz de usuario de este sistema. El estudiante ingresa al sistema y selecciona el curso que desea estudiar. Posteriormente, el sistema despliega la lista de temas y los recursos educativos disponibles; el usuario selecciona el objeto de aprendizaje que le interesa y después escribe una opinión textual que va acompañado de un emoticono que representa la emoción que el estudiante expresa en ese momento.

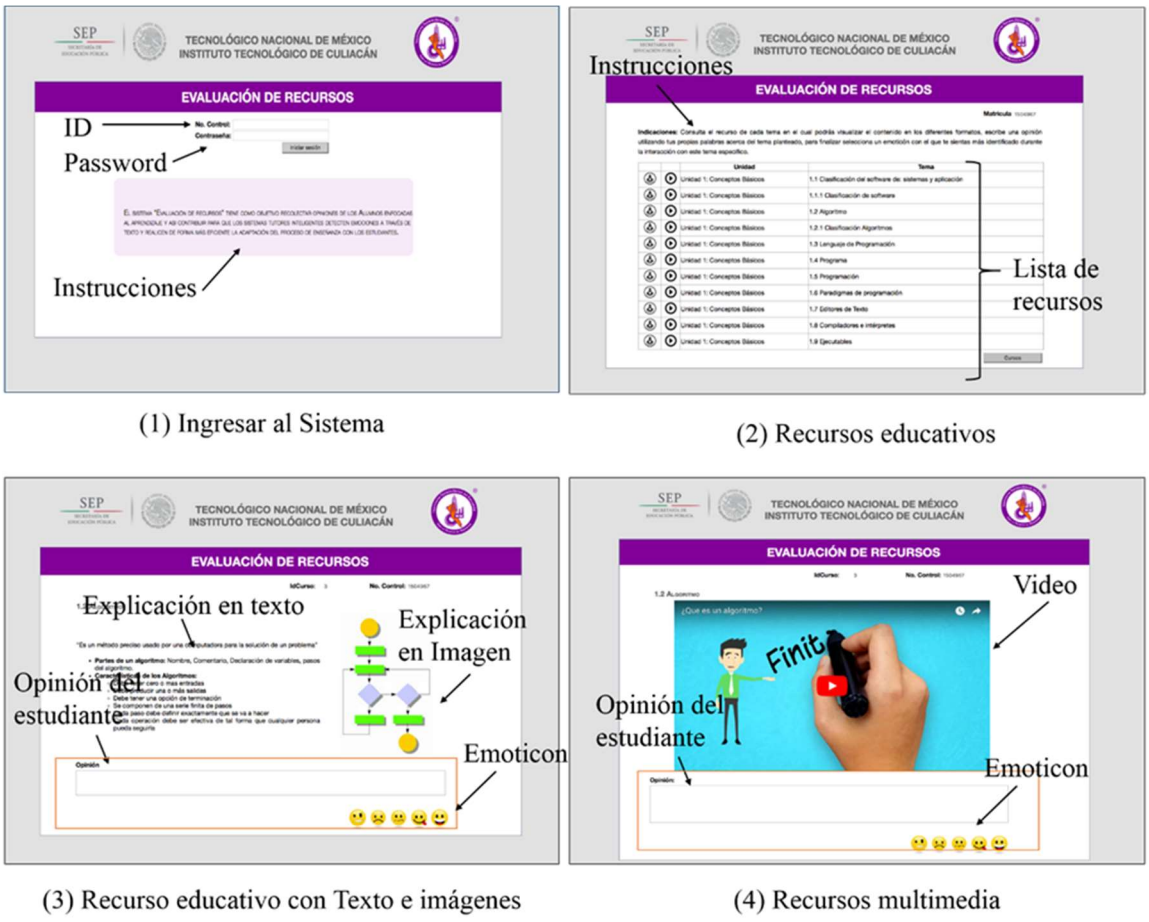
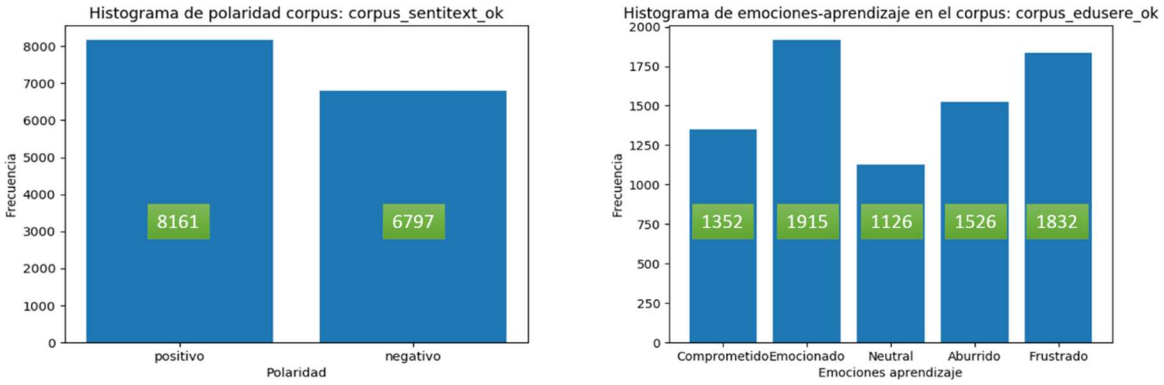


Figura 4-2. Interfaz SERE (Raúl Oramas 2018).

De estas tareas se obtuvieron dos corpus de texto. El primero denominado corpus SentiText que contiene frases orientadas a la polaridad y el segundo corpus EduSere con frases orientadas a emociones educativas. Estas opiniones fueron clasificadas manualmente por un grupo de profesores de la asignatura de lenguajes de programación. SentiText contiene 14958 frases y EduSere contiene 7751 frases. La distribución de cada categoría se muestra en la Figura 4-3.



(a) SentiText (b) EduSere

Figura 4-3. Categorización de los corpus SentiText y EduSere (Oramas-Bustillos et al. 2018).

4.1.3. Alineamiento de Datos

Dado que las emociones en el aprendizaje utilizadas para la creación de los conjuntos de datos son diferentes, estas deben ser alineadas para propósitos de su uso en la clasificación multimodal. Para esto se utiliza el modelo de Russell para realizar un mapeo de los estados cognitivo-afectivos centrados en el aprendizaje de los diferentes conjuntos de datos. El modelo de Russell es presentado en la Figura 4-4.

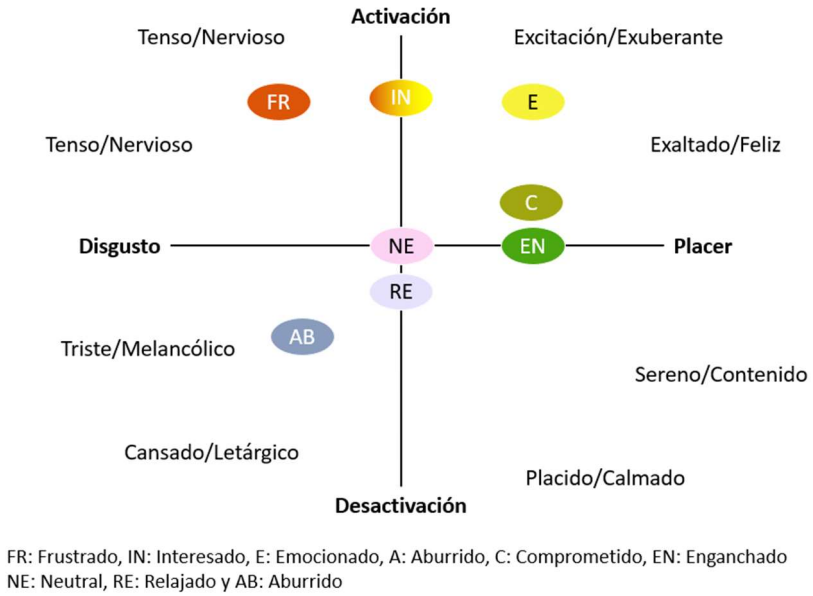


Figura 4-4. Circunflejo de emociones basado en el modelo de Russell.

La tabla 4-2 muestra las diferentes emociones etiquetadas en los conjuntos de datos de imagen y de texto (EmotivInsight y EduSere respectivamente), con las emociones más cercanas según el modelo de Russell.

Tabla 4-2. Emociones etiquetadas en los diferentes conjuntos de datos y sus emociones más cercanas según el modelo de Russell.

EmotivInsight	EduSere
Aburrido	Aburrido
Enganchado	Comprometido
Emocionado	Emocionado
Enfocado	-------
Interesado	Frustrado
Relajado	Neutral

De esta manera, es posible el realizar una aproximación semántica de las diferentes clasificaciones de estos dos conjuntos de datos, a través de la observación de su dinámica con respecto al proceso cognitivo de tareas complejas como lo muestran (D'Mello and Graesser 2012) en su artículo de las dinámicas de la emoción en el proceso cognitivo complejas. La Figura 4-5 muestra el modelo hipotético de D'Mello.

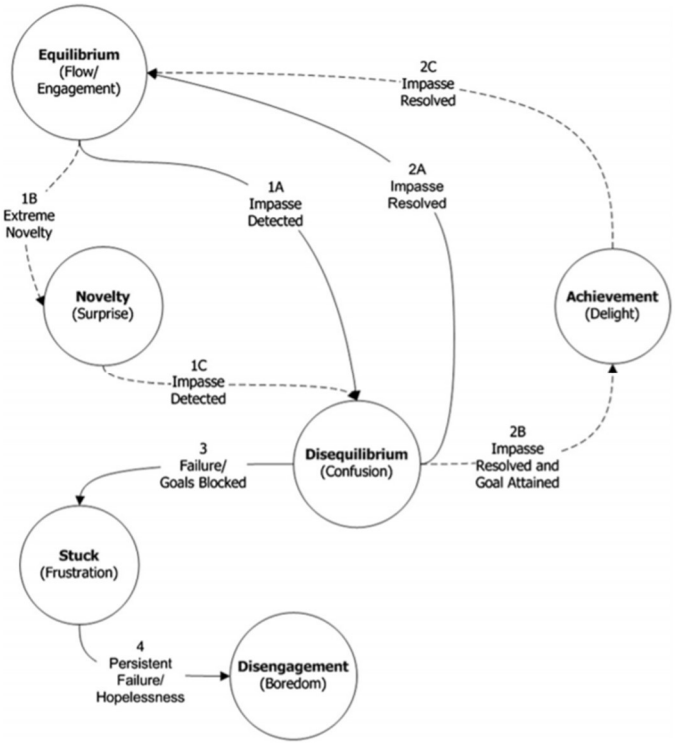


Figura 4-5. Modelo de D'Mello de la dinámica del proceso cognitivo (D'Mello and Graesser 2012).

En este modelo se observa que la emoción con un valor mayor cognitivo es *enganchamiento* o *comprometido*, una emoción que mantiene el flujo del proceso cognitivo es la confusión. En este caso esa emoción se puede mapear de manera negativa a *frustración* en el modelo de Russell y de manera positiva a *sorpresa*, por lo cual se utilizan estas emociones como equivalentes de la confusión en el modelo dinámico de D'Mello. *Emocionado* es mapeado directamente a *deleite. relajado* y *neutral* se mantienen como una sola emoción *neutral*, es decir no se detecta un cambio de emoción. De la misma manera, *aburrido* se mantiene bajo su misma etiqueta, de manera que las emociones quedan mapeadas como se describe en la Figura 4-6.

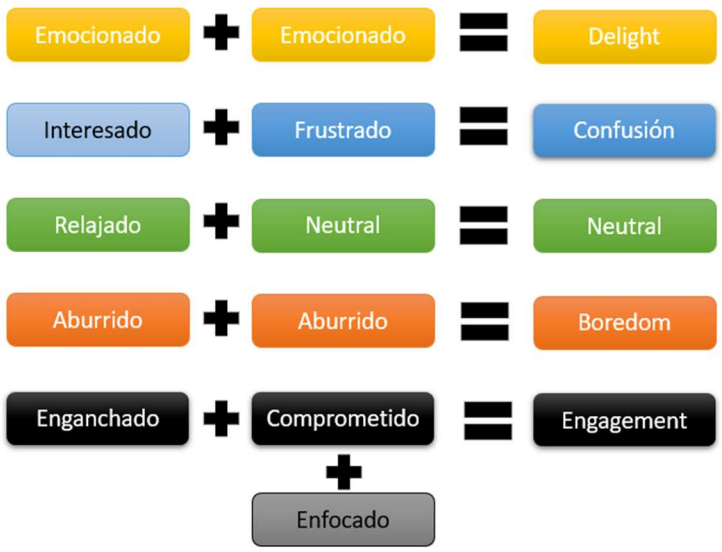


Figura 4-6. Relación de emociones discretas en los conjuntos de datos.

Por lo tanto, las emociones de ambos conjuntos de datos pueden ser representadas en equivalencia con las emociones dentro de la dinámica del proceso cognitivo.

Además de esto dentro del modelo de Russell también se pueden definir diferentes emociones a través de los cambios de *placer* y *disgusto*, que son mapeados directamente a sentimiento positivo o negativo. De esta manera emociones como *neutral* con sentimiento negativo puede ser mapeado como *aburrimiento,* y *neutral* con sentimiento positivo puede ser mapeado a *emocionado* en la Figura 4-4, y estas emociones presentan una clara relación con el modelo

teórico de D'Mello presentado en la Figura 4-5. De esta manera se logra el alineamiento de los datos a través del modelo de D'Mello.

4.2. Técnicas de representación

La representación de los datos como se menciona en el capítulo 2, es una parte fundamental en la creación de modelos de ML y DL. La representación de los datos están fuertemente ligadas a los diferentes algoritmos utilizados en esta área para la extracción de características y predicción.

Es por esto que para solucionar el problema de las representaciones se adaptaron las representaciones de los datos de múltiples modalidades a modelos computacionales previamente utilizados en DL para la clasificación de emociones orientadas al aprendizaje, en imágenes (Zatarain Cabada, Ramón Lucia et al. 2018) y en texto (Oramas-Bustillos et al. 2018). El primer modelo es una CNN para operaciones matriciales. Esta red fue evaluada originalmente con imágenes de estudiantes (Emotiv Insight) como entrada para la extracción de características y predicción de emociones orientadas al aprendizaje. El segundo modelo es usado en el procesamiento de lenguaje natural, utilizando una red de LSTM-CNN con representación de vectores texto en embebidos N-dimensionales para crear una matriz de relación directa entre palabras e imágenes, con la cual se evaluaron polaridad (SentiText) y emociones orientadas al aprendizaje (EduSere).

Las técnicas de representación fueron definidas como imágenes para todos los valores multimodales a utilizar en CNN, y vectores de embebidos de información para la red o modelo de LSTM-CNN. Para llevar a cabo la conversión se definió primero un sistema de representación intermedio de la información. Toda la información independiente de la modalidad puede ser representada con un valor numérico en un vector o matriz mapeada, siendo esta una representación intermedia perfecta.

Para el proceso de transformación de representación de datos se siguieron dos pasos: Una traducción de los datos unimodales a una representación intermedia a través de un preprocesamiento de la información y una conversión o alineamiento de representaciones a

través de diferentes modelos o diccionarios para obtener la representación modal a utilizar para los clasificadores.

Una vez obtenidas estas representaciones finales, se procedió a realizar una fusión de representaciones. Este proceso se presenta en la Figura 4-7 y los específicos del proceso se presentan en las subsecciones siguientes.

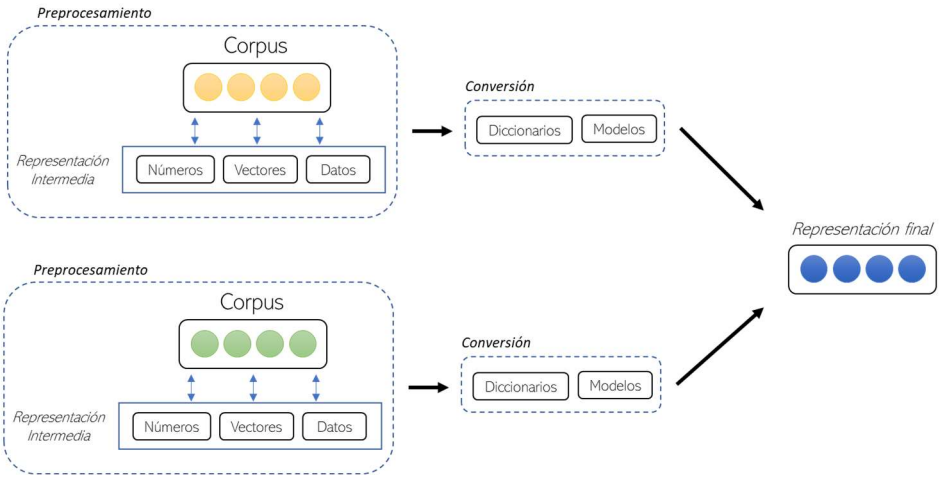


Figura 4-7. Proceso de transformación y fusión de representaciones.

4.2.1. Preprocesamiento de representaciones

El preprocesamiento de las representaciones tiene como objetivo obtener representaciones intermedias de los datos de las diferentes modalidades del experimento. En la modalidad de imagen, se utiliza la escala RGB y la escala de grises para modelar los datos a una representación intermedia de la información. En la Figura 4-8 se observa un ejemplo de las diferentes representaciones de los datos

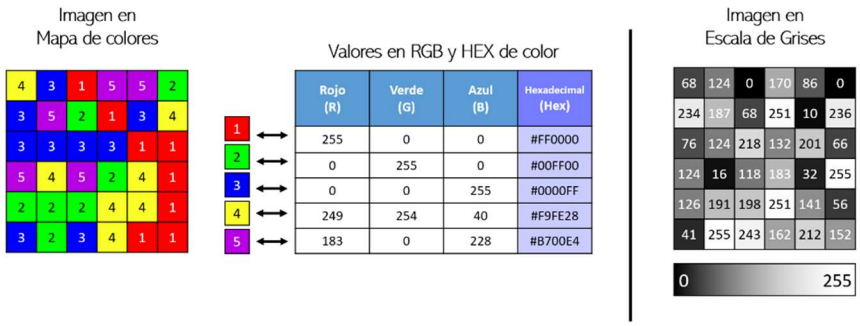


Figura 4-8. Representaciones de imagen en dos formatos diferentes.

Mientras que en el medio de texto se utilizan diccionarios para el modelado de información, estos diccionarios reemplazan a cada palabra dentro de los documentos con valores numéricos que representan un valor de posición en el orden de aparición de las palabras, sin discernir significado, únicamente creando vectores representativos de diversos documentos. Un ejemplo del funcionamiento de estos diccionarios puede observarse en la Figura 4-69.

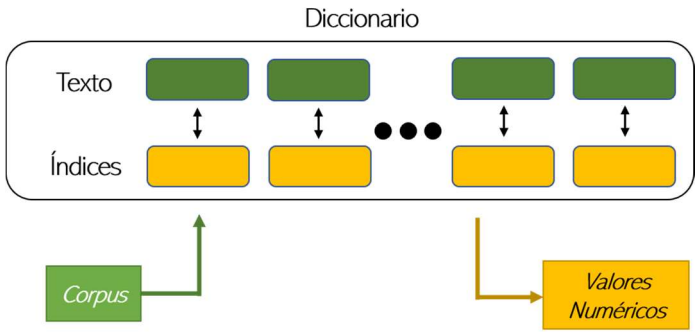


Figura 4-9. Representación de diccionarios de texto.

El utilizar modelos de representación de datos permite crear vectores con valores numéricos enteros que representaran de manera traducible la información de las diferentes modalidades en una llamada representación intermedia, con la cual se realizó la fusión de datos.

4.2.2. Modelos de conversión

Los modelos de conversión son definidos en función de los reconocedores que se utilizarían para la clasificación multimodal de emociones. Dado que las arquitecturas de DL utilizadas para el experimento realizan operaciones de convolución para la extracción de características como su metodología central, y las convoluciones se realizan a través de operaciones matriciales, se define que los modelos de conversión terminan en representaciones matriciales de los datos.

A su vez las arquitecturas de DL objetivo para la adaptación de las representaciones tienen un fin específico, y como tal, se delimitan las representaciones de los datos directamente a cada una de las arquitecturas.

La primera arquitectura es una CNN que utiliza la modalidad de imagen para realizar la clasificación, debido a que una de las modalidades se encuentra en formato de imagen únicamente se realiza la conversión de la modalidad de texto. Para la conversión de texto a imagen se utilizan los vectores de índices de los documentos y con el uso del modelo hexadecimal, se convierten los valores de los índices a valores hexadecimales y posteriormente se utiliza la escala RGB para representar los valores del vector hexadecimal como una línea de pixeles, logrando así representar el texto en un formato de imagen para su posterior fusión. Este proceso se muestra en la Figura 4-10.

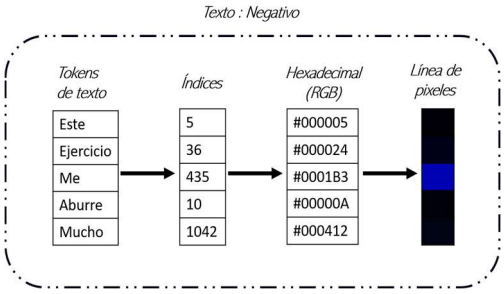


Figura 4-10. Proceso de conversión de texto a imagen.

Para la arquitectura de CNN-LSTM se define una matriz de embebidos (*Word Embeddings*) como la representación final de las modalidades de texto e imagen y debido a esto se define el uso de un diccionario universal para todas las modalidades. Se reservan los primeros 255 índices del diccionario para la representación de los valores de pixeles de una imagen en escala de grises. Esta representación de imagen en vector de valores enteros se presenta en la Figura 4-11.

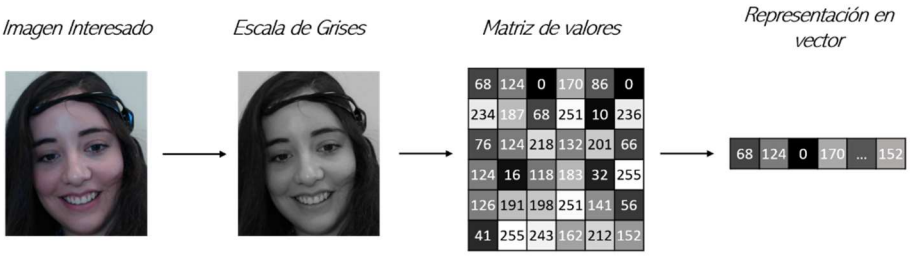


Figura 4-11. Proceso de conversión de imagen a vector.

Con el uso del mismo diccionario también se definen los indicies de las palabras en los documentos de texto y a su vez se crean las representaciones intermedias de documentos en formato de vector. Estos vectores son modelados en sus representaciones de embebidos de acuerdo con la fórmula de probabilidad de embebidos de palabras de espacio continuo como se muestra en la formula.

$$P(w_t|w_{t-k},\ldots,w_{t-1},w_{t+1},\ldots,w_{t+k}).$$

Donde P representa la probabilidad del embebido en la dimensión analizada, w representa el contexto o la repetición de palabras con respecto al valor o índice relacionado.

4.2.3. Fusión de representaciones

La fusión de representaciones se lleva a cabo utilizando las representaciones intermedias de cada una de las modalidades de texto e imagen. Esta se realiza a través de la concatenación directa de los vectores de datos de imagen y texto, creando nuevos vectores que representan la información de ambas modalidades fusionadas, y, de la misma manera se fusionan las etiquetas de cada uno de estos vectores.

De esta manera se crean 2 nuevos conjuntos de datos que fusionan la información de los corpus CEI y CES. Estos conjuntos de datos se definen en base a la modalidad final de representación de los datos. El primer conjunto está basado en la representación de Imágenes y el segundo está basado en la representación de vectores de datos, que se traducen en embebidos de información en espacios N-dimensionales.

4.3. Sistemas multimodales

Se crearon tres diferentes sistemas multimodales para la detección de emociones en texto e imagen: los primeros dos, utilizan las técnicas de representación, modelado y fusión de datos descritas en el subcapítulo anterior para realizar una fusión de datos, y el ultimo utiliza un sistema de fusión de características, basado en las arquitecturas de CNN y LSTM-CNN. En las siguientes subsecciones, se describe la metodología de la creación, e implementación de los diferentes sistemas.

4.3.1. Sistemas basados en representaciones conjuntas

Los sistemas basados en representaciones conjuntas utilizan la fusión de representaciones para llevar a cabo la tarea de fusión de datos. Estos sistemas tienen la característica de haber sido diseñados con el modelo o arquitectura utilizado para la toma de decisiones de las representaciones correctas de los datos. Dos sistemas basados en representación conjunta fueron creados; el primer sistema fue un sistema con representación basada en imágenes (SRBI) y el segundo fue un sistema con representación basado en embebidos (SRBE).

Sistema con representación basada en imágenes (SRBI)

Para la creación de este sistema se define en primera instancia el modelo de CNN como técnica de DL para la clasificación de la información. Se utilizan los conjuntos de datos de CEI y CST. Dado que el corpus de CEI contiene imágenes con una de cinco etiquetas (emociones) y el corpus de CST contiene mensajes de texto con una de dos etiquetas (polaridades), se fusionan datos y etiquetas, obteniendo un total de diez clases diferentes.

El algoritmo 1 muestra el proceso de creación del sistema SRBI.

Algoritmo 1 Sistema de representación basado en Imágenes

```
BEGIN /* Sistema de representación basada en imágenes*/
    Cargar corpus CEI
    Cargar corpus CST
    BEGIN /* Vectorización de CST */
    FOR i en CST DO
        Limpiar (i)
        Tokenizar (i)
        Agregar tokens al diccionario
        Anexar los índices de tokens a un vector x
        Padding (x)
        IntegerToHex(x)
        Anexar a la lista a de pares de x y su etiqueta
    END
    BEGIN /* Preprocesamiento CEI*/
    FOR j en CEI DO
```

```
16:       Detectar rostro (j)
17:           Redimensionar imagen (j)
18.           Anexar a la lista b de pares de j y etiqueta
19:       END
20:       BEGIN /* Representación conjunta en imagen */
21:           FOR k en lista b DO
20:               Obtener par aleatorio z de lista a
21:               Fusionar (z, k) //datos y etiquetas
22:               Anexar fusión a CMI
23:               Obtener par aleatorio z' de lista a
24:               WHILE etiqueta z = etiqueta z' DO
25:                   Obtener par aleatorio z' de lista a
26:               END
27:               Fusionar (z', k)
28:               Anexar fusión a CMI
29:           END
30:       END
31:       BEGIN /* Red Neuronal Convolucionada */
32:           Agregar Conv3D (Tamaño de imagen)
33:           Agregar MaxPool3D
34:           Agregar Conv3D
35:           Agregar MaxPool3D
36:           Agregar Conv3D
37:           Agregar MaxPool3D
38:           Agregar Flatten
39:           Agregar Dense
40:           Agregar Dropout
41:           Agregar Dense
42:           Agregar Dropout
43:           Agregar Dense (Número de clases, softmax)
44:           Compilar CNN (Optimizador)
45:           Entrenar (CMI, epoch, batch, validation)
46:       END
47:   END
```

Primero, se cargan los conjuntos CEI y CST preprocesamiento y la creación de la representación basada en imágenes (1-3).

Para llevar a cabo la fusión de los dos conjuntos de datos, se llevan a cabo dos pasos: el primer paso es la conversión del CST en una representación intermedia vectorizada. Para esto, se lleva a cabo un proceso de vectorización de las palabras contenidas en el corpus CST, donde se crean los vectores que contienen tokens que representan las palabras de cada frase y se lleva a cabo un proceso de *padding* para el rellenado de los vectores hasta conseguir 150 espacios de tamaño. A continuación, los valores decimales de los tokens se convierten a un valor hexadecimal de formato RGB, con el formato #RRGGBB, y se crea un vector de tamaño 150 con la representación de este texto. Esto también puede representarse como una imagen de 1 x 150 píxeles de tamaño como se muestra en la Figura 4-10, donde se crean un par de vectores y etiquetas (4-13).

Para el segundo paso se obtiene una imagen del conjunto CEI a través del preprocesamiento de imágenes. El sistema identifica el rostro del usuario en la imagen y la redimensiona a un tamaño de 150x150x3. Se crean un par de cara y etiqueta (14-19).

Una vez que la representación de imagen de ambas modalidades está disponible, se realiza el proceso de fusión, mediante la obtención de la imagen del rostro y la adición de la representación en imagen de un texto con polaridad, utilizando el formato RRGGBB para las 3 dimensiones de color, por lo tanto, creando una imagen de tamaño 151 x 150 x 3. De la misma manera, las etiquetas de la imagen de la cara y el texto se adjuntan para formar un par de imagen y representación de etiqueta. Esta pareja es posteriormente almacenada en un conjunto multimodal con representación basada en imágenes (CMI). Posteriormente se obtiene una representación en imagen de un texto del CST con la polaridad opuesta y se repiten los pasos anteriores. Fusionamos las representaciones de imagen y las etiquetas creando un nuevo par que se anexa al CMI (19-30). Finalmente, se crea un modelo CNN el cual utiliza el CMI para realizar el entrenamiento y la validación (31-47).

El proceso de manera gráfica se muestra en la Figura 4-12.

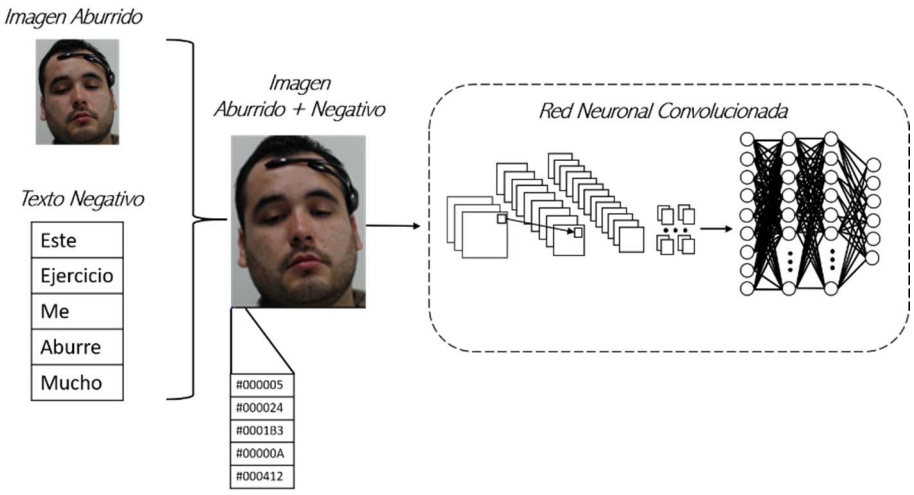


Figura 4-12. Proceso de general del sistema SRBI.

Sistema con representación basada en embebidos (SRBE)

Para la creación de este sistema se define un modelo de CNN-LSTM como técnica de DL para la clasificación de las emociones orientadas al aprendizaje donde se utilizaron los conjuntos CEI y CST. CEI es un corpus utilizado para clasificar expresiones faciales en cinco emociones orientadas al aprendizaje y CST es un corpus para clasificar expresiones textuales

en dos polaridades. En la fusión de rostro con texto, los datos de ambos conjuntos se concatenan mediante una representación vectorial. Luego, se usa un espacio de embebidos de información para crear un nuevo corpus con 10 diferentes etiquetas.

El algoritmo 2 muestra el proceso de creación del sistema SRBE.

Algoritmo 2 Sistema con representación basada en Embebidos

```
 1: BEGIN /* Sistema con representación basada en Embebidos*/
 2:     Cargar corpus CEI
 3:     Cargar corpus CST
 4:     Crear un diccionario con 255 índices reservados
 5:     BEGIN /* Vectorización de CST */
 6:     FOR i en CST DO
 7:         Limpiar (i)
 8:         Tokenizar (i)
 9:         Agregar tokens al diccionario
10:         Anexar los índices de tokens a un vector x
11:         Padding (x)
12:         Anexar a la lista a de pares de x y su etiqueta
13:     END
14:     BEGIN /* Preprocesamiento CEI*/
15:     FOR j en CEI DO
16:         Detectar rostro (j)
17:         Redimensionar imagen (j)
18.         Convertir a escala de grises (j)
19:         Vectorizar (j) en vector x
20:         Anexar a la lista b de pares de x y etiqueta
21:     END
20:     BEGIN /* Representación conjunta en embebido */
21:         FOR k en lista b DO
20:             Obtener par aleatorio z de lista a
21:             Fusionar (z, k) //datos y etiquetas
22:             Anexar fusión a CME
23:             Obtener par aleatorio z' de lista a
24:             WHILE etiqueta z = etiqueta z' DO
25:                 Obtener par aleatorio z' de lista a
26:             END
27:             Fusionar (z', k)
28:             Anexar fusión a CME
29:         END
30:     END
31:     BEGIN /* CNN+LSTM con capa de embebidos*/
32:         Agregar Embedding
33:         Agregar Dropout
34:         Agregar Conv1D
35:         Agregar MaxPool1D
36:         Agregar Conv1D
37:         Agregar MaxPool1D
38:         Agregar Dropout
39:         Agregar LSTM
40:         Agregar Dense
41:         Compilar
42:         Entrenar con CME
43:         Evaluar
```

```
44:         END
45:   END
```

Primero, se cargan los conjuntos CEI y CST y se crea un diccionario con los primeros 255 índices reservados para la representación de la imagen (1-4). Para llevar a cabo la fusión de los dos conjuntos de datos, se realizan dos pasos: el primero es la conversión del conjunto CST a una representación intermedia vectorizada. Para esto, se lleva a cabo un proceso de vectorización de las palabras contenidas en el corpus CST, donde se crean vectores que contienen tokens que representan las palabras de los documentos. Estos tokens se agregan al diccionario para su indexación. Luego, los índices de tokens se utilizan para crear una representación de vector del texto. A continuación, se utiliza un proceso de padding para rellenar el vector a un tamaño de 150 espacios y se crea un par de vector y etiquetas (5-13).

En el segundo paso, se pre procesan las imágenes del conjunto CEI. Para esto se detecta el rostro de la persona en una imagen, se redimensiona a 3x150x150 pixeles de tamaño, y se realiza el proceso de conversión a escala de grises. Después, se transforma la imagen en una matriz de números enteros y, posteriormente, en un vector, intercambiando los valores de cada píxel a un valor entero correspondiente (14-21). Este proceso se muestra en la Figura 4-8.

Una vez que la representación de imagen de ambos conjuntos de datos está disponible en su formato de vector, procedemos a realizar el proceso de fusión. Para esto se obtiene un vector que representa la imagen del rostro y se le concatena la representación vectorial de un texto con polaridad. De la misma manera, las etiquetas de la imagen y del texto se concatenan para formar un par de vector y etiqueta. Este par es agregado al conjunto de datos CME. Posteriormente obtenemos una representación en la imagen de texto con la polaridad opuesta, y se repite la fusión (20-30).

Por último, se crea a un modelo CNN-LSTM con una capa de embebidos como primera entrada en la cual se lleva a cabo la conversión de los vectores en embebidos. El corpus de vectores CME se utiliza para rellenar la capa de embebidos N-dimensional, que posteriormente se utiliza para entrenar el modelo CNN-LSTM (31-45).

Este proceso se puede observar gráficamente en la Figura 4-13.

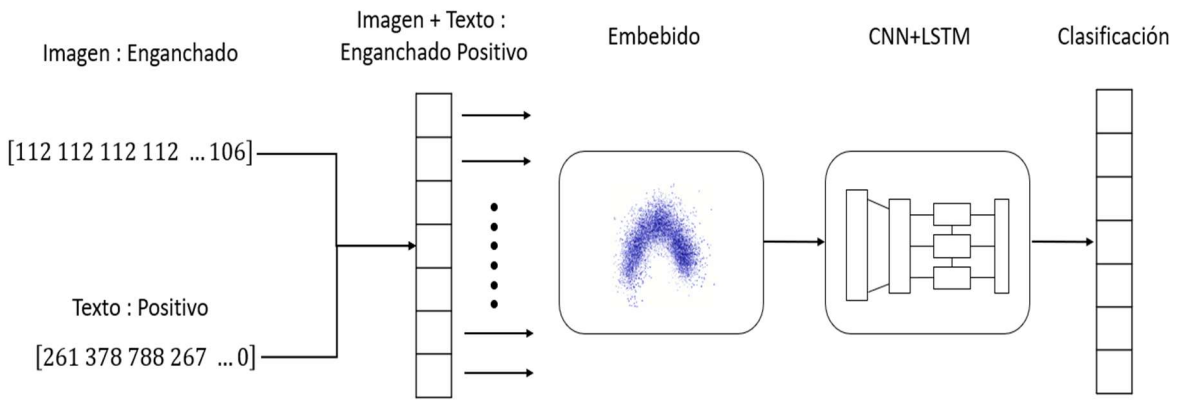


Figura 4-13. Proceso completo del SRBE.

4.3.2. Sistema basado en fusión de características

A diferencia de los sistemas basados en representaciones conjuntas, el sistema basado en fusión de características utiliza ambas arquitecturas CNN y LSTM-CNN antes descritas, reutilizando las etapas de extracción de características y agregando una capa para la concatenación de estas. A diferencia de los otros sistemas este sistema realiza una fusión de características extraídas de las imágenes y el texto, y por lo tanto no crea un nuevo conjunto de datos para realizar la clasificación.

El sistema de fusión de características (SFC) utiliza los modelos de CNN y LSTM-CNN previamente entrenados para la detección de emociones orientadas al aprendizaje utilizando los conjuntos de datos CEI y CES respectivamente. Ambos con cinco etiquetas de emociones orientadas al aprendizaje.

El algoritmo 3 muestra el proceso de creación del SFC.

Algoritmo 3 Fusión de características

```
 1:  BEGIN /* Fusión de características*/
 2:      Cargar CEI
 3:      Cargar CES
 4:      BEGIN /* Preentrenamiento CNN */
 5:          Pre procesar el CEI
 6:          Crear arquitectura CNN
 7:          Compilar
 8:          Entrenar con CEI
 9:          Salvar el modelo
10:      END
```

```
11:      BEGIN /* Preentrenamiento CNN + LSTM*/
12:          Pre procesar el CES
13:          Crear arquitectura CNN + LSTM
14:          Compilar
15:          Entrenar con CES
16:          Salvar el modelo
17:      END
18:      BEGIN /* Crear red de Fusión */
19:          Crear modelo secuencial
20.          Agregar capas de topologías idénticas a la red
             CNN
21:          Agregar capas de topologías idénticas a la red
             CNN+LSTM
22:          Agregar capa concatenación características
23:          Agregar capas densamente conectadas
24:          Inicializar los pesos
25:          Compilar red
26:          Entrenar con CEI y CES
27:          Evaluar modelo
28:          END
29:  END
```

Primero, se cargan los conjuntos CEI y CES (2-3). Luego, se crean dos modelos de clasificación: el primero es para la detección de emociones orientadas al aprendizaje en imágenes, utilizando una arquitectura de CNN para la clasificación de cinco clases utilizando el corpus CEI. Una vez que se realiza el entrenamiento, se guarda la configuración del grafo y los pesos del modelo de clasificación. (4-10).

Después se crea un segundo modelo para la detección de emociones orientadas al aprendizaje en textos, utilizando una arquitectura de CNN + LSTM con una capa de embebidos como entrada, entrenado con el corpus CES. Se realiza el entrenamiento y se guarda la configuración del grafo y los pesos del modelo de clasificación. (11-17).

A continuación, se crea un nuevo modelo híbrido (una combinación de modelos CNN y CNN-LSTM), en el que se reproducen las topologías de las capas de extracción de características de los dos modelos previamente entrenados, utilizando los grafos previamente entrenados (18-21).

Posteriormente se crea una capa de concatenación, en la que se concatenan las características extraídas por ambos sistemas de extracción de características, y se crea una red densamente conectada para la clasificación de las emociones en ambas modalidades (22-23).

Por último, se inicializan los pesos de las capas de extracción de características de los modelos guardados anteriormente, se compila, se entrena y se evalúa el modelo (24-29). Se presenta la arquitectura final del modelo en la Figura 4-11.

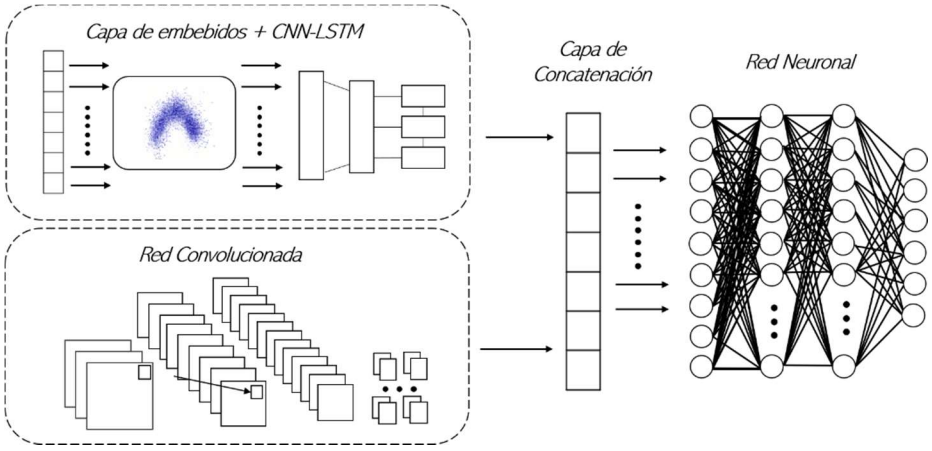


Figura 4-14. Proceso de conversión de imagen a vector.

Debido a que los conjuntos de datos provienen de diferentes fuentes de información, es importante remarcar que para el entrenamiento es necesario utilizar etiquetas idénticas en ambas representaciones unimodales de entrada.

4.4. Aplicación de modelos multimodales en ambientes de aprendizaje

Para el proceso generalizado de implementación de modelos multimodales en ambientes de aprendizaje se divide el trabajo por módulos, facilitando las pruebas y su intercambiabilidad.

Se utilizan dos módulos de recolección de datos dentro del sistema, un módulo de obtención imagen y otro modulo para la obtención de la opinión en texto del usuario. La información de estos módulos es vaciada en un módulo de preprocesamiento, el cual convierte las representaciones a través del modelo de representación utilizado. Esto puede ser la representación basada en imágenes, basada en embebidos de información o la representación en embebidos de palabras para el sistema de fusión de características. Por último, se utiliza un módulo de clasificación con el cual se obtiene la clase a la que pertenece la emoción del usuario, como se muestra en la Figura 4-12.

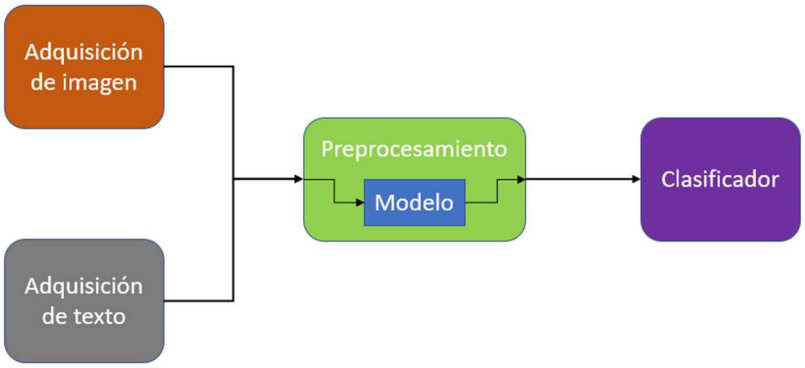


Figura 4-15. Diagrama de módulos para la implementación de clasificadores multimodales.

Una vez realizada la clasificación, esta se utiliza en un algoritmo de toma de decisiones. En este caso se utiliza un sistema de lógica difusa, el cual utiliza la emoción clasificada junto con diferentes valores cognitivos del usuario para la toma de decisiones. La Figura 4-13 16 muestra el sistema implementado con la clasificación multimodal de emociones implementando un sistema de lógica difusa.

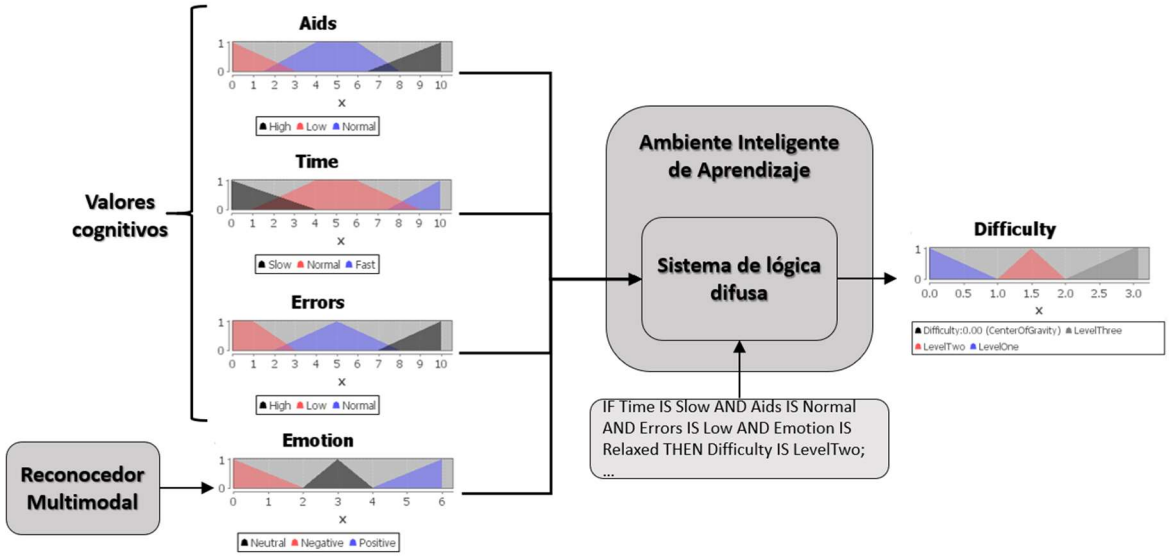


Figura 4-16. Sistema con implementación de reconocimiento multimodal.

Capítulo 5

5.Experimento

En esta parte se explica el proceso de haber sometido los diferentes modelos diseñados durante el capítulo 4 a pruebas, los resultados obtenidos y las diferentes metodologías utilizadas para verificar su validez.

5.1. Pruebas de modelos multimodales

Las métricas de las pruebas de los sistemas multimodales están basadas en las métricas de valor de precisión y valor de pérdida (descritas en el capítulo 2.3) utilizadas para la evaluación de modelos de ML.

Se evaluaron las arquitecturas descritas en los trabajos anteriores para crear modelos unimodales que permitieran poner una línea base para la comparación de las métricas de precisión y perdida, a su vez se realizaron experimentos con las nuevas arquitecturas multimodales evaluando las mismas métricas para crear tablas comparativas.

Una vez que se tenían resultados se realizó un benchmarking comparativo con los trabajos de (González-Hernández et al. 2018) con los sistemas multimodales de DL, descritos en este trabajo con la finalidad de observar las diferencias en la precisión de clasificación con respecto a los trabajos anteriores.

5.1.1. Preparación

Para la preparación del experimento se separaron los conjuntos de datos con una validación cruzada estratificada. El objetivo fue realizar una separación uniforme de las clases de manera porcentualmente equitativa, con un criterio de separación de 80% de los datos en un subconjunto para realizar el entrenamiento, y 20% de los datos en otro subconjunto para la evaluación. Esto con el fin de reducir el ovefitting en la clasificación. Para ello, se realizaron cinco iteraciones de validación cruzada por experimento.

De la misma manera se realizó el alineamiento de los datos utilizando las técnicas descritas en el capítulo 4.

5.1.2. Ejecución

Todos los experimentos se repitieron cinco veces, obteniendo un promedio del valor de la precisión de los diferentes modelos, así como su valor de perdida en la clasificación.

Los experimentos fueron realizados en una computadora laptop con un procesador Intel® Core™ I7-6700HQ con un GPU Nvidia Gtx 1060 y 16 gb de ram. Cada experimento tomó aproximadamente 30 minutos, exceptuando los experimentos que utilizaron la arquitectura CNN-LSTM con una capa de embebidos, los cuales tomaron entre 4 y 12 horas.

5.1.3. Resultados

Los resultados de las pruebas se muestran a continuación en la tabla 5-1. En esta tabla se pueden observar el conjunto de datos utilizados, el modelo implementado, la precisión y perdida obtenida en promedio, el tamaño de los conjuntos y el número de clases que manejó cada clasificador.

Tabla 5-1. Comparativa de resultados de experimentación con arquitecturas unimodales y multimodales

Conjunto	Modelo	Precisión	Perdida	Tamaño	Clases
CST 2018	CNN+LSTM	79.08%	1.72%	7492	2
CST 2019a	CNN+LSTM	87.28%	1.25%	16499	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	CNN + AG	82.13%	1.53%	5056	5
(CEI+CST2019b)	SRBI	62.13%	4.20%	50164	10
(CEI +CST2019b)	SRBE	86.20%	1.35%	50164	10
(CEI+CES2019b)	SFC	81.16%	1.57%	9560	5

De la misma manera, con los resultados se realizó un benchmarking con trabajos anteriores en función del porcentaje de precisión en la clasificación de los modelos. Los resultados de este benchmarking se muestran en la tabla 5-2.

Tabla 5-2. Benchmarking de sistemas unimodales de clasificación de emociones en rostro contra sistemas multimodales

Algoritmo de Clasificación	LBP	GB	CF
KNN	70%	61%	65%
ANN	74%	51%	61%
SVM	69%	61%	63%

CNN	-	-	74%
CNN+AG	-	-	82%
SRBI	-	-	62%
SRBE	-	-	86%
SFC	-	-	81%

5.1.4. Discusión

De los experimentos realizados podemos concluir que el uso de técnicas de fusión temprana depende en un alto grado de las técnicas de representación y alineamiento de los datos. Esto resulta lógico debido a que la precisión en la clasificación está fuertemente ligada a las características obtenidas de los datos, y estas a su vez están fuertemente ligadas a la metodología de extracción de características. Debido a que cada arquitectura utiliza una técnica diferente (Ej. Convolución en CNN), es importante representar la información en un formato que permita la correcta extracción de estas características.

Entre los modelos de representación utilizados (Imagen y Embebidos) el modelo de embebidos de información resultó ser el que arrojó mejores resultados, debido en gran parte a la cantidad y formato de la información presentada al clasificador durante su entrenamiento y validación (matriz de embebidos de información).

En los experimentos de evaluación el modelo multimodal SRBI presentó los peores resultados (62.13 %). El modelo multimodal SRBE presentó resultados prometedores con 86.20 % de precisión en la etapa de validación. Los modelos unimodales de arquitectura CNN-LSTM que utilizaron SentiText obtuvieron 91% de precisión en clasificación de polaridad (solamente dos clases) en texto usando el CST2019b y 69.27% de precisión con el conjunto de datos EduSere clasificando 5 clases. Los modelos de CNN obtuvieron 74.16% en clasificación de emociones para expresiones faciales (cinco clases) utilizando el conjunto EmotivInsight.

Por otro lado, el modelo SFC presentó buenos resultados con una precisión del 81.16 % comparado con los modelos unimodales para el reconocimiento de emociones en imágenes y texto. Este último aplicando técnicas de fusión de características a los mismos conjuntos de datos, a diferencia del SRBI y SRBE que realizan fusión de datos.

5.2. Caso de estudio de reconocedores multimodales para ambientes de aprendizaje

Para probar el proceso de reconocimiento multimodal de las expresiones faciales con opiniones, se integró el sistema en un entorno de aprendizaje inteligente (ILE) desarrollado por estudiantes del Instituto Tecnológico de Culiacán, para la práctica del lenguaje de programación Java. Dentro del ILE, un estudiante escribe código en Java para resolver diferentes problemas con diferentes niveles de complejidad. Cuando el estudiante está editando, compilando y ejecutando un programa, el ILE toma en cuenta varias variables pedagógicas, como el tiempo, los errores y las ayudas, y la variable afectiva Emoción obtenida por el reconocedor multimodal. La Figura 5-1 presenta 4 escenarios diferentes donde el estudiante escribe el código, lo compila y lo ejecuta con éxito, recibiendo mensajes del ILE, que obtiene las imágenes del estudiante de su rostro y las entradas de texto con opiniones, además de otras variables ya mencionadas.

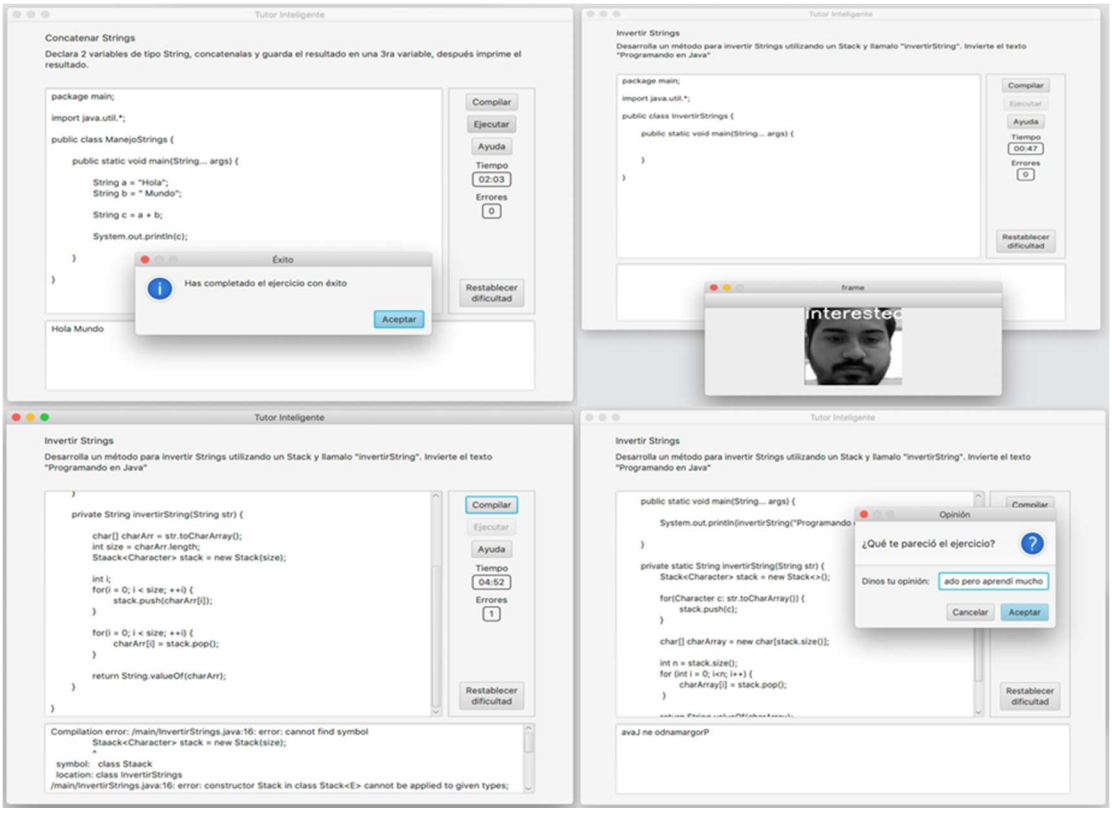


Figura 5-1. Ejemplo del ILE para reforzar programación en java.

Dentro de la ILE hay un módulo que determina, mediante lógica difusa, la complejidad del siguiente ejercicio para el estudiante. Para esto, el módulo tiene en cuenta los valores de las

variables pedagógicas (por ejemplo, el tiempo) y el valor de la variable afectiva Emoción. El módulo tiene un conjunto de reglas difusas que llevan a cabo mediante inferencias la determinación del siguiente problema. Hay tres niveles de ejercicios: básico, intermedio y avanzado.

Un ejemplo de una regla difusa es:

Si (el tiempo es grande) y (los errores son numerosos) y (Las asistencias son varias) y (La emoción es aburrida) Luego (El nivel de estudiante es básico)

En este caso, la regla difusa determina que el estudiante continúa en el nivel básico debido a sus malos resultados durante la solución del ejercicio.

Capítulo 6

6.Conclusiones y trabajo futuro

En este capítulo se presentan la conclusión de los resultados obtenidos de este proyecto de tesis de maestría. En este trabajo se presentan todas las tareas necesarias para realizar un experimento de reconocimiento de emociones multimodal orientadas al aprendizaje basadas en técnicas de fusión temprana.

6.1. Conclusiones del proyecto

Este proyecto consistió en la creación de una metodología de fusión y alineamiento de datos multimodales de texto y rostro para su implementación en ambientes inteligentes de aprendizaje, buscando agregar la parte emotiva a las variables cognitivas para la toma de decisiones.

La experimentación consistió en la creación de modelos de DL para la clasificación de las diferentes modalidades anteriormente fusionadas, utilizando modelos anteriormente probados en las modalidades de texto e imagen para realizar un benchmarking comparativo del desempeño de estos sistemas.

Esta metodología diseñada e implementada a través de los experimentos realizados mostró una mejoría significativa en el proceso de clasificación de las emociones de los estudiantes en ambientes controlados, demostrando un avance en la precisión de clasificación de hasta un 8% con respecto a contrapartes unimodales.

A su vez, se mostró que, con la creación de estos sistemas de fusión, es posible fusionar más de un tipo de información, y gracias a esto es posible decir que de agregar nuevas modalidades la precisión en estos modelos de clasificación podrá ser aún mayor.

6.2. Aportaciones y limitaciones

Este trabajo aporta por un lado una metodología nueva para la creación de sistemas de fusión temprana basada en las representaciones, adaptando representaciones a modelos computacionales de aprendizaje profundo de manera que, no es necesario crear nuevas arquitecturas para realizar el aprendizaje máquina multimodal. Además, demuestra el hecho de que es importante el medio de representación de los datos para poder llevar a cabo el entrenamiento de un modelo computacional de DL. Un ejemplo claro de esto es observar al sistema con representación basado en imágenes, que presenta una pérdida en la precisión de clasificación debido a la ambigüedad de la representación en imagen del texto.

También, este trabajo presenta un aporte en el área de fusión de conjuntos de datos unimodales, debido a que se ha demostrado que mientras se respete una correcta alineación de los datos, es decir un modelo sobre el cual se alinean datos que son semánticamente similares, es posible realizar la fusión de conjuntos de datos elaborados con diferentes muestras de individuos, en diferente tiempo, y aun así mantener una clasificación satisfactoria.

Por otro lado, el trabajo también presenta varias limitaciones. La primera de ellas es el tiempo de entrenamiento de los diferentes modelos computacionales, siendo este uno de los puntos importantes a tratar en trabajos futuros (los sistemas con capas de embebidos toman demasiado tiempo de entrenar y por lo tanto las pruebas fueron limitadas). Por otra parte, los conjuntos de datos presentan un desbalance que entorpece la clasificación de diversas clases, especialmente las clases que tienen menos cantidad de datos, ya que estas presentan menor incidencia y por lo tanto es importante expandir estas bases de datos.

6.3. Trabajo a futuro

Para trabajo futuro se propone el diseño de un framework para la creación de modelos de reconocimiento de emociones multimodales con el fin de obtener un modelo, que utilice todos los sensores de una computadora personal para realizar la evaluación de emociones del humano. Además de esto se propone la evaluación de este framework a través del modelo de dinámicas del proceso cognitivo, que permita la creación de relaciones de acción/reacción de

los sistemas, relacionando directamente las acciones de los tutores afectivos con los diferentes estados emocionales cambiantes de los usuarios en el uso de ellos, permitiendo así un control más específico de esta dinámica cognitiva propuesta por D'Mello.

Una mejora importante a este trabajo será el uso de técnicas de reforzamiento de aprendizaje, que permitan mejorar el uso de los reconocedores en usuarios específicos, utilizando los modelos de este experimento como base para su implementación y adaptación a cada usuario específicamente.

Finalmente, otro trabajo futuro posible es la implementación de los modelos obtenidos para su implementación en dispositivos Android, con el fin de utilizarlos dentro de aplicaciones móviles para realizar pruebas en ambientes no controlados.

Bibliografía

Baltrušaitis, Tadas, Chaitanya Ahuja, and Louis-Philippe Morency. 2016. “Multimodal Machine Learning: A Survey and Taxonomy.” https://arxiv.org/pdf/1705.09406.pdf.

Binali, Haji, Chen Wu, and Vidyasagar Potdar. 2010. “Computational Approaches for Emotion Detection in Text.” In *4th IEEE International Conference on Digital Ecosystems and Technologies*, 172–77. IEEE. https://doi.org/10.1109/DEST.2010.5610650.

Bishop, Christopher M. 2006. *Pattern Recognition and Machine Learning*. http://users.isr.ist.utl.pt/~wurmd/Livros/school/Bishop - Pattern Recognition And Machine Learning - Springer 2006.pdf.

Coen, Michael H. 1998. “Design Principles for Intelligent Environments.” www.aaai.org.

Cun, Y Le. 1989. “Generalization and Network Design Strategies.” http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.476.479&rep=rep1&type=pdf.

D’Mello, Sidney, and Art Graesser. 2012. “Dynamics of Affective States during Complex Learning.” *Learning and Instruction* 22 (2): 145–57. https://doi.org/10.1016/J.LEARNINSTRUC.2011.10.001.

Darwin, Charles, M.A., and F.R.S. 1872. *Expression of the Emotions in Man and Anmials*. London. https://pure.mpg.de/rest/items/item_2309885/component/file_2309884/content.

Dodge, Yadolah. 2008. *The Concise Encyclopedia of Statistics*. Springer.

Ekman, Paul., and Erika L. Rosenberg. 1997. *What the Face Reveals : Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*. Oxford University Press.

Ekman, Paul. 1992. “An Argument for Basic Emotions.” *Cognition and Emotion* 6 (3–4): 169–200. https://doi.org/10.1080/02699939208411068.

Erb, Randall J. 1993. “Introduction to Backpropagation Neural Network Computation.” *Springer*. https://link.springer.com/article/10.1023/A:1018966222807.

Gers, Felix A., Jürgen Schmidhuber, and Fred Cummins. 2000. “Learning to Forget: Continual Prediction with LSTM.” *Neural Computation* 12 (10): 2451–71. https://doi.org/10.1162/089976600300015015.

Gogate, Mandar, Ahsan Adeel, and Amir Hussain. 2017. “A Novel Brain-Inspired Compression-Based Optimised Multimodal Fusion for Emotion Recognition.” In *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, 1–7. IEEE. https://doi.org/10.1109/SSCI.2017.8285377.

Gomez-Uribe, Carlos A., and Neil Hunt. 2015. “The Netflix Recommender System.” *ACM Transactions on Management Information Systems* 6 (4): 1–19. https://doi.org/10.1145/2843948.

González-Hernández, Francisco, Ramon Zatarain-Cabada, Maria Lucia Barrón-Estrada, and Hector

Rodríguez-Rangel. 2018. “Recognition of Learning-Centered Emotions Using a Convolutional Neural Network.” Edited by David Pinto, Vivek Kumar Singh, Aline Villavicencio, Philipp Mayr-Schlegel, and Efstathios Stamatatos. *Journal of Intelligent & Fuzzy Systems* 34 (5): 3325–36. https://doi.org/10.3233/JIFS-169514.

Hecht-Nielsen, R. 1988. “Neurocomputing: Picking the Human Brain.” *IEEE Spectrum* 25 (3): 36–41. https://doi.org/10.1109/6.4520.

Himani Sharma, Sunil Kumar. 2016. “A Survey on Decision Tree Algorithms of Classification in Data Mining.” https://www.researchgate.net/publication/324941161_A_Survey_on_Decision_Tree_Algorithms_of_Classification_in_Data_Mining.

Hochreiter, Sepp, and J J Urgen Schmidhuber. 1997. “Long Short-Term Memory.” *MEMORY Neural Computation* 9 (8): 1735–80. http://www7.informatik.tu-muenchen.de/~hochreithttp://www.idsia.ch/~juergen.

Hu, Anthony, and Seth Flaxman. 2018. “Multimodal Sentiment Analysis To Explore the Structure of Emotions,” May. https://doi.org/10.1145/3219819.3219853.

Huang, Yongrui, Jianhao Yang, Siyu Liu, Jiahui Pan, Yongrui Huang, Jianhao Yang, Siyu Liu, and Jiahui Pan. 2019. “Combining Facial Expressions and Electroencephalography to Enhance Emotion Recognition.” *Future Internet* 11 (5): 105. https://doi.org/10.3390/fi11050105.

Ian Goodfellow, Yoshua Bengio, Aaron Courville. 2016. “Deep Learning.” 2016. https://books.google.com.mx/books?hl=es&lr=&id=omivDQAAQBAJ&oi=fnd&pg=PR5&dq=deep+learning+introduction&ots=MMR5euoITT&sig=mLSUIWruXgGJUc5Hu0yoGyMtgPw#v=onepage&q=deep learning&f=false.

Ioffe, Sergey, and Christian Szegedy. 2015. “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift,” February. http://arxiv.org/abs/1502.03167.

Jain, A.K., Jianchang Mao, and K.M. Mohiuddin. 1996. “Artificial Neural Networks: A Tutorial.” *Computer* 29 (3): 31–44. https://doi.org/10.1109/2.485891.

Kanjo, Eiman, Eman M.G. Younis, and Chee Siang Ang. 2019. “Deep Learning Analysis of Mobile Physiological, Environmental and Location Sensor Data for Emotion Detection.” *Information Fusion* 49 (September): 46–56. https://doi.org/10.1016/J.INFFUS.2018.09.001.

Kecman, V. (Vojislav). 2001. *Learning and Soft Computing : Support Vector Machines, Neural Networks, and Fuzzy Logic Models*. MIT Press.

Khorrami, Pooya, Thomas Paine, and Thomas Huang. 2015. “Do Deep Neural Networks Learn Facial Action Units When Doing Expression Recognition?” https://www.cv-foundation.org/openaccess/content_iccv_2015_workshops/w2/html/Khorrami_Do_Deep_Neural_ICCV_2015_paper.html.

Kim, Eesung, and Jong Won Shin. 2019. “DNN-Based Emotion Recognition Based on Bottleneck Acoustic Features and Lexical Features.” In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6720–24. IEEE. https://doi.org/10.1109/ICASSP.2019.8683077.

Leal Hernandez, Daniel. 2017. “Reconocimiento de Emociones Centradas En El Aprendizaje Por

Medio de Expresiones Faciales." Tecnologico de Culiacan.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. 2015. "Deep Learning." *Nature* 521 (7553): 436–44. https://doi.org/10.1038/nature14539.

Liang, Jingjun, Shizhe Chen, Jinming Zhao, Qin Jin, Haibo Liu, and Li Lu. 2019. "Cross-Culture Multimodal Emotion Recognition with Adversarial Learning." In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4000–4004. IEEE. https://doi.org/10.1109/ICASSP.2019.8683725.

Miao, Haotian, Yifei Zhang, Weipeng Li, Haoran Zhang, Daling Wang, and Shi Feng. 2018. "Chinese Multimodal Emotion Recognition in Deep and Traditional Machine Leaming Approaches." In *2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia)*, 1–6. IEEE. https://doi.org/10.1109/ACIIAsia.2018.8470379.

Negnevitsky, Michael Negnevitsky. 2002. *Artificial Intelligence A Guide to Intelligent Systems Artificial Intelligence Second Edition*. www.pearsoned.co.uk.

Oramas-Bustillos, Raul, Maria Lucia Barron-Estrada, Ramon Zatarain-Cabada, and Sandra Lucia Ramirez-Avila. 2018. "A Corpus for Sentiment Analysis and Emotion Recognition for a Learning Environment." In *2018 IEEE 18th International Conference on Advanced Learning Technologies (ICALT)*, 431–35. IEEE. https://doi.org/10.1109/ICALT.2018.00109.

Patwardhan, Amol S. 2017. "Multimodal Mixed Emotion Detection." In *2017 2nd International Conference on Communication and Electronics Systems (ICCES)*, 139–43. IEEE. https://doi.org/10.1109/CESYS.2017.8321250.

Picard, Rosalind W. 1997. *Affective Computing Focus on Emotion Expression, Synthesis and Recognition*. http://zums.ac.ir/files/research/site/ebooks/Computer Science and Engineering/Affective Computing.pdf.

Popova, Anastasiya S., Alexandr G. Rassadin, and Alexander A. Ponomarenko. 2018. "Emotion Recognition in Sound," 117–24. https://doi.org/10.1007/978-3-319-66604-4_18.

Powers, and David M W. 2011. "EVALUATION: FROM PRECISION, RECALL AND F-MEASURE TO ROC, INFORMEDNESS, MARKEDNESS & CORRELATION." *Journal of Machine Learning Technologies* 2 (1): 37–63. http://dspace.flinders.edu.au/dspace/http://www.bioinfo.in/contents.php?id=51.

Quinlan, J R. 1986. "Induction of Decision Trees." *Machine Learning*. Vol. 1. http://hunch.net/~coms-4771/quinlan.pdf.

Sammut, Claude, and Geoffrey I. Webb, eds. 2010. *Encyclopedia of Machine Learning*. Boston, MA: Springer US. https://doi.org/10.1007/978-0-387-30164-8.

Schalkoff, RJ. 1997. "Artificial Neural Networks." https://scholar.google.com/scholar_lookup?title=Artificial Neural Networks&publication_year=1997&author=R.J Schalkoff.

Seyeditabari, Armin, Narges Tabari, and Wlodek Zadrozny. 2018. "Emotion Detection in Text: A Review." https://arxiv.org/pdf/1806.00674.pdf.

Silva, L.C. De, T. Miyasato, and R. Nakatsu. 1999. "Facial Emotion Recognition Using Multi-Modal Information." In *Proceedings of ICICS, 1997 International Conference on Information,*

Communications and Signal Processing. Theme: Trends in Information Systems Engineering and Wireless Multimedia Communications (Cat. No.97TH8237), 1:397–401. IEEE. https://doi.org/10.1109/ICICS.1997.647126.

Valstar, M. F., M. Mehu, Bihan Jiang, M. Pantic, and K. Scherer. 2012. "Meta-Analysis of the First Facial Expression Recognition Challenge." *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 42 (4): 966–79. https://doi.org/10.1109/TSMCB.2012.2200675.

Wei, Wei, Qingxuan Jia, and Yongli Feng. 2017. "Emotion Recognition Based on Feedback Weighted Fusion of Multimodal Emotion Data." In *2017 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, 1682–87. IEEE. https://doi.org/10.1109/ROBIO.2017.8324660.

Yang, Lingzhi, Xiaojuan Ban, Michele Mukeshimana, and Zhe Chen. 2019. "Multimodal Emotion Recognition Using the Symmetric S-ELM-LUPI Paradigm." *Symmetry* 11 (4): 487. https://doi.org/10.3390/sym11040487.

Zatarain Cabada, Ramón Lucia, Maria, Barron Estrada, Hector Rodriguez Rangel, Francisco González-Hernández, Ramon Zatarain-Cabada, Maria Lucia Barrón-Estrada, and Hector Rodrí-Guez-Rangel. 2018. "Recognition of Learning-Centered Emotions Using a Convolutional Neural Network." *Article in Journal of Intelligent and Fuzzy Systems*. https://doi.org/10.3233/JIFS-169514.

Zhou, and Chellappa. 1988. "Computation of Optical Flow Using a Neural Network." In *IEEE International Conference on Neural Networks*, 71–78 vol.2. IEEE. https://doi.org/10.1109/ICNN.1988.23914.

Zhou, Jian, Xianwei Wei, Chunling Cheng, Qidong Yang, and Qun Li. 2019. "Multimodal Emotion Recognition Method Based on Convolutional Auto-Encoder." *International Journal of Computational Intelligence Systems* 12 (1): 351. https://doi.org/10.2991/ijcis.2019.125905651.