UNIVERSIDAD DEL PACÍFICO

TRABAJO FINAL DE INGENIERÍA DE LA INFORMACIÓN

Clasificación de emociones a partir de la voz en llamadas telefónicas usando Convolutional Neural Network

---

Autores:
Inche, Bryan
Tiger, Jacomo
Uriarte, Erick

Profesor del Curso:
Victor Hugo Ayma

1 de octubre de 2021

Índice general

Índice de figuras

Índice de cuadros

Capítulo 1

Introducción

1.1. Contexto General

Actualmente, en la sociedad se tiene como tendencia el uso del teléfono móvil e internet. Este último ha sido una de las tecnologías que ha crecido de manera más rápida en los últimos años [1]. El número de usuarios entre los años 2000 al 2016 ha pasado de 413 millones a 3.4 billones respectivamente [1]. Según *Statista*, el número de personas que utiliza el internet actualmente es, incluso, de aproximadamente 4.6 billones de personas [2]. Hoy en día, las personas se pueden conectar a la web cómodamente desde los dispositivos móviles, causando así un gran aumento de usuarios. Según datos del Banco Mundial, el uso de los dispositivos móviles para ingresar a la web ha aumentado aproximadamente en un 55 % en la última década [3]. El uso de la red en dispositivos móviles afecta distintos sectores tales como educación, entretenimiento, comunicación y ventas. En el caso específico de ventas, el uso de dispositivos causa una gran oportunidad para las empresas debido a que estas pueden llegar a vender sus productos por la web para poder aprovechar su mercado potencial al máximo [4].

El intercambio de servicios o productos con la ayuda de una red privada o pública, incluyendo el internet, para obtener un valor agregado, se conoce como comercio electrónico [5]. Esto puede ser muy ventajoso para el consumidor igualmente, pues desde la comodidad de su casa puede escoger de diferentes plataformas en línea un producto que desee y puede compararlo con otros productos parecidos para poder finalmente conseguir la mejor opción [5]. El comercio electrónico ha sido tendencia de igual manera en los últimos años, aumentando en un 151 % desde el 2014 al 2019 [6]. En el 2019, de los 3,354 billones de dólares en ventas globalmente de comercio electrónico, 60 % se dio desde dispositivos móviles. Esto quiere decir que gracias al impulso del uso de estos dispositivos, se tiene como tendencia las compras por internet.

Esta tendencia, así como otros comportamientos de la sociedad, fueron duramente impactados a finales del 2019 por el COVID-19 [7]. Este virus es altamente contagioso y puede causar neumonía, bronquitis e incluso la muerte [8]. Alrededor del mundo, han muerto aproximadamente 3.8 millones de personas desde inicios del 2020 a quincena de junio del 2021. [9]. Solamente en Estados Unidos, al menos a 316 millones de personas se les pidió quedarse en sus casas para disminuir el contagio. Debido a restricciones por acercamiento, las personas han optado por comprar de manera virtual [10]. Es decir, que el comercio electrónico se ha vuelto aún mayor tendencia. El virus ha sido un acelerador del cambio de estructura de consumo

y ha aumentado la cantidad de usuarios y ventas del comercio electrónico [10]. Incluso, el World Trade Organization indica que el comercio electrónico ha sido una herramienta y solución importante para los consumidores en tiempos de crisis [11].

Alrededor del mundo se evidencia el aumento de compras en línea debido a la pandemia [7]. Globalmente, se puede evidenciar un aumento en las compras en línea en los sectores de libros y literatura (+16 %), medicina (+9 %), decoraciones de hogar (+7 %), ventas minoristas (+6 %) y moda (+5 %) [12]. En el caso de Estados Unidos, las ventas de *Walmart* por comercio electrónico incrementaron en un 74 %. Asimismo, las compañías de comunicación tales como *Facebook*, *Zoom*, y *Google Meets* también incrementaron su tráfico [7]. Igualmente, en Colombia se pudo observar un aumento entre 50 % y 80 % de uso de comercio electrónico según la Cámara Colombiana de Comercio Electrónico [13] . De la misma manera, en el caso de Pakistán, los usuarios aumentaron en un 10 % [7]. Finalmente, en el caso de Perú, antes de la pandemia el 1.5 % de las ventas en los comercios era a través de *e-commerce*. Posteriormente, las empresas que entraron al comercio electrónico se cuadruplicaron y al cierre del 2020, el 5 % ya vendía por internet [14]. Según la Cámara Peruana de Comercio Electrónico (CAPECE), la transformación digital esperada de 5 años, se dio en 6 meses debido a la pandemia. Se proyecta que para el 2021 la participación del comercio electrónico sobre el total del consumo sea aproximadamente 40 % [14]. La Cámara de Comercio de Lima indicó que 9 millones de peruanos inclusive compraron por internet durante la pandemia.

Hoy por hoy, el estudio del comportamiento del consumidor se ha vuelto relevante y significativo pues la demanda masiva del comercio electrónico ha generado cambios importantes en el contexto actual de la pandemia [7]. El comportamiento del consumidor es definido como "el comportamiento que los consumidores muestran al buscar, comprar, usar, evaluar y rechazar productos y/o servicios que esperan satisfagan sus necesidades" [15]. Esto se da no solo de forma endógena sino exógena también. Dado que el consumidor no solo se enfrenta al producto/servicio sino a un entorno que lo rodea para establecer relaciones con estos cuando se enfrenta a la decisión de compra final. Asimismo, incluye el qué, porqué, cuándo, dónde lo compran. Además, con qué frecuencia y cómo es usado, cómo lo evalúan luego de realizar la compra y qué efectos trae para compras futuras y, finalmente, cómo lo desechan. Es de importancia conocer también que todos los seres humanos somos consumidores y que las decisiones sobre las compras influyen en más eslabones de una cadena amplia de agentes [16]: desde la extracción de materias primas hasta la puesta del producto en una tienda.

Los cambios en el contexto de COVID-19 se están produciendo en todos los ámbitos en la vida de los consumidores. Tal como lo menciona la consultora *McKinsey* son 8 las áreas en donde han emergido cambios importantes en la vida del consumidor. Por ejemplo, en el ámbito de compras y consumo se evidencia un aumento en el comercio electrónico, se ha dado preferencias a marcas confiables, el gasto discrecional ha bajado, compras más grandes pero reducción en las frecuencias de compras, preferencias a tiendas cercanas y polarización de la sostenibilidad son los síntomas más resaltantes en este apartado. Asimismo, la tendencia en el formato de *delivery* de compras online realmente ha significado una aceleración de algo que ya antes ha ido existiendo. Diez años en 8 semanas ha sido la aceleración en entregas de *delivery* por *e-commerce* en el mundo [17]

En un estudio publicado en el *Transnational Marketing Journal* [18] se demuestra que características demográficas como nivel de ingresos, clase social y estado marital son factores importantes que muestran una frecuencia en las decisiones de compra de los consumidores. Asimismo, el factor ansiedad muestra un impacto en los efectos de estas características. La ansiedad demuestra un efecto claro en personas de alto nivel de ingresos y por tanto indirectamente en su comportamiento de compras [18]. Además, los consumidores se vuelven pesimistas debido a la incertidumbre y la inseguridad durante la pandemia de Covid-19, y sus comportamientos de compra, en consecuencia, cambian. Esto demuestra que el estudio de las emociones en el comportamiento del consumidor toma mayor importancia en estos tiempos de incertidumbre.

Se puede observar en estudios en diferentes lugares de el mundo que la salud mental de los ciudadanos ha sido afectada debido a la pandemia. El Departamento de Salud Mental y Abuso de Sustancias de la OMS reconoce que el número de personas expuestas a los factores estresantes extremos es grande, y que la exposición a estos constituye un factor de riesgo para el desarrollo de problemas sociales y de salud mental. De igual manera, existen circunstancias en la vida de los individuos, que pueden propiciar un mayor riesgo psicosocial ante la pandemia de la COVID-19 [19]. Estos factores que estresan a la población pueden ser la duración de la cuarentena, miedo al contagio, aburrimiento en el hogar, no tener lo necesario para sobrevivir, no estar informado correctamente, y la pérdida de ingresos o educación [20].

En en estudio de Puerto Rico se concluye que globalmente han aumentado los niveles de ansiedad, depresión, insomnio y temores generales. Esto se da más aún en trabajadores de salud que enfrentan a la enfermedad [21]. En el caso de El Salvador, se pudo observar que la población compraba solamente por pánico [22]. En Changzhi, China, Cao *et al.* [23] demostraron un aumento en ansiedad grave, leve y moderada. En España aumentó de igual manera el nivel de estrés, depresión, ansiedad, y miedo a contagiarse del virus [24]. Finalmente, en el Perú Essalud advirtió que en el 2021, hubo un incremento en la hospitalización de niños y adolescentes con depresión, el cual llego inclusive al 50 % [25].

Las emociones tienen un rol clave en la vida humana, y la voz es un medio por el cual las personas las expresan para poder dar a entender al receptor cómo se sienten [26]. Actualmente, debido a que muchas personas están conectadas a la red, y que las relaciones entre vendedores y compradores son, por una gran parte, mediante el comercio electrónico, existe un gran uso de herramientas de comunicación [27]. Mediante estas herramientas, las personas utilizan su voz, y por lo tanto es pertinente plantear un modelo que permita analizar las emociones de una persona mediante este canal. En los últimos años ha sido tendencia el análisis de los emociones mediante la voz e imágenes [27]. Estos recientes análisis tienen distintas aplicaciones tales como mejorar la comunicación entre las máquinas y las personas, mejorar el trato a los clientes [26], y finalmente se puede analizar desde un punto psiquiátrico en el que se puede distinguir entre personas que tienen o no depresión [28].

Para realizar un reconocimiento de emociones de la voz se deben analizar las características pertinentes, pues existen diversas características que serían insignificantes para el caso. Estas características se pueden observar en distintos dominios

tales como tiempo, frecuencia, imagen (spectograma), cepstral, etc [29]. Cada dominio maneja diversas técnicas para extraer lo requerido tal como la extracción de los coeficientes cepstrales en la frecuencia de mel para el dominio cepstral [29]. Por otro lado, se tienen distintos modelos de machine learning para realizar la clasificación de emociones de acuerdo a estas características extraídas. Estas incluyen *Support Vector Machines (SVM)* y *Convolutional Neural Networks (CNN)*, entre otras. Esta última técnica puede ser utilizada para caracterizar señales bidimensionales de manera efectiva gracias a que elimina la dependencia de la subjetividad o la experiencia humana [30].

1.2. Situación Problemática

Debido a la pandemia del COVID-19, se han producido dos principales fenómenos en la sociedad. En primer lugar, debido al distanciamiento social obligatorio, se ha reducido la cantidad de ventas presenciales, y se ha optado por una opción en línea (*e-commerce*) [10]. En distintos lugares del mundo se evidencia este proceso evolutivo acelerado hacia esta tendencia. Globalmente, las empresas en los distintos sectores tales como medicina, moda y *retail*, han evidenciado un aumento en las ventas en línea. En Estados Unidos, las ventas virtuales de *Walmart* aumentaron en un 74 % y las personas utilizan de mayor manera herramientas de comunicación tales como *Facebook*, *Zoom* y *Google Meets* [7]. Adicionalmente, en Colombia se observa un número de usuarios récord en cuanto a comercio electrónico y finalmente en Perú se llego a cuadruplicar el número de empresas que vende en línea.

En segundo lugar, la pandemia ha afectado la salud mental de las personas. Debido a esta, se han producido efectos psicológicos negativos que están relacionados directamente a las condiciones causadas por el propio confinamiento. [31]. Estas condiciones incluyen las posibles cuarentenas, el posible contagio, el posible contagio de un familiar, la separación y distancia necesaria entre personas y en el peor de los casos, la muerte.

En un estudio realizado en El Salvador, debido a la pandemia COVID-19 al menos un 40 % de la muestra participante habría incurrido en comportamientos propios de compras por pánico. Estas se correlacionan con síntomas emocionales como la depresión, la ansiedad y el estrés, así como con el interés en el tema de la pandemia [22]. Adicionalmente, en un estudio de España se analizó el perfil emocional de la población según los distintos grupos de edad [24]. Finalmente, se concluyó que los grupos de menor edad tuvieron mayores problemas de sueño en compraración con los otros [24].

Debido a que se ha vuelto una tendencia el uso de los canales no presenciales para la compra, las empresas se deben adaptar para conocer a los clientes y cómo se sienten estos con el producto y en general según distintos contextos. Es de suma importancia resaltar que en todas las áreas de negocio se busca una buena retención del cliente, el cual proviene de una buena atención. En el caso específico del comercio electrónico, todas las empresas están intentando diferenciarse y destacar, por lo tanto es vital tener una buena atención al cliente y calidad de servicio. Este incluye

no solamente la página web, sino todos los medios necesarios para cubrir necesidades. El principal medio de asistencia al comercio electrónico es el uso de llamadas telefónicas. [32]

Los resultados de un estudio realizado en Chile muestran que la mayor parte de las empresas en la etapa de la pandemia COVID-19 han estado monitoreando la actividad de venta y los resultados comerciales con medidas relacionadas con llamadas a clientes existentes, oportunidades de negocios creadas y cotizaciones enviadas. Asimismo, en el estudio mencionado se concluyó que las pandemias generan caídas en la productividad agregada de las economías de los países, lo que a su vez impacta en una caída mayoritaria en la actividad comercial y en las ventas de las organizaciones [33].

En ese sentido, a partir de la problemática planteada anteriormente se formula la siguiente pregunta de investigacion, ¿En qué medida las emociones a través de la voz producidas en las llamadas telefónicas afectan a las ventas en los canales no presenciales (Online)?

1.3. Objetivos

El presente trabajo tiene como objetivo principal determinar la importancia de las emociones clasificadas a partir de la voz en llamadas de ventas telefónicas.

- Objetivo específico 1: Construir un modelo de *machine learning* para clasificar emociones a partir de la voz.
- Objetivo específico 2: Construir un conjunto de datos de voz de llamadas telefónicas de ventas.
- Objetivo específico 3: Aplicar el modelo entrenado previamente en el conjunto de datos de ventas construido.
- Objetivo específico 4: Analizar los resultados de la clasificación de emociones en la base de datos de llamadas telefónicas.

1.4. Justificación

Para poder evidenciar la relevancia e importancia del presente trabajo, se utilizará como referencia los criterios formulados por Ackoff [34] y Miller [35]. Estos se pueden sintetizar principalmente en 5 ejes: conveniencia, relevancia social, implicaciones prácticas, valor teórico y utilidad metodológica.

En primer lugar, la investigación se estima conveniente porque sus resultados podrían ayudar a empresas a conocer con mayor profundidad a sus clientes. Debido a que está incrementando el comercio en línea (como vimos en la sección de problemática), la comunicación entre los clientes y los vendedores es a través de llamadas en muchos casos (normalmente en atención al cliente). Por este medio, es de gran importancia conocer cómo se está sintiendo el cliente durante la llamada. Así, se ayudará a empresas a mantener una mejor comunicación y un mejor trato hacia el cliente.

En segundo lugar, la investigación tiene relevancia social, pues el gran mercado creciente en línea podrá conocer de mejor manera al cliente, mejorando así finalmente las ventas. Adicionalmente, los compradores serán comprendidos de mejor manera y mejorarán las relaciones con los vendedores. Esto finalmente incitará a una mayor compra y un mayor contento por ambos lados.

Por último, el trabajo contiene utilidad metodológica, pues ayuda a crear un clasificador de emociones a partir de la voz proveniente de una base de datos (en español) de llamadas entre vendedores y clientes. Es posible que este estudio favorezca el avance en experimentos de marketing de empresas para una mayor comprensión y un mejor trato hacia el cliente. El presente trabajo cuenta con bases y técnicas sólidas que justifican su realización. Finalmente, la metodología usada podría ser replicada en distintos campos para poder clasificar emociones según distintos contextos.

Capítulo 2

Marco Teórico

A continuación, se presentará el marco teórico dividido en dos secciones. En primer lugar, se encuentran las bases teóricas para poder tener una mejor comprensión respecto a los temas expuestos. Incluidas en estas, se encuentra el funcionamiento de las técnicas utilizadas y sus fundamentos. En segundo lugar, se encuentra el estado del arte, el cual es el sustento del presente trabajo. Se explicarán y compararán trabajos pasados con el trabajo presente, para así resaltar la importancia de este y los futuros aportes posibles a la literatura.

2.1. Comercio Electrónico

Se define como comercio electrónico el intercambio de información, sin la necesidad de papel, entre compradores y vendedores utilizando solamente datos electrónicos, correos electrónicos, boletines electrónicos, el internet, u otras tecnologías basadas en redes [36]. El comercio electrónico trae consigo distintas ventajas para el vendedor. En primer lugar, se tiene obicuidad. Es decir, que el producto puede estar presente en cualquier lugar de la red, todo el tiempo [37]. En segundo lugar, el vendedor puede llegar a contactarse con un público más amplio, e incluso global, debido a que no tiene limitaciones geográficas [37]. En tercer lugar, se puede obtener información del cliente para un futuro análisis [37]. En cuarto lugar, los costos operacionales para el vendedor pueden disminuir, pues no se necesita una ubicación física para poder vender el producto [38]. Finalmente, se pueden automatizar procesos manuales o transacciones que anteriormente necesitaban de papel [36]. De igual manera, existen ventajas hacia el comprador, ya que este no tiene que lidiar con largas colas para un producto, tiene una mayor facilidad para poder comparar precios, mayor customización de parte del vendedor, y finalmente puede recibir el producto desde el confort de su hogar [38].

El comercio electrónico, sin embargo, trae consigo ciertas desventajas que se tienen que tener en cuenta de igual manera. En el caso del vendedor, existe cierto distanciamiento con el cliente, pues no existe una relación directa y presencial, por lo que el vendedor tiene que optar por distintos medios de comunicación [38]. Asimismo, existe una mayor competencia debido a los distintos productos de distintas marcas ofrecidos en la web [38]. También, los productos no pueden ser probados por los clientes previamente causando una posible duda en la compra [38]. Respecto al cliente, este tiene la necesidad de tener internet, y existe una mayor posibilidad de caer en fraude y no llega a recibir el producto mostrado [38].

2.1.1. Dimensiones del comercio electrónico

Un mercado está compuesto por tres distintos componentes, los cuales son los agentes, los productos y los procesos [39]. Los agentes son los vendedores, compradores, e intermediarios. Los productos vendrían a ser los tangibles que son intercambiados y finalmente, las interacciones entre agentes respecto a los productos u otras actividades vendrían a ser los procesos. Estos pueden incluir por ejemplo, la selección de un producto, producción, pago, cobranza, etc. [39]. Cada uno de estos tres agentes puede ser digital o físico, creando así el cubo mostrado a continuación en la figura 2.1.

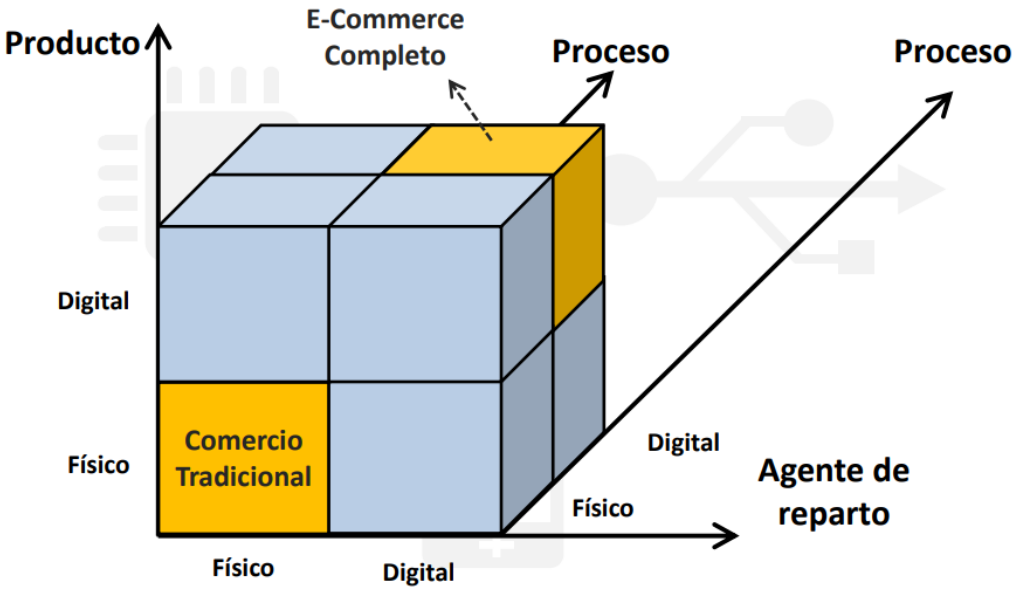

FIGURA 2.1: Dimensiones de mercado
[39]

Como podemos observar en la figura 2.1, cuando se tiene un agente de reparto físico, un producto físico y un proceso físico, se tiene un comercio tradicional. Esto puede ser por ejemplo, una tienda física de ropa en la cual el cliente tendrá que visitar físicamente para realizar la compra. Por lo contrario, se tiene el mercado de comercio electrónico completo. Este es cuando el proceso, el agente de reparto y el producto son digitales. Esto puede ser, por ejemplo, la compra y lectura de un libro en línea. Las empresas relevantes para el trabajo serían las que no tienen un proceso totalmente digitalizado. Es decir, que el contacto del vendedor con el cliente no es en línea por lo cual puede ser por llamada para el posible análisis. Cabe recalcar que la mayoría de los mercados en línea están bajo el área gris [39].

2.1.2. Tipos de comercio electrónico

Como se describió previamente, el comercio electrónico es simplemente el intercambio de información y la realización de una compra mediante un canal electrónico, sin la necesidad de un papel [36]. Existen ciertos tipos de compradores, así como vendedores, por lo que se separan los diferentes tipos de comercio electrónico que se pueden realizar [36].

En primer lugar, se tiene el *Business-to-Consumer* (B2C). En este caso, sería una empresa que vende sus productos o servicios mediante el internet para que el consumidor (una persona) pueda aprovechar los beneficios [36]. Un ejemplo de este caso sería *Amazon*, donde la empresa tiene una página en la que oferta sus productos y personas pueden llegar a comprarlos. En segundo lugar, se tiene *Business-to-Business* (B2B). En este caso, una empresa oferta un producto, servicio o información para que otra empresa finalmente lo compre [36]. Por ejemplo, *Cisco* oferta en línea productos, mantenimiento, o consultoría a empresas de telecomunicaciones.

En tercer lugar, se tiene el *Consumer-to-Consumer* (C2C). En este caso, tanto los compradores como los vendedores son personas individuales. Para este caso, un *marketplace* virtual sirve como nexo para conectar a los compradores y vendedores [36]. Un ejemplo para este caso es *Ebay*. Esta página permite que distintas personas ingresen distintos productos para que otras puedan comprarlos. Finalmente se tiene el último caso, *Consumer-to-Business*. En este caso, los individuos que utilizan el internet, tratan de vender sus productos o servicios a organizaciones o empresas que los necesitan [37]. Un ejemplo en esta situación es *CrowdCube*, una página en la que empresas pueden invertir en proyectos de individuos.

Si se toma en cuenta el gobierno, existen distintos tipos de comercio electrónico a tomar en cuenta para estos casos [36]. En primer lugar, el *Government-to-Government* (G2G). En este caso hay un intercambio de servicios entre gobiernos, ya sean locales o regionales [36]. En segundo lugar, se tiene el *Government-to-Customer* (G2C). Para este caso, el gobierno le ofrece un servicio virtual a una persona. Esto puede ser obtener un certificado de nacimiento [36]. En tercer lugar, se tiene *Business-to-Government* (B2G). En este caso, la empresa mediante un canal virtual le ofrece un producto o servicio al gobierno [40]. Finalmente, se tiene el caso de *Government-to-Business* (G2B). En este caso, el gobierno le ofrece información a las empresas acerca de temas gubernamentales [40].

2.1.3. Tipos de Organizaciones en *E-Commerce*

Existen distintos tipos de empresas que pueden llevar a cabo comercio electrónico [37]. En primer lugar, se encuentran las *Brick and Mortar*. Estas empresas utilizan el comercio tradicional, mas no el virtual. Es decir, no utilizan el comercio electrónico. Este tipo de empresas en su mayoría tienen productos físicos en tiendas físicas [37]. Por ejemplo, un vendedor de fruta que solamente vende físicamente en el mercado.

En segundo lugar, se tiene a las empresas *Click and Mortar*. Estas empresas tienen gran parte de sus ventas en línea (utilizando el comercio electrónico), mas no son todas [37]. Estas empresas pueden tener tanto tiendas virtuales como físicas. Por ejemplo, el caso de *Walmart*, que puede llegar a vender mediante el comercio electrónico, o mediante sus tiendas físicas. Finalmente, las empresas virtuales, son las que conducen sus negocios de manera totalmente en línea, sin una tienda física [37]. Esto podría ser una cocina que reparte comida solamente comprada en línea. Es importante tener en cuenta que una empresa virtual no necesariamente tiene todos los procesos de manera electrónica, y por lo tanto, no necesariamente es una *E-Business*.

2.2. Emociones

Las emociones han sido un campo de exploración desde hace miles de años, y se han visto a través de la historia desde diferentes puntos de vista [41]. A través de los años, las han estudiado distintos filósofos y psicólogos tales como Aristóteles, Descartes, David Hume, William James, Robert Solomon, y muchos otros [41]. Estos distintos puntos de vista causan ciertos debates teóricos. Por ejemplo, David Hume clasifica las emociones en calmadas y violentas. Las calmadas responden a sentmientos psicológicos como la aprobación moral o el goce estético, mientras que las violentas no necesitan ir acompañadas de sensaciones físicas definidas. Posteriormente, William James descarta todas las teorías anteriores que abarcan a la emoción como un concepto más o menos inteligente. Él define a las emociones simplemente como reacciones fisiológicas tales como aumento en la velocidad del latido del corazón, dilatación en los vasos sanguíneos, etc. [41].

Para el presente trabajo se tomará en cuenta la definición de la enciclopedia *Britannica*, la cual la define como una experiencia compleja de consciencia, sensación del cuerpo y comportamiento que refleja el significado personal que uno le da a las cosas, eventos o diversas situaciones [42]. Se toma como base la teoría de William James, la cual considera lás emociones como reacciones fisiológicas de un individuo respecto a su entorno [42]. Es importante tener en cuenta que las emociones no son lo mismo a los sentimientos. Las emociones son un tipo de afecto más intensas y complejas que implican manifestaciones expresivas, tienen corta duración y están centradas en un objeto que interrumpe la cognición [43]. En contraparte, los sentimientos son reacciones subjetivas moderadas de placer y displacer. Tienen intensidad media, y tienen una mayor duración [43]. Por ejemplo, la rabia y el miedo son emociones, mientras que la irritabilidad y el rencor son sentimientos [43].

Es evidente que las emociones son vitales para los seres humanos, tienen un rol muy importante en nuestras vidas y están presentes en diversas situaciones [44]. Las emociones se pueden transmitir de manera implicita mediante el lenguaje corporal, la voz, las expresiones faciales y el comportamiento [42]. Recientemente, se ha vuelto tendencia para los científicos de computación estudiar y reconocer automáticamente emociones e incorporar esa tecnología al mundo real [44]. Adicionalmente, se ha establecido que el habla contiene información relevante sobre el sistema nervioso central y por lo tanto, contiene información de la emoción del individuo [44].

A continuación se presentará el modelo discreto para poder seleccionar las emociones para el futuro análisis. Debido a que no existe un conjunto definido de emociones universal que establece una correspondencia entre emociones y voz, se crean varios modelos para representar las emociones [44]. La categorización de emociones es subjetiva, pues los investigadores no coinciden en un conjunto de etiquetas definido [44]. Estos modelos pueden contener valores discretos o continuos para categorizar las emociones. En el presente trabajo se trabajará, como se mencionó previamente, con el modelo discreto.

2.2.1. Modelo Discreto

Los modelos discretos están basados en emociones básicas. Estas son las formas más intensas de las emociones y a partir de éstas se generan las demás mediante

variaciones o combinaciones entre estas [44]. Las emociones de este tipo también se pueden distinguir claramente una de otra por la mayoría de gente y están asociadas con funciones cerebrales que evolucionaron para lidiar con diferentes situaciones [44]. Debido a que son tan claras de distinguir, para su clasificación según las expresiones (en este caso voz), se podría facilitar un poco el proceso respecto al modelo continuo [44]. A continuación, podemos observar los distintos conjuntos que distintos autores definen como emociones básicas.

Autor	Emociones Básicas	Base de Inclusión
Plutchik	Acceptance, anger, anticipation, disgust, joy, fear, sadness, surprise	Relation to adaptive biological process
Ekman, Friesen, Ellsworth	Anger, disgust, fear, joy, sadness, surprise	Universal facial Expressions
Gray	Rage and terror, anxiety, joy	Hardwired
Izard	Anger, contempt, disgust, distress, fear, guilt, interest, joy, shame, surprise	Hardwired
James	Fear, grief, love, rage	Bodily involvement
Mowner	Pain, pleasure	Unlearned emotional states
Oatley and Johnson-Laird	Anger, disgust, anxiety, happiness, sadness	Do not require propositional content
Paksepp	Expectancy, fear, range, panic	Harwired
Tomkins	Anger, interestm contempt, disgust, distress, fear, joy, shame, surprise	Density on neutral firing
Watson	Fear, love, rage	Hardwired
Weiner and Graham	Happiness, sadness	Attribution independent

FIGURA 2.2: Conjuntos de emociones básicas según distintos autores. [44]

En el caso del presente trabajo, se utilizarán las emociones básicas propuestas en la base de datos de EmoFilm. Esto es debido a que se entrena un clasificador de emociones según esta base de datos. Las emociones utilizadas para etiquetar en esta base de datos fueron felicidad, tristeza, miedo, ira, y desprecio. [45]. Estas emociones están basadas en las 6 emociones básicas propuestas por Paul Ekman (observables en la figura 2.2), pero decidieron quedarse con las 5 mencionadas debido a las frecuencias de estas en la base de datos.

2.3. Procesamiento y reconocimiento de la voz

En 1872, Darwin señalaba sobre la importancia de la voz en la comunicación afectiva entre animales y que la eficiencia de los órganos vocales está en el más alto grado como medio de expresión. Asimismo, ésta se encontraba al nivel de las expresiones faciales y de postura como uno de los medios más importantes por los cuales los animales (incluidos los humanos) expresaban emociones [46]. A continuación, se dará una breve explicación sobre los mecanismos que participan en la producción y el reconocimiento de la voz.

2.3.1. La voz humana

La voz humana es el sonido básico que luego es modificado por las características únicas de una lengua [47]. Estas características contemplan, por ejemplo, la fonología y fonética.

En investigaciones sobre el habla y la voz humana es necesario distinguir aspectos de corto y largo plazo. En el primero usualmente se transmite información lingüística como el contenido fonético sean estos aspectos de gramática y contenido del lenguaje; el segundo, más bien transmite información no lingüística como indicadores de género y edad del hablante, y, sobre todo lo que se atañe a esta investigación, el estado emocional [48]. En la expresión de las emociones también son expresadas estructuras lingüísticas y el contenido semántico del habla. Sin embargo, la influencia más directa de las emociones en la voz es la forma en que ellas afecta características suprasegmentales (o prosódicas) del habla (acento, entonación, ritmo, duración y otros) a través de los cambios en los mecanismos fisiológicos de la voz.

Producción de la voz

El sistema vocal humano contempla acciones coordinadas entre tres sistemas fisiológicos: el sistema respiratorio, vocal (de fonación) y de resonancia. A continuación, se dará una breve explicación sobre lo que compone cada sistema. El primero está compuesto por los pulmones, la tráquea, la caja torácica y el diafragma. Estos se encargan del equilibrio de la presión en los pulmones con la ayuda de los músculos inspiratorio y espiratorio. Asimismo, este sistema proporciona la presión necesaria y regulada para impulsar el sistema de fonación[49].

El segundo consiste principalmente en la laringe, cuya estructura incluye las cuerdas vocales y la glotis (una abertura entre las cuerdas vocales en donde el flujo de aire circula desde la tráquea hasta la faringe). Mientras que en una situación tranquila las cuerdas vocales están separadas y el aire fluye a través de la glotis; durante la fonación, las cuerdas se juntan y se tensan. El aire se obstruye y la presión del aire es acumulado bajo las cuerdas vocales y produce una separación de las mismas. Cuando el aire fluye a través de la glotis, la presión del aire entre las cuerdas vocales baja, lo que hace que las cuerdas vocales se cierren y el ciclo se repite. El resultado de este proceso es una fluctuación periódica en la presión del aire que corresponde a un sonido con una frecuencia base llamada frecuencia fundamental (f0) y muchos armónicos, los cuales tienen frecuencias que son múltiplos enteros de f0. Los cambios tanto en la presión de aire bajo la laringe como la tensión y posición de cuerda vocales producirán variaciones en la intensidad de f0 y la distribución armónica de energía del sonido. Por ejemplo, cuando las cuerdas vocales están en alta tensión y la presión subglótica es alta debido a un gran esfuerzo espiratorio, las cuerdas vocales se cerrarán más repentinamente, lo que provocará un aumento no solo en la intensidad general, sino también en f0 y la energía en los armónicos. Tal configuración vocal podría esperarse para ciertas emociones de alta excitación, como la ira [49].

Finalmente, el sistema de resonancia, que comprende el resto del tracto vocal, se extiende desde la glotis, a través de la faringe hasta las cavidades oral y nasal, para luego filtrarse el sonido. La forma y longitud de este sistema depende de la fisonomía de los articuladores (lengua, velo del paladar, dientes y labios), y estos determinan cómo ciertos armónicos se amplifican y otros se atenúan, dando lugar a un sonido de habla irradiado muy complejo de comprender. Solo un número pequeño de patrones armónico atenuados y amplificados llamados formantes corresponden a las diferentes vocales y consonantes vocalizadas.

Percepción de la voz

Hay dos componentes principales en el sistema de percepción auditiva: los órganos auditivos periféricos (oídos) y el sistema nervioso auditivo (cerebro). El oído procesa una señal de presión acústica transformándola primero en un patrón de vibración mecánica en la membrana basilar y luego representando el patrón mediante una serie de pulsos que serán transmitidos por el nervio auditivo. La información de percepción se extrae en varias etapas del sistema nervioso auditivo. En esta sección nos centraremos principalmente en los órganos auditivos[50].

La estructura relevante del oído interno para la percepción del sonido es la cóclea, que se comunica directamente con el nervio auditivo, llevando una representación del sonido al cerebro. La cóclea es un tubo en espiral de unos 3,5 cm de largo, que se enrolla unas 2,6 veces. La espiral está dividida, principalmente por la membrana basilar que se extiende a lo largo, en dos cámaras llenas de líquido. La cóclea se puede considerar aproximadamente como un banco de filtros, cuyas salidas están ordenadas por ubicación, de modo que se logra una transformación de frecuencia a lugar. Los filtros más cercanos a la base coclear responden a las frecuencias más altas y los más cercanos a su ápice responden a las más bajas [50].

2.3.2. Reconocimiento de la voz

El primer paso en cualquiera de los sistemas de reconocimiento automático de la voz es la extracción de características. Es decir, la identificación de los componentes de la señal de audio que son buenos y que permiten identificar el contenido fonológico y fonético y descartar todo lo demás que lleve información innecesaria como el ruido [49].

Primero, para la comprensión del proceso de producción de la voz es que los sonidos generados por el ser humano son filtrados por el sistema fonatorio tal como se vio en el subcapítulo de producción de la voz. Este determina qué sonido sale. Si se pudiera determinar la forma con precisión, esto debería dar una representación precisa del fonema que se produce. La forma del tracto vocal se manifiesta en la envolvente del espectro de potencia de tiempo corto [51]. Si bien existen numerosos métodos de estimación de características aplicados al procesamiento de voz como el análisis cepstral [52] [53], el Linear Predictive Coding (LPC, por sus siglas en inglés) y el MFCC [54], el principal método de extracción de características de señales de voz, se tienen los coeficientes cepstrales de las frecuencias de Mel [55]. Por esta

razón, en este trabajo se hará uso de la técnica de MFCC, que ha sido usada en numerosos intentos para realizar la identificación de voz y habla, como el presentado en el estudio de [56] donde combinan MFCC con una técnica de coincidencia de caracterisiticas conocida como *Dynamic Time Warping* (DTW) para realizar la comparaciones de patrones de voz.

2.3.3. Extracción de características: Coeficientes Cepstrales en las Frecuencias de Mel

La extracción de los coeficientes cepstrales en las frecuencias de Mel (MFCC, por sus siglas en inglés) es uno de los métodos más comunes para la extracción de

características de la voz humana [55]. Fueron presentados por Davis y Mermelstein en la década de 1980, y han sido lo último en tecnología desde entonces. Antes de la introducción de los MFCC, los Coeficientes de Predicción Lineal (LPC, por sus siglas en inglés) y los Coeficientes Cepstrales de Predicción Lineal (LPCC, por sus siglas en inglés) fueron el tipo de método principal para el reconocimiento automático de voz (ASR, por sus siglas en inglés) especialmente con clasificadores HMM. En los sistemas ASR se identifican tres fases importantes para el procesamiento de habla y tienen que ver con el análisis de las características de las señales de habla (o la voz), la clasificación y reconocimiento de patrones y la verificación de pronunciación de las palabras reconocidas por el sistema [57]. En esta investigación se usará el análisis de las características de las señales de voz del paradigma ASR.

Etapas para la extracción de los MFCC

En este acápite, el objetivo principal es describir cómo se transforma la forma de la onda de entrada en una secuencia de vectores característicos acústicos. Cada uno de los vectores representa la información en una pequeña ventana de tiempo de la señal. Los MFCC se basan en la idea del *Cepstrum* (resultado de calcular la inversa de la transformada de Fourier del espectro de la señal estudiada en escala logarítmica). Invirtiendo el orden de las primeras cuatro letras, de esta forma se refuerza la idea de inversa del *Spectrum* [58].

Como se recordará, el primer paso en el procesamiento de la voz es la conversión de las representaciones análogicas en una digital [58]. Este proceso tiene dos pasos: el muestreo y cuantificación. Una señal pasa por un proceso de muestreo al medir su amplitud en un momento determinado [59]. La frecuencia de muestreo es el número de muestras tomadas por segundo. Para realizar una medición precisa de una onda se necesita tener al menos dos muestras en cada ciclo: una parte positiva y otra negativa. Más de dos muestras aumentaría la precisión de la amplitud; sin embargo, menos de dos harían que la frecuencia se pierda completamente. Así pues, la onda de frecuencia máxima que es posible medir es aquella donde la frecuencia es la mitad de la frecuencia de muestreo. Esta máxima frecuencia para una tasa de muestreo dada se llama frecuencia Nyquist. La mayor parte de la voz humana se encuentra debajo de los 10Khz; por lo tanto, se necesitaría una frecuencia de 20Khz para obtener una precisión aceptable [60]. Por lo general, para almacenar los valores de mediciones de amplitud se representan con números tanto de 8 o 16 bits [58]. Este proceso de representación con números enteros se llama cuantificación pues hay granularidad mínima (tamaño cuántico) y todos los valores que están más cerca entre sí que este tamaño cuántico se representan de manera idéntica.

A cada muestra en forma de onda cuantificada digitalizada se referirá como x[n], donde n es un índice a lo largo del tiempo. Ahora que se tiene una representación cuantificada y digitalizada de la forma de la onda, es momento de la extracción de características de MFCC. Son seis pasos que componen este proceso y serán descritas a continuación.

Preénfasis

La primera etapa en la extracción de características MFCC es el proceso de aumentar la cantidad de energía en frecuencias altas. Esto se da ya que en el espectro de segmentos sonoros como los son las vocales, hay más energía en las frecuencias más bajas que en las altas. Esta caída de energía en las frecuencias (llamada inclinación espectral) es causada por la naturaleza del pulso glótico [58]. El aumento de la energía de alta frecuencia causa que la información de estos formantes (una banda de frecuencia particularmente amplificada por el tracto vocal) superiores esté más y mejor disponible para el modelo acústico y mejora la precisión de detección de la voz [61].

Este énfasis se realiza usando un filtro. Se define el filtro como se muestra en la Ecuación 2.1.

$$p(z) = 1 - 0{,}97z^{-}1 \tag{2.1}$$

Segmentación y Ventaneo

El objetivo de extracción de características es proporcionar características espectrales que puedan ayudar a construir clasificadores provenientes de la voz mediante un sensor (micrófono o teléfono). Por tanto, una conversación completa implicaría que el espectro cambie rápidamente. Se entiende, entonces, que el habla es una señal no estacionaria lo que significa que sus propiedades no son constantes a lo largo del tiempo. Más bien, lo que se necesita extraer son características espectrales de una ventana pequeña de la voz. Este proceso se hará usando una ventana que no se haga cero dentro de alguna región y cero en otra zona, ejecutando esta ventana a través de la señal de voz y extrayendo la forma de la onda dentro de cada ventana. Se puede representar este proceso de ventanas por tres parámetros: el ancho de la

ventana (en milisengundos), el desplazamiento entre ventanas sucesivas y la forma de la ventana. El frame se le llama a la voz extraída en cada ventana y al numero de milisengundos en cada marco el tamaño del frame y al número de milisegundos entre los bordes izquierdos de las ventanas sucesivas el cambio del frame [58].

Espectro de Potencia

El paso siguiente es la extracción de información espectral para la señal en la ventana. Se necesita saber cuánta energía contiene la señal en diferentes bandas de frecuencias. La herramienta a usar para la extracción información espectral para una señal de tiempo discreto (es decir, muestreada) es la Transformada Discreta de Fourier (DFT, por sus siglas en inglés) [58].

La entrada a la DFT es una señal en ventana $x[n]...x[m]$, y la salida, para cada una de las N bandas de frecuencia discretas, es un número complejo $X[k]$ que representa la magnitud y la fase de ese componente de frecuencia en la señal original. Si se graficara la magnitud contra la frecuencia, se puede visualizar el espectro. Por ejemplo, la figura se muestra una porción de una señal con ventana Hamming de 25 ms y su espectro calculado por una DFT (con algo de suavizado adicional) [58].

El preénfasis es aplicado a las señales de voz, para incrementar la magnitud de la cantidad de energía presente en las altas frecuencias de la señal y hacer de este parámetro detectable en posteriores fases de procesamiento

El ventaneo de la señal, se realiza al dividir la señal de entrada $s[n]$ en tramas cortas $x_i[n]$.

Banco de filtros de Mel y Transformada discreta de Coseno (DCT)

El siguiente paso para la obtención de los coeficientes de Mel, corresponde al cálculo del periodograma de la señal, con esto se busca encontrar la cantidad de energía presente en cada una de las bandas de frecuencia en las que la señal se encuentra ubicada (Ecuación 2.1).

$$P_i[k] = \frac{1}{N}|S_i[k]|^2 \tag{2.2}$$

Después de hallar la densidad de potencia presente en las bandas de frecuencia de la señal, se filtra el vector de datos resultante, utilizando un banco de filtros de Mel.

Por último, se aplica la operación logarítmica al resultado del proceso de filtrado, para extraer los coeficientes de Mel utilizando la transformada discreta de coseno (DCT), cuyo objetivo es de correlacionar las cantidades estimadas a partir de la aplicación de los filtros de Mel [54].

El resultado final después de aplicar la DCT, condensa los valores de coeficientes por ventana, de acuerdo al número de filtros definidos para la extracción. Estos datos corresponden a los coeficientes estáticos de las frecuencias de Mel, y son utilizados como parámetros de entrenamiento de la red neuronal [62].

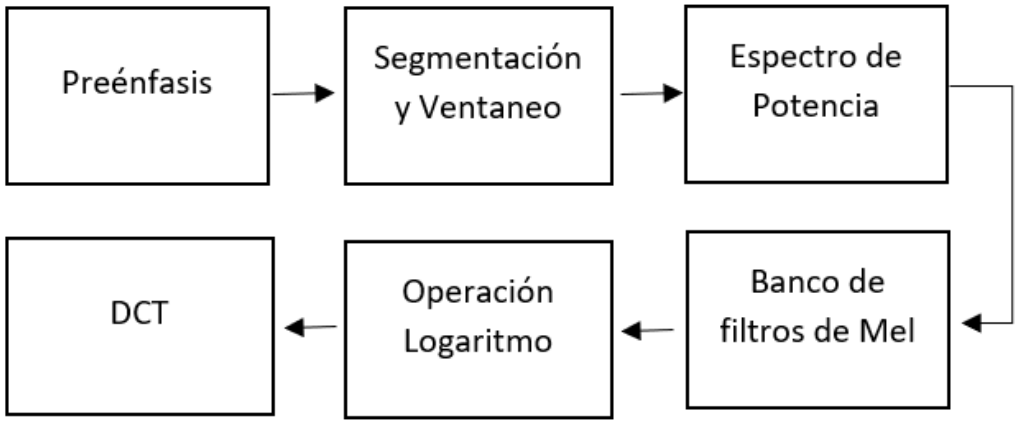

FIGURA 2.3: Procesos involucrados para el cálculo de los MFCC [62]

En la Figura 2.3, se observa el flujo y cada una de las etapas que estan involucrados para el calculo del MFCC que han sido descritas con mayor profundudidad en los parrafos anteriores; donde se han descrito los procesos de Preénfasis, Segmentación y ventaneo, Espectro de Potencia, Banco de filtros de Mel, Operación Logaritmo, y la DCT.

2.4. Redes Neuronales Convolucionales

En primer lugar, se define las Redes Neuronales Artificiales (RNA). Las RNA, forman parte de la Inteligencia Artificial [63]. Según Oliveira *et al* [64] las RNA son redes entrenadas a través de las entradas obtenidas a partir de escenarios externos o internos en el sistema y estas entradas se multiplican por pesos asignados al azar [65]. Asimismo, Callejas *et al*[63], afirman que las RNA son una familia de técnicas de procesamiento de información inspirado por la forma de procesar información del sistema nervioso biológico porque se inspira en el sistema nervioso de un ser vivo, tratando de imitar el comportamiento del cerebro humano [66].

En la Figura 2.4, podemos observar un ejemplo de una RNA, con una sola neurona, que tambien se conoce como Perceptron. Ahi se observa que las entradas a la RNA, son los valores de x, mientras que los w son los pesos que se le asignan a cada una de las entradas x. Luego en el siguiente paso se aplica una sumatoria ponderada de los pesos y las entradas. Finalmente la salida $out(t)$ se obtiene, aplicando una funcion de activación, que determina un umbral que generará una salida binaria de dos categorías.

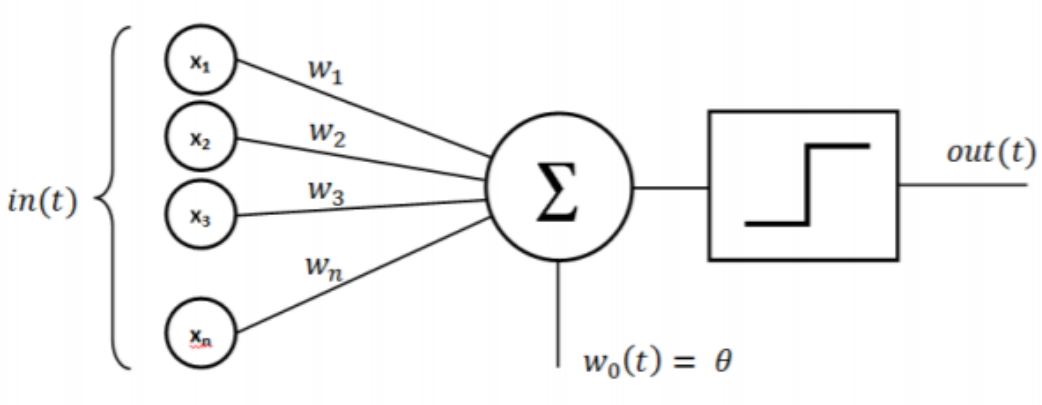

FIGURA 2.4: Ejemplo de un Perceptron simple
[67]

Las redes neuronales convolucionales (CNN) son algoritmos de aprendizaje profundo que pueden entrenar grandes conjuntos de datos con millones de parámetros, en forma de imágenes 2D como entrada y convolucionarlas con filtros para producir los resultados deseados[68].

CNN tiene múltiples capas; entre ellas la capa de convolución, la capa de *pooling*, la capa de no linealidad(*e.g. ReLU*), y la capa *fully-connected*[69]. Las dos primeras, capas de convolución y *pooling*, realizan la extracción de características, mientras que la ultima capa *fully-connected*, mapea las características extraídas en la salida final, como la clasificación[70].

En las imágenes digitales, los valores de los píxeles se almacenan en una cuadrícula bidimensional (2D), es decir, una matriz de números (Figura. 2.5), y se aplica una pequeña cuadrícula de parámetros llamada kernel, un extractor de características optimizable, en cada posición de la imagen, lo que hace que las CNN sean altamente eficientes para el procesamiento de imágenes, ya que una característica puede ocurrir en cualquier parte de la imagen [70].

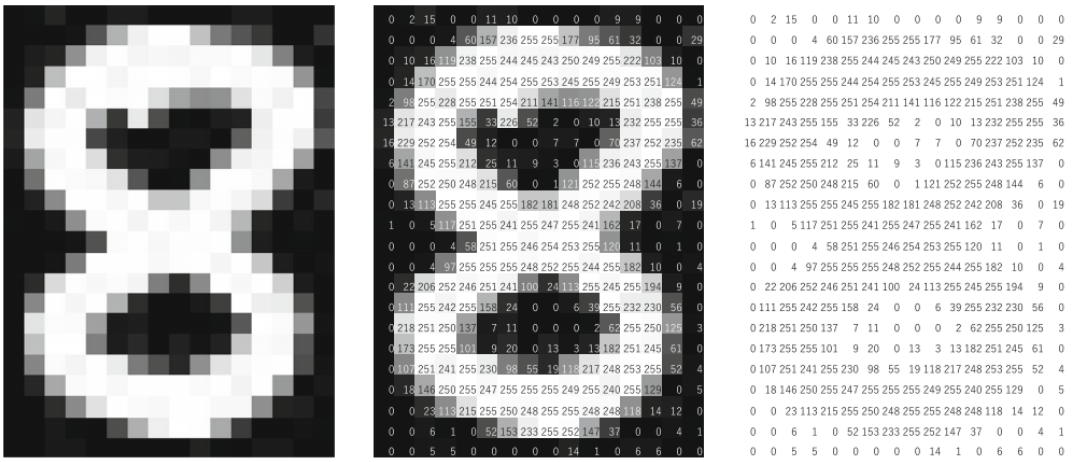

FIGURA 2.5: La computadora ve una imagen como una matriz de números. La matriz de la derecha contiene números entre 0 y 255, cada uno de los cuales corresponde al brillo de píxeles de la imagen de la izquierda. Ambos se superponen en la imagen del medio. La imagen de origen se descargó a través de http://yann.lecun.com/exdb/mnist

2.4.1. Capa de Convolución

La convolución $s(t)$ en términos generales se define como la integral del producto de dos funciones $x(a)$ y $w(a)$ despues de realizar una reflexión horizontal a la segunda función y desplazarla t unidades [71] [Ecuación 2.3]. En terminología de redes convolucionales, el primer argumento en este caso, la función x de la convolución se suele denominar entrada y el segundo argumento en este caso, la función w como *kernel*. La salida es a definido como el *feature map*[71].

$$s(t) = \int x(a)w(t-a), \forall t, a \in R \tag{2.3}$$

En aplicaciones de aprendizaje automático, la entrada suele ser una matriz multidimensional de datos y el *kernel* suele ser una matriz multidimensional de parámetros aprendibles. Nos referiremos a estos arreglos multidimensionales como tensores porque cada elemento de la entrada y el *kernel* deben almacenarse explícitamente por separado, normalmente. Por ello, a menudo se usa convoluciones sobre más de un eje a la vez. En el siguiente ejemplo, se usa una imagen bidimensional I como entrada, y un *kernel* de bidimensional K [71][Ecuación 2.4].

$$s[i,j] = (I \times K)[i,j] \tag{2.4}$$

Donde, I es una matriz bidemensional que representa una imagen, y K es una matriz bidimensional que representa el Kernel de indices i, j.

Las capas convolucionales también pueden reducir significativamente la complejidad del modelo a través de la optimización de su salida. Estos están optimizados a través tres hiperparámetros [72].

- *Depth*: La *depth* (profundidad) del volumen de salida producido por las capas convolucionales puede ser establecer manualmente a través del número de neuronas dentro de la capa a la misma región de la entrada. Esto se puede ver con otras formas de ANN (*Artificial Neural Network*), donde todas las neuronas en la capa oculta están directamente conectadas a cada neurona de antemano. Reducir este hiperparámetro puede minimizar significativamente el número total de neuronas de la red, pero también puede reducir significativamente las capacidades de reconocimiento de patrones del modelo.
- *Stride*: El *stride* (paso) en el que establecemos la *depth* alrededor del espacio dimensional de la entrada para ubicar el campo receptivo. Por ejemplo si íbamos a establecer un paso como 1, entonces tendríamos un campo receptivo muy superpuesto que produce activaciones extremadamente grandes. Alternativamente, estableciendo el stride en un mayor número reducirá la cantidad de superposición y producirá una salida de dimensiones espaciales más bajas.
- *Zero-padding*: El *Zero-padding* (relleno de ceros) es el proceso simple de rellenar el borde de la entrada y es un método eficaz para dar un mayor control en cuanto a la dimensionalidad de la volúmenes de salida.

Mediante el uso de estas técnicas, modificaremos la dimensionalidad espacial de la salida de capas convolucionales. Para calcular esto, se puede hacer uso de la siguiente fórmula[72][Ecuación 2.5].

$$\frac{(V - R) + 2Z}{S + 1} \tag{2.5}$$

Donde V representa el tamaño del volumen de entrada ($height \times width \times depth$), R representa el tamaño del campo receptivo, Z es la cantidad de ajuste de *Zero-padding* y S se refiere al *Stride*. Si el resultado calculado de esta ecuación no es igual a un entero, el *Stride* se ha establecido incorrectamente, ya que las neuronas no podrán encajar perfectamente en la entrada dada.

En la Figura 2.6 se observa un ejemplo de aplicación de la convolución. Se observa un *input*, que viene a ser una matriz de dimension 3×4, y un *kernel* de dimension 2×2. Asimismo, para obtener el *Output*, se realiza la operación de multiplicación de elemento por elemento de las matrices, y finalmente una sumatoria. El *Output* que se genera en la Figura 2.6, es una matriz de dimensión 2×3.

2.4.2. *Pooling*

Una capa *Pooling* proporciona una operación de reducción de resolución típica que reduce la dimensionalidad en el plano de los mapas de características, para introducir una invariancia de traducción a pequeños cambios y distorsiones, que disminuyen el número de parámetros aprendibles subsiguientes [70].

La capa *Pooling* realiza operaciones comunes que vednrian a ser:

- *Max pooling*: La forma más popular de operación de agrupación es la *Max pooling*, que extrae parches de los mapas de características de entrada, genera el

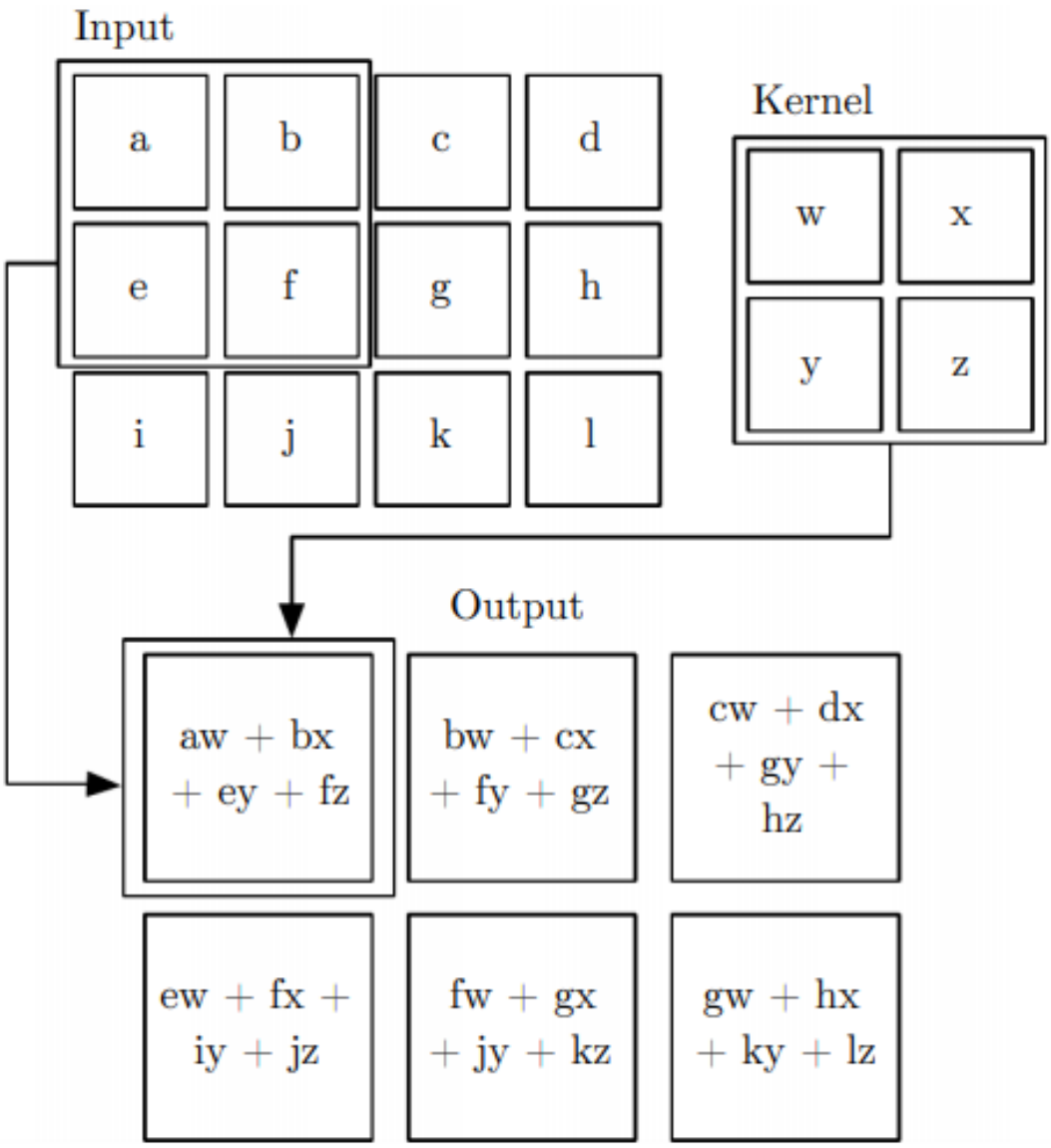

FIGURA 2.6: Un ejemplo de convolución. En este caso restringimos la salida a solo posiciones donde el *kernel* se encuentra completamente dentro de la imagen, llamadas convoluciones válidas en algunos contextos. Se dibujan los recuadros con flechas para indicar cómo se forma el elemento superior izquierdo del tensor de salida aplicando el *kernel* a la correspondiente región superior izquierda del tensor de entrada.

valor máximo en cada parche y descarta todos los demás. En la Figura. 2.7, se observa un ejemplo del uso de la *Max pooling* con un filtro de tamaño 2 × 2 con una stride de 2. Esto reduce la dimensión en el plano de los mapas de características. A diferencia de la altura y el ancho, la dimensión de *depth* de los mapas de características permanece sin cambios [70].

- *Global average pooling*: Otra operación de agrupación que vale la pena mencionar es la *Global average pooling* [73]. Esta realiza un tipo extremo de reducción de resolución, donde un mapa de características con un tamaño de altura × ancho se reduce a una matriz de 11 simplemente tomando el promedio de todos los elementos en cada mapa de características, mientras que la *depth* de

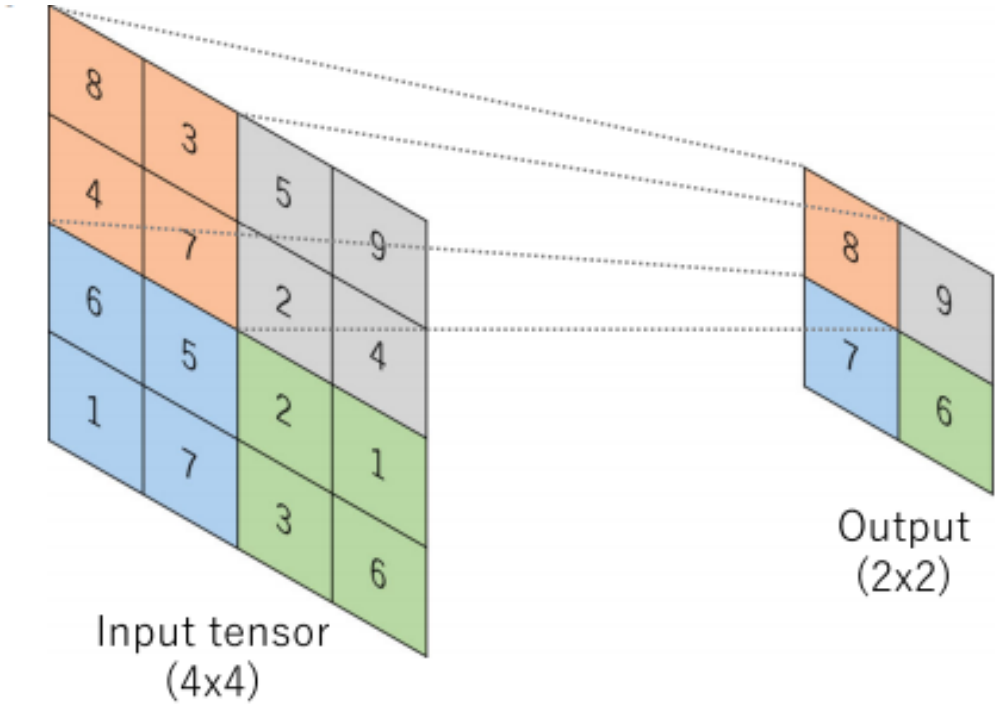

FIGURA 2.7: Un ejemplo de operación de *Max pooling* con un tamaño de filtro de 2 × 2, sin *padding* y un *stride* de 2, que extrae 2 × 2 parches de los tensores de entrada, genera el valor máximo en cada parche y descarta todos los demás valores, resultando en una reducción de la dimensión en el plano de un tensor de entrada por un factor de 2. [70]

los mapas de características es retenido. Las ventajas de aplicar la agrupación de promedios globales son las siguientes: (1) reduce el número de parámetros que se pueden aprender y (2) permite que la CNN acepte entradas de tamaño variable [70].

2.4.3. *Fully Connected*

Los mapas de características de salida de la convolución final o la capa de *pooling* generalmente se aplanan, es decir, se transforman en una matriz unidimensional (1D) de números (o vector), y conectada a una o más capas *fully connected*, también conocidas como capas densas, en la que cada entrada está conectada a cada salida por un peso que se puede aprender. La capa final *fully connected* normalmente tiene el mismo número de nodos de salida que el número de clases. A cada capa *fully connected* le sigue una función no lineal(e.g. ReLU) [70].

- Función de activación de última capa: Una función de activación aplicada a la tarea de clasificación multiclase es una función *softmax* que normaliza los valores reales de salida de la última capa *fully connected* a las probabilidades de la clase objetivo, donde cada valor varía entre 0 y 1 y todos los valores suman 1 [70].

En la Figura. 2.8, se presenta la descripción de una arquitectura de CNN con cada una de las capas: Capa Convolución, Capa *Pooling* y Capa *Fully Connected*. Donde la arquitectura de CNN, tiene como primera capa, una de convolución en conjunto con

una funcion de activación, y luego la capa que se utiliza es una *pooling*. En la ultima capa se utiliza la *fully connected* que determina la salida de la red. Asimismo, la CNN utiliza como algoritmo de aprendizaje de Backpropagation que actualiza los pesos de la red desde la salida de la red hasta la entrada de la red(hacia atras).

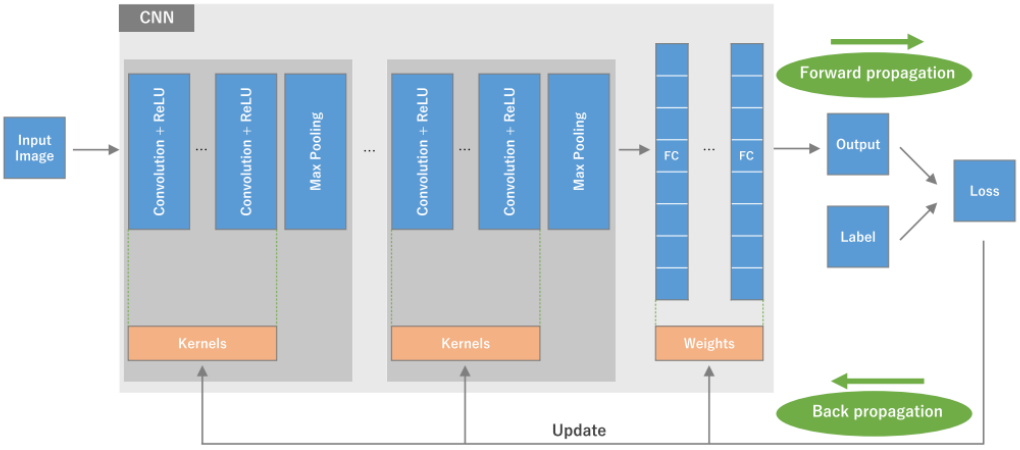

FIGURA 2.8: Una descripción general de la arquitectura de una red neuronal convolucional (CNN) y el proceso de entrenamiento. Una CNN se compone de un apilamiento de varios bloques de construcción: capas de convolución, capas *pooling* (por ejemplo, *max pooling*) y capas *fully connected* (FC).
[70]

2.5. Algoritmo de *Backpropagation*

En esta seccion del trabajo se pasara a hacer una descripcion del algoritmo de *backpropagation*. Para ello, primero se explicara el concepto del descenso de la gradiente, la cual es la parte fundamental para el algoritmo de *backpropagation*.

2.5.1. Descenso de la gradiente

El objetivo del descenso de la gradiente es poder encontrar los 'pesos' que mejor se ajusten a los datos de entrenamiento para minimizar el error de salida [74]. En primer lugar, se define el error estimado del modelo (e.g. Error cuadrático medio). Por tanto, se tiene lo siguiente:

$$E(\vec{w}) = \frac{1}{2} \sum_{d \in D} (t_d - o_d)^2 \tag{2.6}$$

donde o_d representa la salida final de la red neuronal para el dato de entrenamiento d y t_d seria el valor objetivo que se desea estimar en la salida de la red. Como o_d se encuentra en función de los 'pesos'(w), entonces el error también esta en función de los pesos (Ecuación 2.6).

Ahora bien, una vez se tenga la la función de error, la idea del algoritmo descenso del gradiente sería reducir el error de manera iterativa hasta lograr un valor minimo. La dirección hacia donde se incrementa el error se define como la derivada parcial

del error con respecto a cada peso, a esto se le conoce como gradiente (Ecuación 2.7) [74].

$$\bigtriangledown E(\vec{w}) \equiv [\frac{\partial E}{\partial w_o}, \frac{\partial E}{\partial w_1}, ..., \frac{\partial E}{\partial w_n}] \tag{2.7}$$

Como el gradiente $\bigtriangledown E(\vec{w})$ indica la dirección hacia donde se incrementa el error, entonces la regla del entrenamiento del descenso de gradiente se define como (Ecuacion 2.8)

$$\bigtriangledown \vec{w} = -\eta \bigtriangledown E(\vec{w}) \tag{2.8}$$

En la Ecuación 2.8 , η es una constante positiva conocida como el ratio de aprendizaje, la cual es importante para determinar la velocidad de convergencia hacia un valor de error mínimo [74].

Uno de las desventajas que tiene el algoritmo de descenso del gradiente es que en algunos casos la función de error podría tener varios mínimos locales y el algoritmo podría converger en alguno de ellos sin alcanzar el error mínimo global deseado. A pesar de este problema, se han obtenido buenos resultados con este algoritmo [74].

En la Figura 2.9, podemos observar un ejemplo de la función de error en el descenso del gradiente.

Ahora bien, continuando con la descripción del algoritmo *Backpropagation*, este vendría a ser un algoritmo de aprendizaje que involucra un procedimiento iterativo para minimizar una función de error, con ajustes a los pesos que se realizan en una secuencia de pasos. A cada uno de estos pasos, podemos distinguir entre dos etapas distintas [75].

- En la primera etapa, se deben evaluar las derivadas de la función de error con respecto a los pesos. Asi pues la contribución de la técnica de *backpropagation* consiste en proporcionar un método computacionalmente eficiente para evaluar tales derivadas. Cabe resaltar que en esta etapa los errores se propagan hacia atrás a través de la red.
- En la segunda etapa, las derivadas se utilizan para calcular los ajustes de los pesos.

Ahora presentaremos cómo trabaja el algoritmo de *backpropagation* para una red neuronal que tiene una topología de *feed-forward*, funciones de activación no lineales y una función de error representada en la Figura. 2.10. [75].

A continuación, se ilustrarán las fórmulas resultantes:

$$E(w) = \sum_{n=1}^{N} E_n(w) \tag{2.9}$$

En la [Ecuación 2.9] se considera el problema de evaluar cada uno de los pesos w en la funcion de error que seria $E_n(w)$.

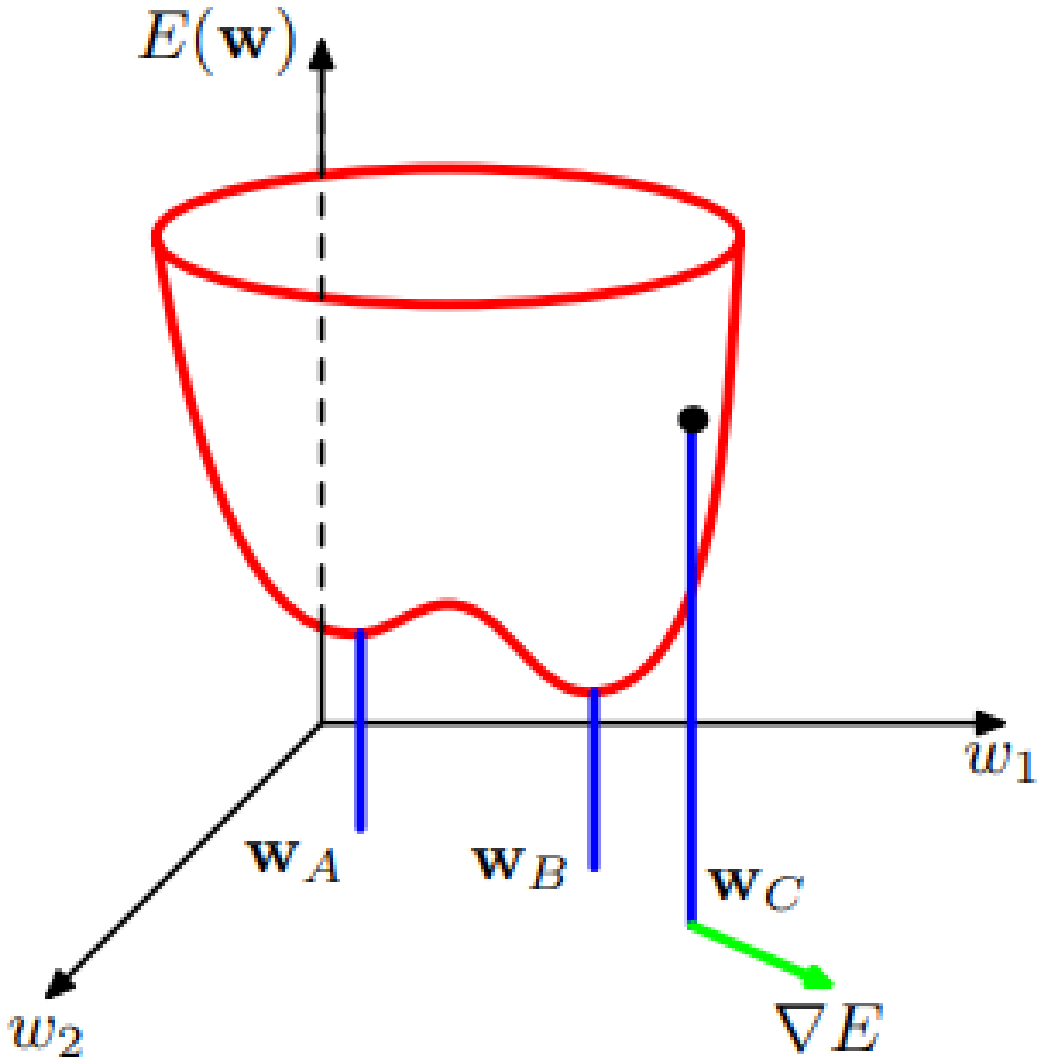

FIGURA 2.9: Vista geométrica de la función de error $E(w)$ como una superficie asentada sobre el espacio de pesos. El punto wA es un mínimo local y wB es el mínimo global.
[75]

Ahora, se considera primero un modelo lineal simple en el que las salidas y_k son combinaciones lineales de las variables de entrada x_i [Ecuación 2.10].

$$y_k = \sum_i w_{ki} x_i \tag{2.10}$$

Por otro lado, el gradiente de una función de error con respecto a un peso w_{ji} viene dado por:

$$\frac{\partial E_n}{\partial w_{ji}} = (y_{nj} - t_{nj}) x_{ni} \tag{2.11}$$

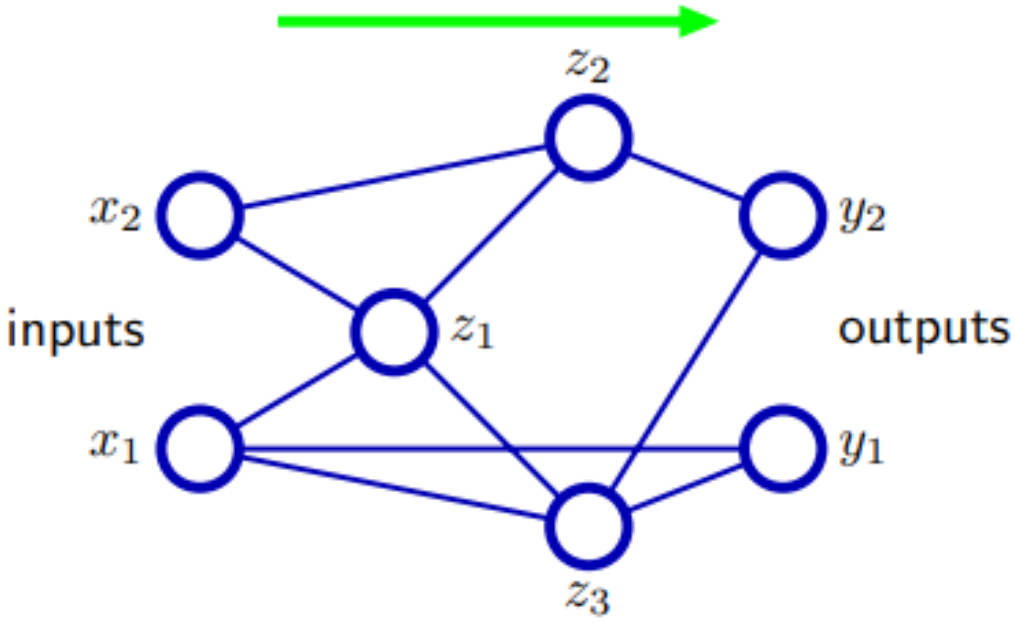

FIGURA 2.10: Ejemplo de una red neuronal que tiene un topología *feed-forward*. Tenga en cuenta que cada unidad oculta y de salida tiene un parámetro de sesgo asociado.
[75]

En una red de *feed-forward*, cada unidad calcula una suma ponderada de sus entradas :

$$y_j = \sum_i w_{ji} x_i \tag{2.12}$$

donde, y_j es la salida de la red, w_j son los pesos asignados a cada una de las entradas x_i.

A continuación se considera la evaluación de la derivada de E_n con respecto a un peso w_{ji}. Las salidas de las diversas unidades dependerán del patrón de entrada particular n. Sin embargo, para mantener la notación ordenada, omitiremos el subíndice n de las variables de la red. Primero notamos que E_n depende del peso w_{ji} solamente a través de la entrada sumada a_j a la unidad j. Por tanto, podemos aplicar la regla de la cadena para derivadas parciales para dar:

$$\frac{\partial E_n}{\partial w_{ji}} = \frac{\partial E_n}{\partial a_j} \frac{\partial a_j}{\partial w_{ji}} \tag{2.13}$$

Introducimos ahora una notación δ:

$$\delta_j = \frac{\partial E_n}{\partial a_j} \tag{2.14}$$

donde los δ a menudo se denominan errores por razones que veremos en breve

$$\frac{\partial a_j}{\partial w_{ji}} = z_i \tag{2.15}$$

Reemplazando la ecuación 2,14 y la ecuación 2,15 en la ecuación 2,13 , se obtiene:

$$\frac{\partial E_n}{\partial w_{ji}} = \delta_j z_i \tag{2.16}$$

Al evaluar los δ para unidades ocultas, de nuevo hacemos uso de la regla de la cadena para derivadas parciales, donde la suma corre sobre todas las unidades k a las que la unidad j envía conexiones (Ecuación 2.17).

$$\delta_j = \frac{\partial E_n}{\partial a_j} = \sum_k \frac{\partial E_n}{\partial a_k} \frac{\partial a_k}{\partial a_j} \tag{2.17}$$

Para obtener la derivada del error total E se puede obtener repitiendo los pasos anteriores para cada patrón en el conjunto de entrenamiento y luego sumando todos los patrones:

$$\frac{\partial E}{\partial w_{ji}} = \sum_n \frac{\partial E_n}{\partial w_{ji}} \tag{2.18}$$

En la derivación anterior se ha asumido implícitamente que cada unidad oculta o de salida en la red tiene la misma función de activación h (·) [75].

2.6. Métricas de evaluación

En esta sección, se analizará las métricas de evaluación que se usan para evaluar el rendimiento de los algoritmos y modelos de *Machine* y *Deep Learning*.

Para resolver un problema de clasificación es necesario seleccionar el algoritmo apropiado mediante una o varias métricas de evaluación del rendimiento en varios algoritmos candidatos [76].

Las métricas más frecuentes utilizadas indistintamente por investigadores en estudios de clasificación predictiva son: *accuracy*, *recall* (o sensibilidad), Matriz de confusión, *F1- score*, curva ROC y AUC [77]. Por tanto, en este trabajo se trabajará principalmente con la Matriz de confusión, *Accuracy*, *Recall* y Especificidad.

2.6.1. Matriz de Confusión

La Matriz de confusión trata de una tabla de frecuencias donde las filas pertenecen a la clase predicha y las columnas a la clase real, representando el número de predicciones de cada clase mutuamente excluyentes (Cuadro 2.1)

	Clase 1: Positivo	Clase 2: Negativo	TOTAL
Clase 1: Positivo	f_{11} = Verdadero Positivo	f_{11} = Falso Negativo	$F1T$
Clase 2: Negativo	f_{01} = Falso Positivo	f_{00} = Verdadero Negativo	$F0T$
	$f1T$	$f0T$	N

CUADRO 2.1: Matriz de confusión para variables dicotómicas. [78]

En el cuadro 2.1 ejemplifica el caso de una clasificación binaria, donde los verdaderos positivo y negativo ($f11$ y $f00$) corresponden al número de instancias correctamente clasificadas en cada clase, mientras que los totales F y f representa las frecuencias marginales correspondientes a: $F1T$ - $F0T$ es igual a la cantidad total de elementos en cada clase, $f1T$ – $f0T$ es el total de instancias clasificadas por el algoritmo como positivo y negativo en el caso binario y finalmente N representa el tamaño muestral utilizado para la clasificación [78].

La *accuracy* es la métrica más utilizada por investigadores en el caso dicotómico y multiclase, debido a su facilidad de cálculo y comprensión para evaluar la efectividad general del algoritmo [79] [80] [81]. Esta medida muestra el porcentaje de casos totales que el modelo ha acertado.

Recall o sensibilidad es una medida que permite conocer la proporción de casos positivos que fueron correctamente clasificados. En un modelo perfecto el *recall* es igual a 1 para cada clase. Desde el punto de vista analítico un investigador busca aumentar el *recall* sin afectar el valor de la *accuracy* [82].

Asimismo, la metrica de Especificidad, viene a ser la proporción de casos negativos que fueron correctamente clasificados [82].

En el cuadro 2.2 se describe las fórmulas utilizadas para el cálculo de cada métrica según los diferentes enfoques presentados [83] [84] [85] todas ellas en un intervalo [0,1] y basados en la nomenclatura de la 2.1.

2.7. Estado del Arte

En esta sección, se realizará la revisión de la literatura del presente trabajo, donde se dividirá en tres partes: en primer lugar, se enfocará en los trabajos desarrollados de acuerdo a un modelado dinámico de características, en segundo lugar, se enfocará en los trabajos desarrollados de acuerdo a un modelado estático de características, y finalmente, se presentará la comparación del presente trabajo y los trabajos revisados. En el Cuadro 2.3 (al final del capítulo), se muestran los tres principales trabajos presentados en el estado del arte.

Métrica	Fórmula	Descripción
Accuracy	$\frac{f_{11}+f_{00}}{N}$	Proporción de clasificaciones predichas de manera correcta sobre el total de instancias.
Recall(Sensibilidad)	$\frac{f_{11}}{f_{11}+f_{01}}$	Proporción de casos positivos bien clasificados.
Especificidad	$\frac{f_{00}}{f_{00}+f_{10}}$	Proporción de casos negativos bien clasificados.

CUADRO 2.2: Métricas para evaluación de rendimiento de clasificadores de Machine y Deep Learning.
[83] [84] [85]

2.7.1. De acuerdo a un modelado dinámico de características

En esta sección se abordarán trabajos de investigación con un enfoque de modelado dinámico de características. En este tipo de modelos el factor tiempo es explícitamente considerado [86]. En este se emplean además características como tono, energía, MFCC y sus derivativas, estos empleados en conjunto con modelos de clasificación como el *Hidden Markov Model* (HMM). [87].

Asimismo, el enfoque de modelado dinámico, realiza un análisis que se hace a nivel de ventanas del mismo tamaño, por lo que para cada elocución se tienen vectores de características de diferentes tamaños dependiendo de su duración [88].

En primer lugar, se presenta el trabajo de Bozkurt *et al* [89] en donde se propone un sistema de reconocimiento de emociones promovido por señales de voz. Para este trabajo se hizo uso del corpus *FAU Aibo Emotion Corpus* [90], que se distribuye a través del *INTERSPEECH 2009 Emotion Challenge*, este corpus contiene grabaciones de voz espontáneas y emocionales.

Asimismo, en el estudio mencionado se plantean diversos desafíos de clasificación y de características, para ello se plantearon problemas de clasificación de dos y cinco clases. En este trabajo se propone la clasificacion emocional del hablante, para ello se presenta características relacionadas con la prosodia, espectrales y aquellas que estan basadas en el *Hidden Markov Model* (HMM).

En la siguiente Figura 2.11, podemos observar un ejemplo de una estructura de HMM con múltiples ramas.

Las caracteristicas de la prosodia consiste en valores medios de tono, primera derivada de tono e intensidad. Las características espectrales consisten en coeficientes cepstrales en escala Mel (MFCC), características de frecuencia espectral de línea (LSF) y sus derivadas. A su vez, para el entrenamiento no supervisado de estructuras se emplea HMM para definir características temporales relacionadas con la prosodia para el reconocimiento de emociones.

Para la evaluación y reconocimiento de emociones en este trabajo se utilizarón

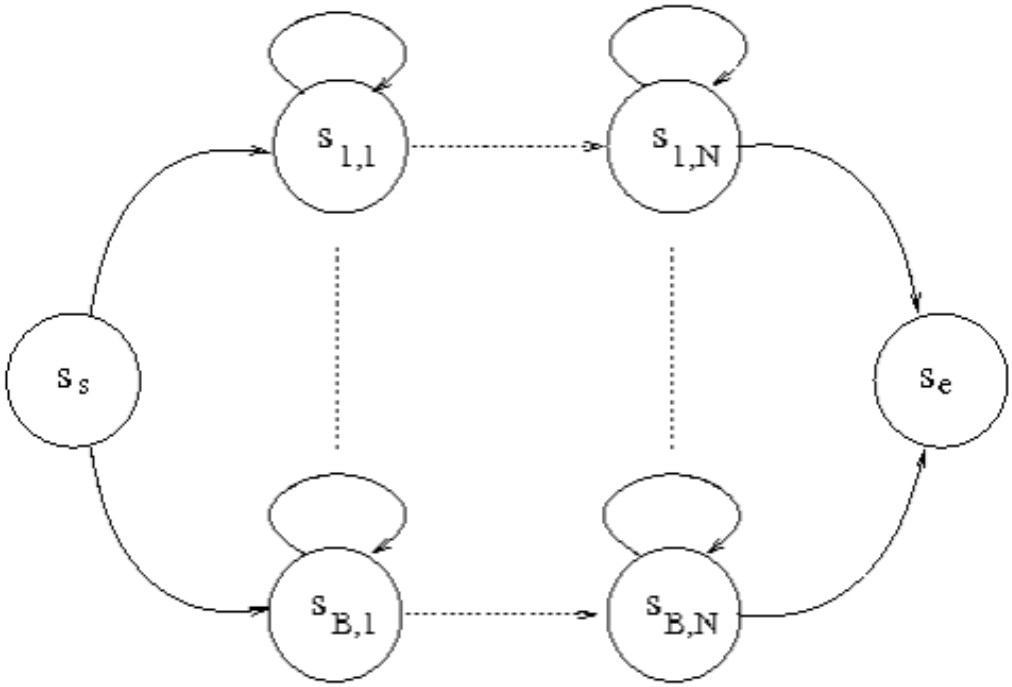

FIGURA 2.11: Estructura HMM de múltiples ramas.
[89]

clasificadores basados en el modelo de mezcla gaussiana (GMM). Donde los resultados fueron presentados para la evalucación de los clasificadores y evaluación de las características. Asimismo, se utilizaron problemas de clasificacion para dos y cinco clases.

Con respecto a evaluación de los clasificadores, los resultados en el estudio de Bozkur *et al* [89], arrojan que, para problemas de clasificación de dos clases y cinco clases, los clasificadores GMM tienen la mayor precision con 65.25 % para la clasificación de dos clases, y 46.66 % para cinco clases. Ahora bien con respecto a la evaluación de características , para los resultados se experimentó con la estructura de HMM con diferentes parámetros variando el número de estados por rama de 3 a 10 y el número de componentes gaussianos por estado hasta 12. Asimismo, se encontró que el conjunto de características con 3 estados por rama y 12 componentes gaussianos por estado produce los mejores resultados con una precisión de 57.43 % para la clsificación de dos clases y 27.48 % para cinco clases.

Como segundo trabajo para el modelado dinámico tenemos el trabajo de Dumouchel *et al.* [91] acerca de cepstrales y características de largo plazo para el reconocimiento de emociones. En el trabajo se realizan 2 distintas tareas. En primer lugar, se hace una clasificación para 2 emociones. Estas son inactivo (ocioso) o negativo. Luego se propone la clasificación para 5 emociones las cuales son Enojo, Empatía, Neutral, Positivo y Descanso. Se propuso fusionar varios sistemas de clasificación que operan a corto plazo, largo plazo y según las características de la voz.

Luego, para cada tarea se utilizaron distintos tipos de extracción de características de la voz y distintos sistemas de clasificación. Dentro de estos sistemas de extracción, tanto para la primera como la segunda tarea se realizó la extracción con MFCC (Coeficiente Cepstral de Frecuencia de Mel) y la extracción del tono de voz, del logaritmo de energía y de los dos primeros formantes utilizando el paquete Praat. Se debe tener en cuenta que el primer método de extracción se utilizó para características en el corto plazo, mientras que el segundo se utilizó para las características de largo plazo. En el caso de MFCC, se utilizaron 12 coeficientes cesptrales de frecuencia de Mel. Cada frame tuvo un tiempo definido de 10ms. Asimismo, los silencios se removieron para la extracción. En el caso de la extracción con el paquete *Praat*, se usaron intervalos de 10ms y que solo se utilizaron secciones en las que aparecía una voz existente.

Adicionalmente, para cada tarea se utilizaron distintas técnicas de clasificación. En la primera se utilizaron Support Vector Machines (SVM), Modelos Mixtos Gaussianos (GMM) y una combinación de estos. Esto significa que existen dos tareas, cada una con modelos de extracción de características de corto y largo plazo, y finalmente con técnicas de clasificación tales como SVM, GMM y sus distintas fusiones. Para la base de datos se utilizó un corpus de emociones proveniente de *FAU AIBO*. Se utilizó 9959 datos de entrenamiento y 8257 datos de prueba. Se utilizaron grabaciones de voces de niños. Se realizó 9 veces el proceso de *cross validation* para los datos y como criterio de separación se tomaron en cuenta 3 identidades diferentes de los niños.

Para medir el rendimiento de los distintos clasificadores, se midieron tanto el *recall* con peso o sin peso, pero se le dio más importancia al *recall* sin peso. En el caso del análisis para 2 emociones, el mejor rendimiento lo tuvo el sistema con extracción de características de largo plazo para el caso del *recall* con peso con el sistema GMM (70.84 %). Sin embargo, no supera al que se obtuvo en el estado del arte del trabajo respectivo (72.6 %). De igual manera, para el recall sin peso, se obtuvo mejores resultados con la fusión de los tres sistemas basados en entrenamiento con regresión logística.

En el caso del análisis para las 5 distintas emociones, se obtuvo el mejor resultado para un *recall* sin peso con una extracción de MFCC para el corto plazo y un sistema de clasificación GMM. Se llegó a un 39.4 % el cual obtuvo un resultado mayor a su respectivo estado del arte en 3.5 %.

2.7.2. De acuerdo a un modelado estático de características

En esta sección se abordarán trabajos de investigación con enfoque de modelado estático de caracteristicas. Un modelo estático es aquel donde el factor tiempo no se tiene en cuenta explícitamente [86]. El problema de clasificación en este tipo de modelamiento se aborda con métodos estáticos como *Support Vector Machines* (SVM) o Redes Neuronales. Asimismo, las caracteristicas son obtenidas de la extracción de *Low Level Descriptors* (LLD), como por ejemplo la entonación, energía, o coeficientes espectrales [92] [93] [94].

En primer lugar, se revisa el trabajo de Tripathi *et.al.* [95] quienes proponen un método de reconocimiento de emociones a través de la voz basado en características del habla y transcripciones de la misma (texto) con experimentaciones usando

arquitecturas de *Deep Neural Network* (DNN, por sus siglas en inglés) en los cuales se toman combinaciones de características del habla y texto como inputs. El primer modelo entrenado es un CNN basado en texto en el que las transcripciones del habla se toman como entradas. El segundo, es un modelo CNN basado en características del habla en donde se usan como entradas tanto espectogramas como MFCCs como información extraída de las señales en bruto del audio para obtener las características necesarias para que junto con las CNN propuestas se realice la detección de emociones. Por último, se realiza un modelo combinado de CNN basado en características de texto y habla donde se combinan los tipos de entradas: Espectrograma y MFCC (1), espectrograma y texto (2) y MFCC y texto (3). En el primero, el canal del espectrograma consta de 4 capas 2D-CNN paralelas con kernels de diferentes tamaños. Al igual que el canal del espectrograma, el canal MFCC también consta de 4 capas 2D-CNN paralelas. Las salidas de ambos canales se alimentan a una capa FC cada una. Las salidas de ambas capas son FC, después de la normalización, se concatenan y alimentan a la segunda capa FC. El último paso es alimentar las salidas de la última capa FC a una capa softmax. Las arquitecturas de los modelos 2 y 3 son similares a las del modelo 1. Ambos modelos tienen un canal de texto que toma incrustaciones de palabras como entrada pero alterna entre espectrograma y MFCC como característica de voz.

El modelo combinado de texto y espectrograma proporciona una precisión de clasificación del 69,5 % y una precisión general del 75,1 %, mientras que el modelo combinado de texto y MFCC también ofrece una precisión de clase del 69,5 %,pero una precisión general del 76,1 %, un 5,6 % y casi 7 % de mejora sobre los índices de referencia actuales, respectivamente. Los modelos propuestos se pueden utilizar para aplicaciones relacionadas con las emociones, como chatbots conversacionales, robots sociales, etc. Identificar las emociones y los sentimientos ocultos en el habla puede desempeñar un papel en la mejor conversación.

Como segundo trabajo, los autores Basu *et al*.[96], realizaron un reconocimiento de las emociones utilizando las señales de voz. Donde utilizó el método de MFCC para la extraccion de características de las señales de voz, a su vez, en este estudio se usó un enfoque basado en los modelos de *Convolutional Neural Network* (CNN) y *Long Short Term Memory* (LSTM) para los problemas de clasificación.

Para este estudio, se utilizó la base de datos de Berlín sobre el habla emocional (EmoDB), que almacena 535 expresiones pronunciadas por 10 actores diferentes. Asimismo, este corpus constaba de siete clases emocionales: feliz, enojado, ansioso, temeroso, aburrido, disgustado y neutral [96].

Para el estudio de Basu *et al* [96], como un paso previo a utilizar el método de MFCC, se uso un procesamiento en el conjunto de datos. Primero se calculo los valores de amplitud de cada archivos de la voz con una frecuencia de muestreo de 16000 muestras por segundo. Este procesamiento previo, generó archivos de voz con el mismo tamaño en un volumen alto en una escala lineal.

En este estudio se dividió el conjunto en 80 % para entrenamiento y el 20 % para prueba. Como siguiente paso, se usó la técnica de MFCC con velocidad y aceleración para cada archivo del conjunto de datos de entrenamiento y de prueba. Luego estas características extraídas sirvieron como entradas para la CNN. Asimismo, Basu [96],

utilizó el modelo CNN, con tres capas de convolución, con 32, 16, 8 filtros respectivamente. Asimismo, en el estudio se usó como función .Adadeltaçomo optimizador, y como función de activación la ReLu. Por otro lado en el estudio se empleó una red LSTM con dos capas ocultas con 50 neuronas en la primera capa y 20 neuronas en la segunda capa. Como función de activación para la salida final se uso la función *Softmax*.

Para este estudio, se obtuvo una precisión de 90 % para la etapa de entrenamiento, mientras que para la etapa de prueba se obtuvo un 80 %. Asimismo, el trabajo concluyó que el uso del modelo CNN-LSTM para el reconocimiento de emociones resulta efectivo y un paso importante hacia la creación de un diseño genérico[96].

Para el tercer trabajo se tomó en cuenta un reconocimiento de emociones a partir de la voz utilizando CNN escrito por Abdul *et al*. [97] Este trabajo es muy importante, pues abarca las técnicas que se van a utilizar en el presente trabajo. Para el trabajo se utilizó la base de datos "*Surrey Audio Visual Expressed Emotion*" (SAVEE), la cual fue grabada por 4 hombres entre 27 y 31 años, que tienen como lengua materna el inglés, y que son graduados e investigadores en la universidad Surrey. Se tienen como emociones ira, disgusto, miedo, felicidad, tristeza y sorpresa. Asimismo, se le agregó neutralidad. Se consideraron 300 datos de entrenamiento, 100 de validación y 80 de prueba.

Para la clasificación de emociones se utilizaron tres distintos clasificadores a parte del CNN para una posible comparación. *Multivariate Linear Regression Classification* (MLR), *Support Vector Machines* (SVM) y *Recurrent Neural Networks* (RNN). Para el CNN se tomaron en cuenta 7 capas de una dimensión convolucional, cada una seguida de una capa de normalización y una capa de *max pooling* con un pool size de 2, excepto por la última capa.

Las grabaciones duraron 8 segundos y las que duraron menos se les completó con vacíos. El número de filtros en las capas fueron 32, 64, 128, 256, 512, 1024 y 1024 respectivamente. Los tamaños del *Kernel* fueron 21, 19, 17, 15, 13, 11 y 9 respectivamente. La última capa de una dimensión fue seguida por una capa *max pooling* y luego por otra capa densa con 128 nodos. La función de activación para todas las capas fue "Relu". La última capa del modelo tiene 7 nodos, debido a que existen 7 salidas, y tienen la función de activación "*Softmax*".

En la etapa del entrenamiento del CNN, se utilizó el optimizador *Adam* con un learning rate de 0.001, beta1 de 0.9, beta2 de 0.999. La categoría de *Loss* fue "*Categorical Crossentropy*". Se entrenó con más de 400 épocas. Finalmente, en los resultados se demuestra que la mejor *accuracy* conseguida fue la perteneciente a la técnica de CNN. Para ira, disgusto, miedo, felicidad, neutral, tristeza y sorpresa se llegaron a *accuracies* de 73.5 %, 80.88 %, 80.12 %, 86.38 %, 89.75 %, 87.25 % y 83.61 % respectivamente.

2.7.3. Técnicas de extracción de características de voz

Para poder tener una mejor idea qué método de extracción de características de voz utilizar, se revisarán trabajos pasados que muestren finalmente cual podría ser la tecnica más adecuada para el caso propuesto. En primer lugar, se tiene el trabajo de Namrata (2013) [98] el cual realiza una red neuronal con 3 parámetros. Estos

parámetros son las 3 distintas técnicas de extracción de características los cuales son *Linear Predictive Codes* (LPC), *Perceptual Linear Prediction* (PLP), y Coeficientes Cepstrales de Frecuencia de Mel (MFCC). El trabajo comenta incluso antes de mostrar los resultados, que la técnica más frecuente y dominante para encontrar características del espectro, es MFCC. Es muy utilizada porque utiliza el dominio de frecuencia utilizando la escala de Mel, la cual está basada en la escala humana. Cuando se utiliza MFCC con el dominio de frecuencia, se tiene una mayor precisión comparada a características con dominio de tiempo. Adicionalmente, estos coeficientes son robustos y fuertes según la variación del audífono y condiciones de grabación de voz [98].

La técnica MFCC extrae parámetros de la voz similares a los utilizados en conversaciones y le quita énfasis a otra información que contiene la grabación. En el trabajo de Namrata, se toman 2 segundos (aproximadamente 128 *frames*) que contienen 128 *samples* cada uno con un tamaño de ventana de 16ms. Se utilizan los primeros 40 *frames* que dan una buena estimación de discurso. Un total de 42 parámetros de MFCC incluyen 12 originales, 12 delta y 12 delta delta, y 3 energía logs.

Finalmente, el trabajo muestra con los resultados que LPC no es una técnica tan aceptable para este caso debido a la naturaleza lineal de computación. Debido a que la voz humana no es lineal por naturaleza, esta técnica no es una buena opción para la estimación del discurso. Fueron de mucha más utilidad lás tecnicas de PLP y MFCC que incluyen el concepto de un sistema de audición humana, y por lo tanto tienen mejores resultados comparados con LPC.

En segundo lugar, el trabajo de Singh *et al.* [99] tiene como principal objetivo realizar una comparativa entre técnicas de extracción de características cepstrales y no cepstrales. La primera, la técnica de extracción a corto plazo llamado el coeficiente cepstral de frecuencia de Mel (MFCC) y, la segunda, las de largo plazo llamadas prosódicas, respectivamente. Las características de MFCC se extraen de los fonemas del hablante en las oraciones del habla preseleccionadas.

En el estudio se encontró que existen factores como la variación de la voz, diferentes micrófonos y auriculares, que afectan la eficiencia del reconocimiento de la voz. Asimismo, se obtiene que hay otros parámetros que afectan a este reconocimiento: Calidad de los datos, lenguaje, ruido en la grabación, longitud de la muestra de voz, variación de la voz.

En conclusión, se demuestra que MFCC es mejor que las técnicas prosódicas en el reconocimiento del hablante y sus características del tracto vocal del mismo. Además, que los sistemas cepstrales a corto plazo generalmente funcionan bien dado que reflejan mejor la fisiología del hablante y no solo en el contenido fonético.

Como último trabajo se tiene al de Martinez [100], donde se utiliza el MFCC para extraer las funciones de voz y la técnica de cuantificación vectorial para identificar al hablante. Esta técnica se utiliza habitualmente en la compresión de datos y permite modelar la función de probabilidad por la distribución de diferentes vectores. Los resultados de este trabajo fueron un 100 % de precisión con una base de datos de 10 hablantes.

Por último, en el trabajo de Martinez [100], se concluye que utilizando la técnica

de MFCC se representa de mejor manera la voz humana, y que la técnica LPCC se utiliza en la comunicación digital, por lo que el propósito principal de esta técnica no es representar la voz, sino para comprimir y transmitir la información que contiene la voz. Por ello, debido a que el MFCC usa la escala mel, la aproximación al comportamiento de la voz humana es buena, y además representa de mejor manera la voz[100].

En conclusión, como podemos observar en los trabajos anteriores, la técnica MFCC tiene muchas venajas para el presente caso y por consiguiente será esta tomada en cuenta para la extracción de datos.

2.7.4. Comparación del presente trabajo y los trabajos revisados

Tomando en cuenta los trabajos presentados en la secciones anteriores, se especifican las técnicas para la clasificación de emociones mediante la voz. De los 3 trabajos presentados en la sección de modelado dinámico de características tienen en común la técnica de MFCC para la extracción de características de las señales de voz. Esto se ve reflejado en el presente trabajo, pues se utilizará la misma técnica. Además, en dichos trabajos se utilizaron técnicas tales como SVM, GMM, y HMM, las cuales tienen de variable de entrada el audio tomando en cuenta el tiempo. Se llegó a la conclusión que la técnica con mejor precisión, tanto para la clasificación de 2 y 5 emociones, es GMM.

Sobre los 3 trabajos analizados en la sección de modelado estático de características, se tiene que estos, al igual que en los trabajos que usan el modelado dinámico obtienen las características de la voz con la técnica de MFCC. Luego, desarrollan la técnica de CNN para la clasificación y reconocimiento de emociones usando las señales de voz. Asimismo, para el presente trabajo nos basaremos en aplicar la técnica de CNN para clasificación de las emociones, debido a los resultados prometedores que pudo conseguir en trabajos pasados.

En base a lo anterior, el presente trabajo se diferencia del resto de trabajos en lo que respecta al idioma de la base de datos a utilizar, pues en el presente trabajo se trabajará una base de datos en el idioma español, a diferencia del resto de trabajos que lo hicieron con el idioma inglés. Adicionalmente,la base de datos en el presente trabajo será extraída de un *call center* lo cual podría tener como posible aplicación ayudar en un futuro análisis en ventas según las emociones de los clientes.

Autor	Título	Objetivo	Técnicas Utilizadas	Resultados
Basu *et al.*(2017)	Emotion recognition from speech using convolutional neural network with recurrent neural network architecture	Propone un modelo de reconocimiento de las emociones utilizando las señales de voz. Donde utilizo el método de MFCC para la extraccion de características de las señales de voz. Se usó un enfoque basado en los modelos de CNN (Convolution Neural Network) y LSTM (Long Short Term Memory) para abordar los problemas de clasificación.	▪ MFCC ▪ CNN ▪ LSTM	El Modelo obtuvo una precisión de 90 % para la etapa de entrenamiento, y 80 % para la etapa de prueba sobre el *dataset* EmoDB.
Tripathi *et al.* (2019)	Deep learning based emotion recognition system using speech features and transcriptions	Propone y enfrenta 4 modelos: Modelo CNN basado en texto, Modelo CNN basado en funciones de voz, Modelo CNN con MFCC de entrada, Modelos CNN combinados basados en funciones de voz y texto para reconocimiento de emociones a través de la voz	▪ MFCC ▪ CNN ▪ LSTM	El resultado sobresaliente fue el uso de la técnica combinada con inputs de voz y text con lo que se consiguió un accuracy de 76.1 %
Abdul *et al.* (2019)	Convolutional Neural Network (CNN) Based Speech-Emotion Recognition	Los autores proponen el Reconocimiento de emociones por medio de la voz bajo un método CNN sin procesamiento previo analizando con cuatro técnicas	▪ MLR ▪ SVM ▪ RNN ▪ CNN	La mejor clasificación se obtuvo con el modelo CNN obteniéndose un accuracy de 83.6 %

CUADRO 2.3: Principales trabajos descritos en el estado del arte

Capítulo 3

Metodología

En el presente capítulo se detalla la metodología propuesta para poder cumplir tanto el objetivo general como los específicos del trabajo. Esta metodología se divide en cuatro etapas: la construcción de un modelo de *machine learning* capaz de clasificar emociones a partir de la voz con la base de datos EmoFilm, la construcción de un conjunto de datos de voz de llamadas telefónicas de ventas, la aplicación del modelo entrenado en la base de datos de ventas construido para refinar el modelo base entrenado en la etapa 1 y finalmente presentar un análisis de resultados de clasificación de emociones en base a datos de llamadas telefónicas. En cada una de las etapas se describe de manera específica los procesos a seguir, la importancia de estos y las técnicas necesarias para poder llevarlos a cabo. En la Figura 3.1 se muestra el diagrama de la metodología en donde se presentan estas etapas y cómo estas se relacionan para poder cumplir los objetivos.

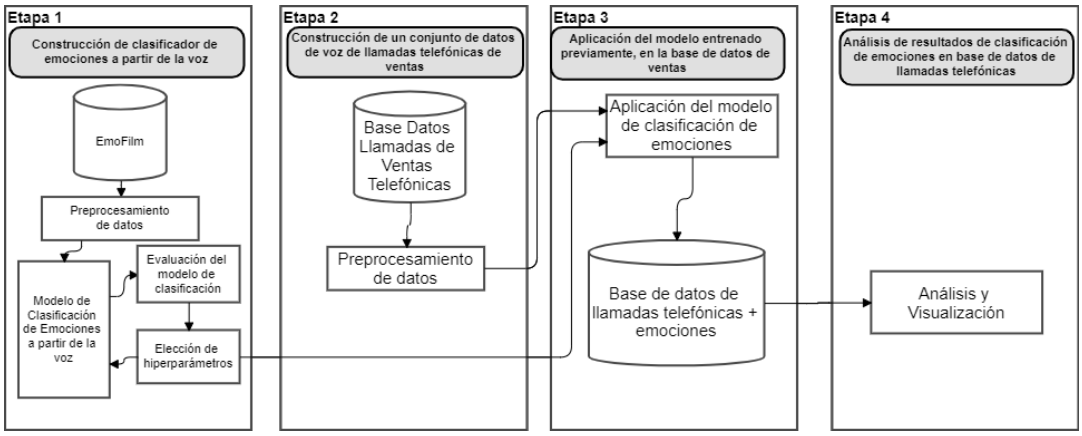

FIGURA 3.1: Diagrama de la metodología

Como se mencionó anteriormente, el propósito de la metodología es formular y aplicar métodos que permitan alcanzar los objetivos general y específicos. Por lo tanto, en el presente trabajo se propone entrenar un modelo clasificador de emociones a partir de la voz, utilizando la base de datos en línea llamada EmoFilm. Una vez entrenado el modelo CNN con la base de EmoFilm, se aplica este modelo a una base de datos de llamadas telefónicas previamente trabajada y construida. Esto tiene el objetivo de obtener una base de datos de llamadas que no solamente indique si se concretó la venta o no se concretó, sino que también incluye las emociones encontradas durante las llamadas con el modelo entrenado en la etapa 1. Finalmente, se relaciona la concreción de la venta con las emociones clasificadas con la ayuda del modelo clasificador de emociones.

3.1. Etapa 1: Construcción de un clasificador de emociones a partir de la voz

Para la construcción de un clasificador de emociones a partir de la voz, se entrena un modelo a partir de la base de datos (BD) EmoFilm, la cual está en el idioma español y contiene audios recolectados de distintas series de televisión, así como etiquetas de emociones encontradas en estos audios. Estas emociones incluyen ira, desprecio, felicidad, miedo, tristeza. En primer lugar, se realizará un preprocesamiento de la BD EmoFilm para poder extraer características de las etiquetas que se encuentran en los nombres de los archivos de audio. Esta operación incluye la selección de audios pues estas contienen audios en tres idiomas y etiquetas de sus respectivas emociones que se encuentran en el mismo nombre de archivo de cada audio. Una vez filtrados los audios correspondientes se realizará la extracción de características de estos con el algoritmo MFCC que se explicó en el marco teórico del presente trabajo. Esta técnica extrae los coeficientes cepstrales en la escala de Mel y su respectivo espectograma considerando la percepción auditiva del ser humano en las frecuencias y separa de manera eficiente las características de voz útiles para el proceso de clasificación. Como se observa en la figura 3.2, se presenta el ejemplo de un espectograma de Mel que es espectograma donde sus frecuencias están en las escala de Mel.

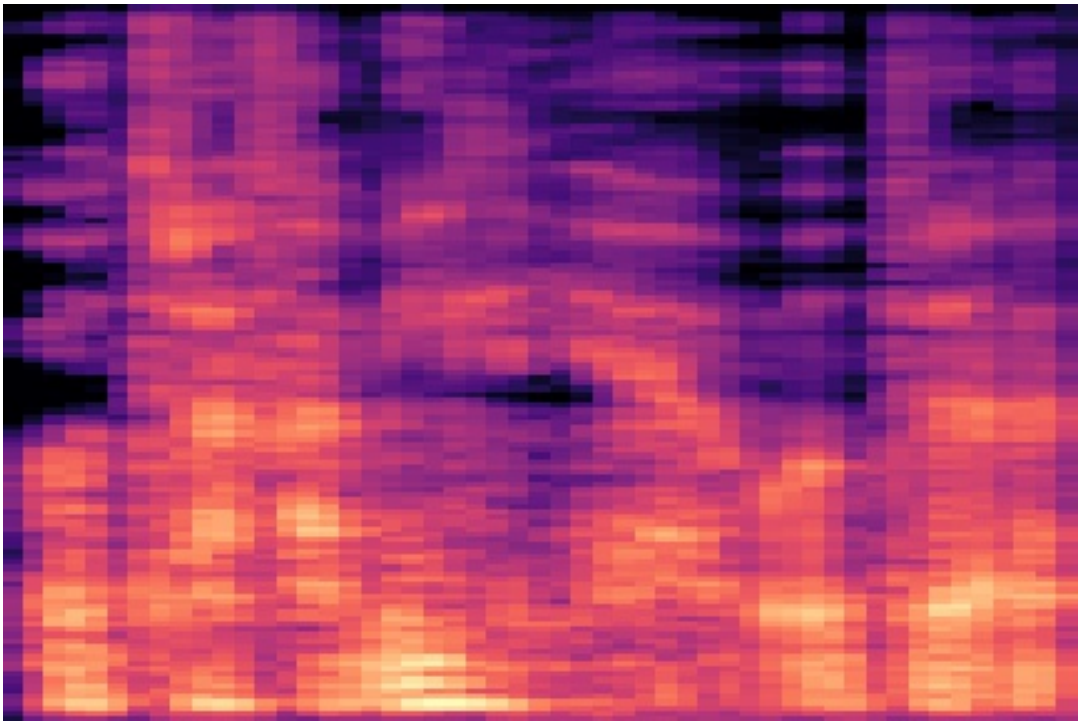

FIGURA 3.2: Imagen de un espectograma de Mel

Una vez utilizada la técnica MFCC, se obtendrán los espectogramas y coeficientes, los cuales servirán de *input* al clasificador de imágenes de redes neuronales convolucionales (CNN). Este modelo se utilizará para cada instancia (cada audio) con el objetivo de clasificar la emoción. Esta técnica ha demostrado ser efectiva en literatura revisada previamente en el Estado del Arte. Finalmente, se evaluará el clasificador de emociones (el CNN), y para ello se utilizará la matriz de confusión y la métrica *accuracy*. Es importante resaltar que si el accuracy del modelo no es mayor a 90 %, se

realizará un proceso iterativo para modificar los parámetros e hiperparámetros para mejorar el resultado. Los parámetros de la librería *Librosa* para extraer los MFCCs y los espectogramas de Mel se componen de los siguientes: *sampling rate* de 20500 y 40 coeficientes cepstrales en la escala de Mel.

En la figura 3.3 se presenta la arquitectura CNN que se aplicará para realizar el modelo clasificador de emociones a través de la voz. En primer lugar, se hará un *resize* de los datos ya que como se ha obtenido coeficientes cepstrales por cada instancia de audio, este solo es de 1 dimensión. Para ello se hizo una función en *Python* para su respectiva conversión en dos dimensiones. En segundo lugar, la arquitectura inicia con una capa convolutiva con con filtro 64 y un kernel de tamaño 5. Luego, se tiene una capa *Dropout* de 0.1 la cual permitirá desactivar neuronas aleatoriamente y así evitar el *overfitting*. En tercer lugar, se tiene su respectiva capa de activación *ReLU* que permitirá a la red estandarizar los valores de los coeficientes cepstrales obtenidos en el preprocesamiento. En cuarto lugar, se tiene una capa *MaxPooling* que permitirá resumir nuestros datos para redimensionarlos y hacer la red más eficiente. En cuarto lugar, se tiene una capa convolutiva para poder realizar nuevamente un muestreo de nuestros datos procesados en los datos resultantes y luego se procede a realizar la misma operación. Luego, para obtener un red capaz de realizar las clasificaciones se procede a convertir los datos con la función Flatten. Asimismo, se crea una capa *Dense* para finalmente realizar las clasificaciones con una capa de activación Softmax. Se usará un optimizador RMSprop con una tasa de aprendizaje de 0.0001 y un rho de 0.9 que es el factor de descuento para los subsiguientes gradientes en la operación de optimización.

3.2. Etapa 2: Construcción de un conjunto de datos de voz de llamadas telefónicas de ventas

En la etapa 2 se procederá a construir la base de datos de llamadas de ventas telefónicas a partir de un conjunto de datos obtenidos de una empresa de *Call Center*. Para ello se utilizará una BD de llamadas de ventas telefónicas (500 donde se concretaron en venta y 500 que no se llegaron a concretar). El preprocesamiento consistirá en segmentar cada llamada haciendo uso de los softwares: *VoxtSort Diaritazion* y *Ocenaudio*. Esta segmentación consiste en identificar las principales frases de las dos personas presentes en la llamada (operador y cliente). Se obtendrán 2 frases de cada uno al principio de la llamada, 2 frases al medio de la llamada y 2 frases al final de la llamada. Esto tiene el objetivo de poder distinguir las diversas emociones que pueda presentar un interlocutor o un cliente a lo largo de la llamada. Finalmente, se tendrán las distintas frases que tienen de llave la llamada, la etiqueta de 'Venta' o 'No Venta', y entre las caracteristicas se tendrá, la fecha, hora y género del hablante.

Una vez segmentadas todas las llamadas en frases, se tomará un subconjunto de 30 llamadas (con 12 frases cada llamada) para poder etiquetar manualmente la emoción según el tono de voz. Para este etiquetado, cada uno de los tres integrantes escucha las frases y manualmente pone la emoción que según la persona debería haber en la llamada (las mismas emociones de las base de EmoFilm). Luego, se comparan los resultados de los 3 integrantes y por mayoría de votación queda la emoción corerespondiente. Si no se llega a una mayoría, la frase se escucha nuevamente en

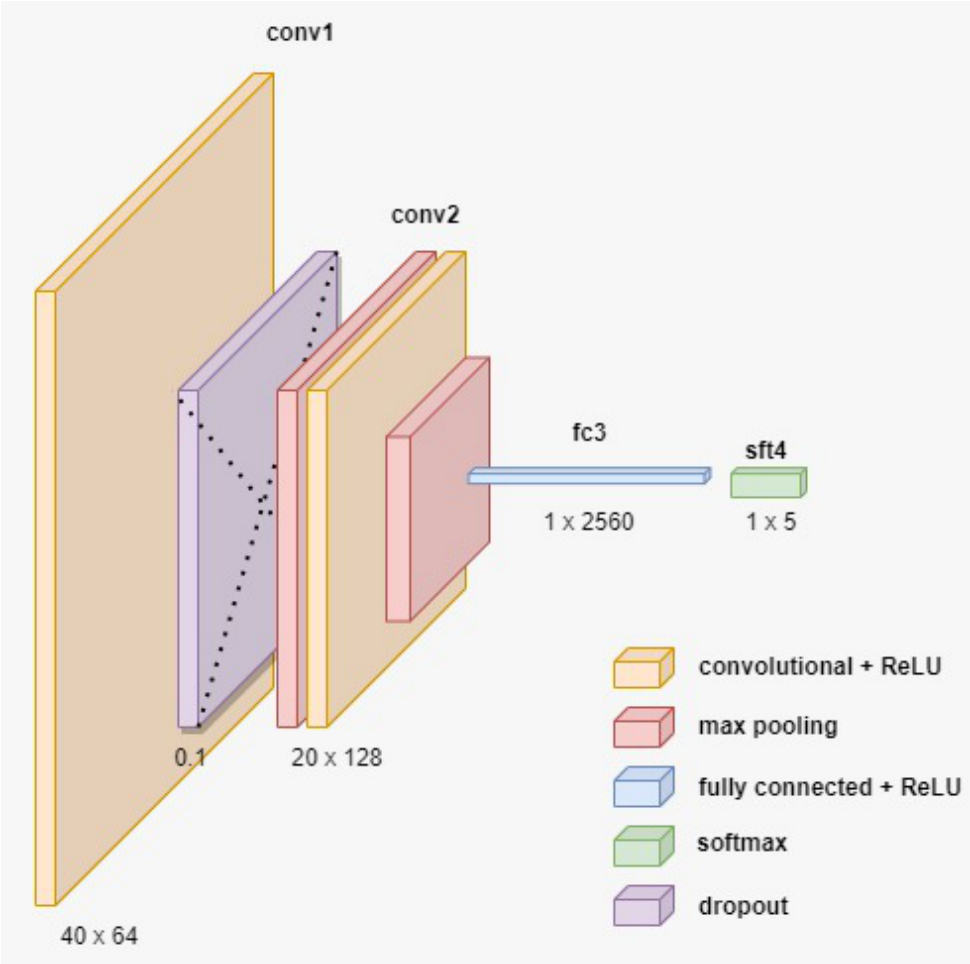

FIGURA 3.3: Modelo CNN clasificador de voz

conjunto y se llega a una conclusión. Esta metodología se encuentra en literatura pasada [101]. Una vez etiquetado el subconjunto, se le aplica el modelo CNN entrenado previamente de la base de datos EmoFilm. Se mide el accuracy del modelo aplicado al subconjunto y si este es mayor que 80 %, entonces se procederá a aplicar el modelo CNN a todas las llamadas telefónicas (1000 audios con 12 frases cada audio).

3.3. Etapa 3: Aplicación del modelo entrenado previamente en la base de datos de ventas construido

A partir de la construcción de un clasificador (CNN) de emociones en la etapa 1, y teniendo la base de datos de llamadas telefónicas preprocesadas en la etapa 2, se aplica el modelo de clasificación de emociones en la base total de datos completa de llamadas de ventas telefónicas preprocesadas.

Finalmente, se tendrá en la base de llamadas de ventas telefónicas, todas las frases con las emociones clasificadas, y si se concretó la venta o no, y las otras características de la fecha, hora y género del hablante.

3.4. Etapa 4: Análisis de resultados de la clasificación de emociones en llamadas telefónicas

En esta última etapa se analizan los resultados de ventas de las llamadas unidas a las emociones asociadas obtenidas con el modelo clasificador de emociones. Se realizará un cuadro de mando con el *software Power BI* para analizar el impacto de las emociones y los porcentajes de concreción de venta. Con ello encontrar patrones que relacionen las variables independientes (emociones) y la dependiente (concreción de venta en la llamada) en el conjunto de datos construido. Asimismo, la evaluación de la clasificación de emociones se realizará principalmente a partir de la métrica accuracy. Esta métrica es la más utilizada por investigadores en el caso dicotómico y multiclase, debido a su facilidad de cálculo y comprensión para evaluar la efectividad general del algoritmo [78] [79] [80]. Este indicador varía entre cero y uno, siendo lo más óptimo tener un modelo con un accuracy cercano al uno, para así tener un modelo bastante robusto.

Bibliografía

[1] H. R. Max Roser y E. Ortiz-Ospina, «Internet,» dirección: https://ourworldindata.org/internet.

[2] J. Johnson, *Global digital population as of January 2021*, 2021. dirección: https://www.statista.com/statistics/617136/digital-population-worldwide/#:~:text=As\%20of\%20January\%202021\%20there,the\%20internet\%20via\%20mobile\%20devices..

[3] The World Bank, *Mobile Cellular Suscriptions*, data retrieved from World Development Indicators, https://data.worldbank.org/indicator/IT.CEL.SETS?end=2019&start=2000, 2018.

[4] Y. Bakos, «The emerging landscape for retail e-commerce,» *Journal of economic perspectives*, vol. 15, n.º 1, págs. 69-80, 2001.

[5] A. Meier y H. Stormer, *eBusiness & eCommerce: managing the digital value chain*. Springer Science & Business Media, 2009.

[6] T. Sabanoglu, *Global retail e-commerce market size 2014-2023*, 2021. dirección: https://www.statista.com/statistics/379046/worldwide-retail-e-commerce-sales/.

[7] A. Bhatti, H. Akram, H. M. Basit, A. U. Khan, S. M. Raza y M. B. Naqvi, «E-commerce trends during COVID-19 Pandemic,» *International Journal of Future Generation Communication and Networking*, vol. 13, n.º 2, págs. 1449-1452, 2020.

[8] World Health Organization, *Coronavirus (CoV) GLOBAL*, 2021. dirección: https://www.who.int/es/health-topics/coronavirus#tab=tab_1.

[9] E. O.-O. Max Roser Hannah Ritchie y J. Hasell, «Coronavirus (COVID-19 Deaths,» *Our World in Data*, 2020, https://ourworldindata.org/covid-deaths?country=OWID$_W$RL.

[10] R. Y. Kim, «The impact of COVID-19 on consumers: Preparing for digital sales,» *IEEE Engineering Management Review*, vol. 48, n.º 3, págs. 212-218, 2020.

[11] World Trade Organization, *WTO report looks at role of e-commerce during the COVID-19 pandemic*. dirección: https://www.wto.org/english/news_e/news20_e/rese_04may20_e.htm.

[12] O. Andrienko, *Ecommerce amp; Consumer Trends During Coronavirus*, 2020. dirección: https://www.semrush.com/blog/ecommerce-covid-19/.

[13] A. Vargas, *Comercio electrónico ha crecido más de 300 % en Latinoamérica en la pandemia*, 2020. dirección: https://www.larepublica.co/globoeconomia/e-commerce-ha-crecido-mas-de-300-en-latinoamerica-en-medio-de-la-pandemia-3000424.

[14] F. Bravo, *Comercio electrónico en Perú: La Guía más completa del mercado*, 2021. dirección: https://www.ecommercenews.pe/ecommerce-insights/2021/crecimiento-del-comercio-electronico-en-peru.html#:~:text=El\%20Per\%C3\%BA\%20tiene\%2011.8\%20millones,a\%20trav\%C3\%A9s\%20de\%20dispositivos\%20m\%C3\%B3viles.&text=El\%20Per\%C3\%BA\%20representa\%20el\%205.3\%25\%20de\%20volumen\%20ecommerce\%20en\%20la\%20regi\%C3\%B3n.

[15] L. G. Schiffman y L Lazar Kanuk, «Comportamiento del Consumidor (DÉCIMA EDICIÓN ed.),» *México DF: Pearson Educación. Recuperado el*, vol. 8, 2010.

[16] A. Deaton y J. Muellbauer, *Economics and consumer behavior*. Cambridge university press, 1980.

[17] S. Kohli, B. Timelin, V. Fabius y S. M. Veranen, *How COVID-19 is changing consumer behavior–now and forever*, 2020.

[18] H. M. Paksoy, Y. Durmaz, F. Çopuroğlu, B. D. Özbezek y col., «The Impact of Anxiety Caused by COVID-19 on Consumer Behaviour,» *Transnational Marketing Journal*, vol. 8, n.º 2, págs. 243-270, 2020.

[19] J. Hernández Rodríguez, «Impacto de la COVID-19 sobre la salud mental de las personas,» *Medicentro Electrónica*, vol. 24, n.º 3, págs. 578-594, 2020.

[20] S. K. Brooks, R. K. Webster, L. E. Smith, L. Woodland, S. Wessely, N. Greenberg y G. J. Rubin, «The psychological impact of quarantine and how to reduce it: rapid review of the evidence,» *The lancet*, vol. 395, n.º 10227, págs. 912-920, 2020.

[21] A. Martínez-Taboas, «Pandemias, COVID-19 y Salud Mental:¿ Qué Sabemos Actualmente?» *Revista Caribeña de Psicología*, págs. 143-152, 2020.

[22] C. I. Orellana y L. M. Orellana, «Síntomas emocionales y compras por pánico durante la pandemia de COVID-19: Un análisis de trayectoria,» *Psicogente*, vol. 24, n.º 45, págs. 1-19, 2021.

[23] W. Cao, Z. Fang, G. Hou, M. Han, X. Xu, J. Dong y J. Zheng, «The psychological impact of the COVID-19 epidemic on college students in China,» *Psychiatry research*, vol. 287, pág. 112 934, 2020.

[24] B. Sandín, R. M. Valiente, J. García-Escalera y P. Chorot, «Impacto psicológico de la pandemia de COVID-19: Efectos negativos y positivos en población española asociados al periodo de confinamiento nacional.,» *Revista de Psicopatología y Psicología Clínica*, vol. 25, n.º 1, 2020.

[25] Essalud, *EsSalud advierte incremento en la hospitalización de niños y adolescentes por diagnóstico de depresión en el último año.* dirección: http://noticias.essalud.gob.pe/?inno-noticia=essalud-advierte-incremento-en-la-hospitalizacion-de-ninos-y-adolescentes-por-diagnostico-de-depresion-en-el-ultimo-ano.

[26] K Tarunika, R. Pradeeba y P Aruna, «Applying machine learning techniques for speech emotion recognition,» en *2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, IEEE, 2018, págs. 1-5.

[27] E. Cambria, D. Hazarika, S. Poria, A. Hussain y R. Subramanyam, «Benchmarking multimodal sentiment analysis,» en *International Conference on Computational Linguistics and Intelligent Text Processing*, Springer, 2017, págs. 166-179.

[28] J. C. Mundt, P. J. Snyder, M. S. Cannizzaro, K. Chappie y D. S. Geralts, «Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology,» *Journal of neurolinguistics*, vol. 20, n.º 1, págs. 50-64, 2007.

[29] F. Alías, J. C. Socoró y X. Sevillano, «A review of physical and perceptual feature extraction techniques for speech, music and environmental sounds,» *Applied Sciences*, vol. 6, n.º 5, pág. 143, 2016.

[30] B. Zhang, C. Quan y F. Ren, «Study on CNN in the recognition of emotion in audio and images,» en *2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS)*, IEEE, 2016, págs. 1-5.

[31] B. Sandín y P. Chorot, «Cuestionario de Sucesos Vitales (CSV): Estructura factorial, propiedades psicométricas y datos normativos,» *Revista de Psicopatología y Psicología Clínica*, vol. 22, n.º 2, págs. 95-115, 2017.

[32] E. Liberos, *Vender a través de la red; el comercio electrónico*. ESIC Editorial, 2016.

[33] J. Bullemore-Campbell y E. Cristóbal-Fransi, «La dirección comercial en época de pandemia: el impacto del covid-19 en la gestión de ventas,» *Información tecnológica*, vol. 32, n.º 1, págs. 199-208, 2021.

[34] R. L. Ackoff, «Science in the systems age: Beyond IE, OR, and MS,» *Operations Research*, vol. 21, n.º 3, págs. 661-671, 1973.

[35] D. C. Miller y N. J. Salkind, *Handbook of research design and social measurement*. Sage, 2002.

[36] K. K. Bajaj, D. Nag y K. K. Bajaj, *E-commerce*. Tata McGraw-Hill Education, 2005.

[37] K. Laudon y C. G. Traver, *E-commerce*. Pearson Educación, 2009.

[38] M Niranjanamurthy, N Kavyashree, S Jagannath y D. Chahar, «Analysis of e-commerce and m-commerce: advantages, limitations and security issues,» *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 2, n.º 6, págs. 2360-2370, 2013.

[39] S.-Y. Choi, D. O. Stahl y A. B. Whinston, *The economics of electronic commerce*. Macmillan Technical Publishing Indianapolis, IN, 1997.

[40] R. Nemat, «Taking a look at different types of e-commerce,» *World Applied Programming*, vol. 1, n.º 2, págs. 100-104, 2011.

[41] L. P. Bolaños Florido, «El estudio socio-histórico de las emociones y los sentimientos en las Ciencias Sociales del siglo XX,» *Revista de Estudios Sociales*, n.º 55, págs. 178-191, 2016.

[42] R. C. Solomon, *Emotion*, 2019. dirección: https://www.britannica.com/science/emotion/The-structure-of-emotions.

[43] J. A. Marina y M. L. Penas, *Diccionario de los sentimientos*. Anagrama, 2000.

[44] H. P. Espinosa, «Reconocimiento de emociones a partir de voz basado en un modelo emocional continuo,» 2010.

[45] E. Parada-Cabaleiro, G. Costantini, A. Batliner, A. Baird y B. Schuller, «EmoFilm - A multilingual emotional speech corpus,» 2018. DOI: 10.5281/zenodo.1326428. dirección: https://doi.org/10.5281/zenodo.1326428.

[46] C. Darwin, «The expression of the emotions in man and animals (1872),» *The Portable Darwin*, págs. 364-393, 1993.

[47] J. Cocteau, G. Kaiser y A. Strindberg, *La voz humana*. Peña, Del Giudice, 1953.

[48] T. Johnstone, «The effect of emotion on voice production and speech acoustics,» 2017.

[49] X. Huang, A. Acero y H.-W. Hon, «Spoken Language Processing. Guide to Algorithms and System Development,» *PH*, 2001.

[50] R. Munkong y B.-H. Juang, «Auditory perception and cognition,» *IEEE signal processing magazine*, vol. 25, n.º 3, págs. 98-117, 2008.

[51] M. A. S. Murillo, J. I. de la Rosa Vargas y A. M. Báez, «COMPARACIÓN DE TÉCNICAS DE PARAMETRIZACIÓN ESPECTRAL PARA RECONOCIMIENTO DE VOZ EN IDIOMA ESPAÑOL,»

[52] R. Maia, M. Akamine y M. J. Gales, «Complex cepstrum for statistical parametric speech synthesis,» *Speech Communication*, vol. 55, n.º 5, págs. 606-618, 2013.

[53] H. Hong, Z. Zhao, X. Wang y Z. Tao, «Detection of dynamic structures of speech fundamental frequency in tonal languages,» *IEEE Signal Processing Letters*, vol. 17, n.º 10, págs. 843-846, 2010.

[54] X.-C. Yuan, C.-M. Pun y C. P. Chen, «Robust Mel-Frequency Cepstral coefficients feature detection and dual-tree complex wavelet transform for digital audio watermarking,» *Information Sciences*, vol. 298, págs. 159-179, 2015.

[55] L. D. Vignolo, H. L. Rufiner, D. H. Milone y J. C. Goddard, «Evolutionary cepstral coefficients,» *Applied Soft Computing*, vol. 11, n.º 4, págs. 3419-3428, 2011.

[56] L. Muda, M. Begam e I. Elamvazuthi, «Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques,» *arXiv preprint arXiv:1003.4083*, 2010.

[57] H. Veisi y H. Sameti, «Speech enhancement using hidden Markov models in Mel-frequency domain,» *Speech Communication*, vol. 55, n.º 2, págs. 205-220, 2013.

[58] R. Klabunde, «Daniel jurafsky/james h. martin, speech and language processing,» *Zeitschrift für Sprachwissenschaft*, vol. 21, n.º 1, págs. 134-135, 2002.

[59] M. S. Santina y A. R. Stubberud, «Basics of sampling and quantization,» en *Handbook of networked and embedded control systems*, Springer, 2005, págs. 45-69.

[60] L. Zhang, S. Tan y J. Yang, «Hearing your voice is not enough: An articulatory gesture based liveness detection for voice authentication,» en *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 2017, págs. 57-71.

[61] Y. Zhang, C.-j. HE, L. Yuan, C. Kai y W.-c. XING, «Improved perceptually non-uniform spectral compression for robust speech recognition,» *The Journal of China Universities of Posts and Telecommunications*, vol. 20, n.º 4, págs. 122-126, 2013.

[62] O. L. Ramos, D. A. Rojas, L. A. Góngora y col., «Reconocimiento de patrones de habla usando MFCC y RNA,» *Visión electrónica*, vol. 10, n.º 1, págs. 5-11, 2016.

[63] C. Ivan, P. Juan, R. Juan, H. Ferney y D. Fabio, «Implementación de una red neuronal artificial tipo SOM en una FPGA para la resolución de trayectorias tipo laberinto,» en *2013 II International Congress of Engineering Mechatronics and Automation (CIIMA)*, IEEE, 2013, págs. 1-6.

[64] W. R. Asanza y B. M. Olivo, «Redes neuronales artificiales aplicadas al reconocimiento de patrones,» *Editorial UTMACH*, 2018.

[65] K. Shaban, A. El-Hag y A. Matveev, «A cascade of artificial neural networks to predict transformers oil parameters,» *IEEE Transactions on Dielectrics and Electrical Insulation*, vol. 16, n.º 2, págs. 516-523, 2009.

[66] M. F. Tahir, M. A. Saqib y col., «Optimal scheduling of electrical power in energy-deficient scenarios using artificial neural network and Bootstrap aggregating,» *International Journal of Electrical Power & Energy Systems*, vol. 83, págs. 49-57, 2016.

[67] C. Antona Cortés, «Herramientas modernas en redes neuronales: la librería keras,» B.S. thesis, 2017.

[68] R. Chauhan, K. K. Ghanshala y R. Joshi, «Convolutional neural network (CNN) for image detection and recognition,» en *2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC)*, IEEE, 2018, págs. 278-282.

[69] S. Albawi, T. A. Mohammed y S. Al-Zawi, «Understanding of a convolutional neural network,» en *2017 International Conference on Engineering and Technology (ICET)*, 2017, págs. 1-6. DOI: 10.1109/ICEngTechnol.2017.8308186.

[70] R. Yamashita, M. Nishio, R. K. G. Do y K. Togashi, «Convolutional neural networks: an overview and application in radiology,» *Insights into imaging*, vol. 9, n.º 4, págs. 611-629, 2018.

[71] Y. Bengio, I. Goodfellow y A. Courville, *Deep learning*. MIT press Massachusetts, USA: 2017, vol. 1.

[72] K. O'Shea y R. Nash, «An introduction to convolutional neural networks,» *arXiv preprint arXiv:1511.08458*, 2015.

[73] M. Lin, Q. Chen y S. Yan, «Network in network,» *arXiv preprint arXiv:1312.4400*, 2013.

[74] T. M. Mitchell y col., «Machine learning,» 1997.

[75] C. M. Bishop, *Pattern recognition and machine learning*. Springer, 2006.

[76] N. Lavesson y P. Davidsson, «Evaluating learning algorithms and classifiers,» *International Journal of Intelligent Information and Database Systems*, vol. 1, n.º 1, págs. 37-52, 2007.

[77] C. Catal, «Performance evaluation metrics for software fault prediction studies,» *Acta Polytechnica Hungarica*, vol. 9, n.º 4, págs. 193-206, 2012.

[78] M. Hossin y M. Sulaiman, «A review on evaluation metrics for data classification evaluations,» *International Journal of Data Mining & Knowledge Management Process*, vol. 5, n.º 2, pág. 1, 2015.

[79] Q. Gu, L. Zhu y Z. Cai, «Evaluation measures of the classification performance of imbalanced data sets,» en *International symposium on intelligence computation and applications*, Springer, 2009, págs. 461-471.

[80] M Hossin, M. Sulaiman, A Mustapha, N Mustapha y R. Rahmat, «A hybrid evaluation metric for optimizing classifier,» en *2011 3rd Conference on Data Mining and Optimization (DMO)*, IEEE, 2011, págs. 165-170.

[81] N. Japkowicz, «Assessment metrics for imbalanced learning,» *Imbalanced learning: Foundations, algorithms, and applications*, págs. 187-206, 2013.

[82] W. Drzewiecki, «Thorough statistical comparison of machine learning regression models and their ensembles for sub-pixel imperviousness and imperviousness change mapping,» *Geodesy and Cartography*, vol. 66, 2017.

[83] K. J. Danjuma, «Performance evaluation of machine learning algorithms in post-operative life expectancy in the lung cancer patients,» *arXiv preprint arXiv:1504.04646*, 2015.

[84] R. S. Borja Robalino, «Método de concordancia bayesiano y su aplicación en problemas de clasificación multiclase con categorías desequilibradas,» Tesis de mtría., Universitat Politècnica de Catalunya, 2019.

[85] A. A. Cárdenas y J. S. Baras, «Evaluation of classifiers: Practical considerations for security applications,» en *Proc. AAAI Workshop Evaluation Methods for Machine Learning*, 2006, págs. 777-780.

[86] N. Safiullin y B. Safiullin, «Static and dynamic models in economics,» en *Journal of Physics: Conference Series*, IOP Publishing, vol. 1015, 2018, pág. 032 117.

[87] A. Pittermann y J. Pittermann, «Getting bored with htk? using hmms for emotion recognition from speech signals,» en *2006 8th international Conference on Signal Processing*, IEEE, vol. 1, 2006.

[88] B. Vlasenko, B. Schuller, A. Wendemuth y G. Rigoll, «Frame vs. turn-level: emotion recognition from speech considering static and dynamic processing,» en *International Conference on Affective Computing and Intelligent Interaction*, Springer, 2007, págs. 139-147.

[89] E. Bozkurt, E. Erzin, Ç. E. Erdem y A. T. Erdem, «Improving automatic emotion recognition from speech signals,» en *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[90] S. Steidl, *Automatic classification of emotion related user states in spontaneous children's speech*. Logos-Verlag, 2009.

[91] P. Dumouchel, N. Dehak, Y. Attabi, R. Dehak y N. Boufaden, «Cepstral and long-term features for emotion recognition,» en *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[92] T. Vogt y E. André, «Exploring the benefits of discretization of acoustic features for speech emotion recognition,» en *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[93] S. Planet, I. Iriondo, J. C. Socoró, C. Monzo y J. Adell, «GTM-URL contribution to the INTERSPEECH 2009 Emotion Challenge,» en *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[94] C.-C. Lee, E. Mower, C. Busso, S. Lee y S. Narayanan, «Emotion recognition using a hierarchical binary decision tree approach,» *Speech Communication*, vol. 53, n.º 9-10, págs. 1162-1171, 2011.

[95] S. Tripathi, A. Kumar, A. Ramesh, C. Singh y P. Yenigalla, «Deep learning based emotion recognition system using speech features and transcriptions,» *arXiv preprint arXiv:1906.05681*, 2019.

[96] S. Basu, J. Chakraborty y M. Aftabuddin, «Emotion recognition from speech using convolutional neural network with recurrent neural network architecture,» en *2017 2nd International Conference on Communication and Electronics Systems (ICCES)*, IEEE, 2017, págs. 333-336.

[97] A. B. A. Qayyum, A. Arefeen y C. Shahnaz, «Convolutional Neural Network (CNN) Based Speech-Emotion Recognition,» en *2019 IEEE International Conference on Signal Processing, Information, Communication & Systems (SPICSCON)*, IEEE, 2019, págs. 122-125.

[98] N. Dave, «Feature extraction methods LPC, PLP and MFCC in speech recognition,» *International journal for advance research in engineering and technology*, vol. 1, n.º 6, págs. 1-4, 2013.

[99] N. Singh, P. R. Khan y R. S. Pandey, «MFCC and Prosodic Feature Extraction Techniques: A Comparative Study,» *International Journal of Computer Applications*, vol. 54, págs. 9-13, sep. de 2012. DOI: 10.5120/8529-2061.

[100] J. Martinez, H. Perez, E. Escamilla y M. M. Suzuki, «Speaker recognition using Mel frequency Cepstral Coefficients (MFCC) and Vector quantization (VQ) techniques,» en *CONIELECOMP 2012, 22nd International Conference on Electrical Communications and Computers*, IEEE, 2012, págs. 248-251.

[101] S. R. Livingstone y F. A. Russo, «The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English,» *PloS one*, vol. 13, n.º 5, e0196391, 2018.