Diseño y desarrollo de una herramienta de detección de la expresión facial de las emociones y su uso en el tratamiento y detección de la Alexitimia

Autor: Arturo Sánchez Palacio

Directores: Gabriel Antonio Valverde Castilla y Raúl Arrabales Moreno

Máster Data Science para Finanzas (2018 -2019)

A Raúl Moreno y Gabriel Antonio Valverde Castilla
por su inestimable ayuda y trabajo.
A todo el equipo de Serendeepia Research.
A mi familia y amigos por su apoyo incondicional.

ÍNDICE

1. INTRODUCCIÓN

1.1 PLANTEAMIENTO DEL PROYECTO

Este Trabajo Fin de Máster se encuentra enmarcado en mis prácticas extracurriculares propias de este máster en la empresa Serendeepia Research. Dentro de dicha empresa mi tutor Raúl Arrabales Moreno desarrolla actualmente una línea de productos destinada al tratamiento y detección de la Alexitimia (trastorno psicológico que se explicará en el Capítulo 3). Este trabajo presenta el desarrollo de uno de dichos productos denominado *Prolexitim Emotions Mirror*.

Este producto consiste en una aplicación para ordenador y tablet (en versiones posteriores podría extenderse también a dispositivos móviles) capaz de detectar la emoción expresada en el rostro del usuario. A grandes rasgos el proceso sería el siguiente:

- La aplicación se conecta a la webcam del dispositivo.
- El dispositivo proyecta en pantalla imagen del usuario.
- Cada cierto tiempo (cada segundo o menos) un fotograma de la cámara es enviado a un modelo de Deep Learning que devuelve la emoción proyectada.
- Esta emoción aparece por pantalla en la aplicación.

Este proyecto plantea dos resultados valiosos, por una parte la aplicación en sí puede ayudar a personas (especialmente niños) con problemas para la expresión de las emociones y por otra parte el modelo matemático subyacente puede ser reutilizado en una gran variedad de campos.

Habitualmente los seres humanos desarrollan la capacidad de transmitir emociones mediante la interacción con otras personas. Los gestos se van adquiriendo de manera natural a medida que los niños crecen y se producen una serie de asociaciones (sonrisa con felicidad, ceño fruncido con ira…) que facilitan en gran medida la comunicación con el resto de personas (diversos autores sostienen que un 65% de la comunicación se produce de manera no verbal y dentro de ese 65% las expresiones faciales son uno de los canales más relevantes).

Algunas personas sin embargo presentan dificultades a la hora de desarrollar estas asociaciones y adquirir dichos hábitos comunicativos, esto supone un círculo vicioso; estas personas tienen problemas para la comunicación, son desplazadas, luego tienen menos posibilidades de comunicarse, luego tienen menos posibilidades para aprender y mejorar su comunicación y así sucesivamente… Diversos psicólogos y psiquiatras teorizan que esta capacidad se puede “entrenar”, es decir, que a base de repetir veces y veces una sonrisa cuando uno se siente feliz el cuerpo termina adoptando dicho hábito y haciéndolo de una manera natural. En esta línea la aplicación podría ser el software perfecto para realizar este entrenamiento sin necesidad de asistencia de un tercero.

Además el modelo matemático subyacente se puede aplicar en campos mucho más generales. A continuación se presentan una serie de posibles aplicaciones en distintas situaciones:

- Seguridad vial. Estos modelos se pueden aplicar para que el coche sea capaz de detectar determinadas emociones asociadas a una mayor probabilidad de accidente como son la ira. Además introduciendo una base de datos en la que se presenten imágenes de personas cansadas este mismo modelo se podría emplear para detectar situaciones en las que el conductor se encuentra cansado o somnoliento indicándole que haga un descanso.
- Entrevistas de trabajo. Monitorear las emociones del entrevistado puede permitir descubrir hasta qué punto la persona se adaptaría al supuesto trabajo o puede estar mintiendo para lograrlo. Este conocimiento enriquecería en gran manera la información extraída a partir de las respuestas a las preguntas.
- Investigación de mercado. La detección de emociones se puede aplicar a distintas subáreas de este campo; desde evaluación de los anuncios publicitarios (se puede monitorear la emoción del espectador a lo largo de un anuncio observando qué puntos tienen una mayor carga emocional y qué emoción se produce) hasta reacciones de personas ante baldas de supermercados.
- Evaluación de creaciones audiovisuales. Es posible monitorear la experiencia de usuario a la hora de navegar por una página web, ver una película o probar un videojuego descubriendo hasta qué punto estas creaciones están dejando una huella en el usuario o son un proceso de interacción vacío. Estos dos últimos puntos plantean un gran valor para el modelo creado en campos de experiencia de cliente.
- Seguridad. En algunos puntos del mundo los atracos en cajeros son sucesos prácticamente del día a día. Se podrían establecer sistemas que activaran alarmas y negaran la disposición de efectivo cuando el cliente que realiza la operación transmite pánico.
- Apoyo en sesiones psicológicas. El modelo permitiría monitorear el estado anímico del paciente detectando expresiones que transmiten sentimientos que el especialista podría haber pasado por alto en un momento dado revelando temas de especial importancia o sensibilidad para el primero.

En las siguientes secciones se presenta el proceso seguido al completo comenzando por el planteamiento de la idea inicial, la elección y justificación del uso de modelos de Deep Learning, la selección de bases de datos para el entrenamiento del modelo, el planteamiento de las dos aplicaciones agrupadas en este trabajo, los resultados obtenidos a partir del modelo en entrenamiento y validación y la puesta en producción de ambas con especial hincapié en la segunda por la gran problemática actual para la puesta en producción de ciertos modelos en Tensorflow.

1.2. ESTRUCTURA DE LA MEMORIA

Tras esta introducción en el segundo capítulo se presentan distintas aplicaciones de la Inteligencia Artificial en el mundo de la salud, uno de los campos que más modificado se verá por la aparición de la Inteligencia Artificial según anuncian los expertos.

En el tercer capítulo se introduce el concepto de Alexitimia en torno al que gira este proyecto. Se presentan algunas aplicaciones de Inteligencia Artificial en el mundo de la psicología. Además se explican los procesos de detección de la Alexitimia más habituales siendo uno de ellos el TAS 20 implementado en este trabajo.

Una vez introducido el problema se presentan los conceptos técnicos básicos para abordar los modelos de redes neuronales en torno a los que gira este trabajo. Se empieza explicando el concepto de Inteligencia Artificial y desde él se profundiza en Aprendizaje Automático y por último en Aprendizaje Profundo.

En el quinto capítulo se presentan las herramientas informáticas más importantes empleadas en este Trabajo Fin de Máster, a saber: PyCharm, TensorFlow, GitHub, Polyaxon y Jekyll.

En el sexto capítulo se presentan las bases de datos sobre las que se entrenarán y validarán los distintos modelos así como la base de datos estructuradas generada a partir de la información obtenida mediante el test de detección de la Alexitimia y las cuestiones sociodemográficas adjuntas a este.

En el séptimo capítulo se explican las decisiones tomadas y el proceso seguido para el desarrollo de este proyecto aclarando puntos como los criterios de validación y bondad del modelo. Además se justifican distintas decisiones de programación y negocio.

En el octavo capítulo se presentan los dos métodos de puesta en producción de los modelos diseñados.

En el noveno capítulo se presentan los resultados más relevantes del trabajo: modelos entrenados junto a sus métricas de validación y enlaces y explicación de las distintas aplicaciones planteadas.

Finalmente en el décimo capítulo se presentan las conclusiones de este trabajo y algunas ideas de mejora que el autor espera poder aplicar en un futuro con el fin de lograr una aplicación plenamente funcional.

2. INTELIGENCIA ARTIFICIAL Y SALUD

En los últimos años la discusión sobre la integración de herramientas tecnológicas avanzadas en los tratamientos médicos ha cobrado importancia. Entre estas herramientas destacan aquellas basadas en técnicas de Inteligencia Artificial. Esto se debe principalmente a dos factores: el aumento exponencial que está experimentando este área en los últimos años viviendo la tercera era de la Inteligencia Artificial y la desmesurada cantidad de datos (estructurados y no estructurados) que genera el sector médico en la actualidad (150 exabytes (10^8 bytes) tan solo en Estados Unidos con una predicción de crecimiento anual del 48%) (Jiang, 2017). Este aumento exponencial de la Inteligencia Artificial ha sido potenciado por el éxito en su aplicación a distintos problemas que previamente no se habían podido resolver, al abaratamiento de los sistemas de procesamiento y almacenamiento de datos y a la democratización del conocimiento.

Algunas de estas ideas ya han sido puestas en práctica en distintos campos de la medicina, por ejemplo, la herramienta Watson for Oncology[1] realiza propuestas de tratamientos para distintos tipos de cáncer coincidiendo en un 99% de los casos con el diagnóstico producido por un especialista. Esta herramienta incluye técnicas de tratamiento del lenguaje natural que han analizado distintos artículos científicos, manuales de buenas prácticas y cuadernos médicos de distintos especialistas en la materia.

Otro ejemplo de uso de Inteligencia Artificial es el chatbot Sensely[2] capaz de recolectar información de los usuarios mediante textos, habla, fotos o vídeos. Distintos algoritmos tratan toda esta información y la comparan con la base de datos principal emitiendo un diagnóstico provisional al paciente y situándolo en una cola para ser atendido por un médico real según la urgencia que revelen sus síntomas.

Por último un campo que está cobrando fuerza en los últimos años con el potente desarrollo del aprendizaje profundo es la creación de historias clínicas electrónicas (Esteva, A. : 2019). Estas bases de datos aumentan de manera astronómica: una organización médica de tamaño grande puede capturar transacciones médicas de más de diez millones de pacientes a través de los años. Una única hospitalización genera en torno a 150.000 datos. Así la agregación de todos los datos en una base explotada por algoritmos de Inteligencia Artificial podría alcanzar el equivalente de 200.000 años de conocimiento médico y 100 millones de años de datos obtenidos a partir del paciente almacenándose en esa base de datos todo tipo de enfermedades incluyendo mutaciones y enfermedades raras. El gran problema al que se enfrenta este campo es la diversidad de estructuras y escrituras presentes en los informes de los distintos profesionales y clínicas (Balmonte. R: 2015).

[1] https://www.ibm.com/es-es/marketplace/clinical-decision-support-oncology

[2] http://www.sensely.com/

En aras de solucionar este problema en Estados Unidos se ha comenzado a implantar un modelo de reporte básico a seguir por todos los hospitales y a la vez se está procediendo al procesado y volcado de los informes escritos hasta ahora en un modelo común abriéndose camino hacia una base de datos común como la mencionada al inicio de esta sección. La siguiente figura (*Figura 1*) presenta una representación superficial de este proceso:

Figura 1. Esquema del proceso de creación de un modelo común de reporting.

Dentro de las herramientas de este mismo campo, la Visión Artificial es una rama de la Inteligencia Artificial cuyo fin es procesar, analizar y comprender mediante ordenadores imágenes del mundo real logrando extraer conocimiento y datos numéricos que puedan ser procesados por un ordenador (Dana: 1982).

En Medicina destaca actualmente la utilización de este tipo de técnicas para la interpretación de radiografías. Dentro de estas técnicas se distinguen dos ramas, aquellas autónomas que no requieren la intervención humana (detección de tumores en radiografías, clasificación de tumores (malignos o benignos), etc. y herramientas de apoyo a los doctores, por ejemplo, detección y segmentación de imágenes para diagnósticos urgentes, por ejemplo en pacientes con obstrucciones en arterias cerebrales que pueden causar daños cerebrales permanentes en cuestión de minutos o detección de áreas problemáticas en radiografías (Liebeskind : 2018).

Otro posible uso de las técnicas de visión artificial son la cirugía no invasiva y la cirugía automatizada mediante robots, por ejemplo, a la hora de suturar una herida un algoritmo de visión artificial podría procesar la imagen procediendo a la colocación de los puntos en la mejor trayectoria posible obteniéndose esta mediante un algoritmo de optimización que considere restricciones como las articulaciones o las particularidades del cuerpo del paciente. De la misma manera mediante el entrenamiento por imágenes, un robot puede lograr poner puntos automáticamente aprendiendo las maniobras físicas a partir de imágenes de médicos haciéndolo (Iscimen: 2015).

Otro de los principales campos de desarrollo de la visión artificial es el desarrollo de sistemas de navegación de espacios interiores para personas con una discapacidad visual grave (Feng: 2018). La creación de software asociado a dispositivos como los teléfonos móviles o las gafas virtuales ha permitido la construcción de interfaces muy cómodos y accesibles para personas con discapacidad visual permitiéndoles una visión mejorada que podría derivar en un futuro muy lejano a la autonomía de dichas personas necesitando solo su smartphone y unas gafas de realidad virtual y pudiendo prescindir de otros instrumentos tradicionales como el bastón o los perros guía. Esta idea atrae en gran medida al colectivo por la discreción que implicaría reduciendo en gran medida su discriminación (miradas indiscretas, exclusión…). La gran dificultad que afrontan estos dispositivos es la implementación de algoritmos de deeplearning lo suficientemente certeros y potentes como para poder ofrecer un servicio funcional en tiempo real.

Los proyectos de Inteligencia Artificial en Salud se pueden dividir en dos grandes vertientes; aquellos que buscan asistir a profesionales en sus labores diarios intentando automatizarlas (la detección de tumores antes mencionada sería un claro ejemplo) y aquellas que intentan mejorar la calidad de vida de las personas con diversidad funcional (como pueden ser las personas con discapacidad visual antes mencionada). Este proyecto se encuentra principalmente encuadrado en esta segunda vertiente buscando ayudar a personas con dificultad para la transmisión de emociones a identificar y combatir este problema. El fin último de este proyecto es aportar un granito de arena a la mejor convivencia y adaptación de las personas que padecen estas dificultades mejorando su independencia y calidad de vida.

3. ENFOQUE DEL PROBLEMA

3.1 INTRODUCCIÓN DEL CONCEPTO DE ALEXITIMIA

El fin último de este trabajo es presentar el desarrollo de una herramienta de Inteligencia Artificial para la detección y el tratamiento de la Alexitimia. En esta sección se ofrece una pequeña introducción a este trastorno. Además se presentan el TAS-20 y el TAT dos herramientas de detección de la Alexitimia. La primera ha sido implementada en este proyecto.

El término Alexitimia es acuñado por el psiquiatra e investigador Peter Emanuel Sifneos (1920 - 2008) en la década de los 70 cuando junto a John Case Nemiah (1918 - 2009) estudia las entrevistas de un grupo de pacientes con trastornos psicosomáticos en las cuales se busca valorar el pensamiento libre e imaginación de los entrevistados. En estos pacientes descubren un gran dificultad para la verbalización de sus emociones, una fantasía limitada y un estilo de pensamiento literal, sin matices y orientado a lo externo. Así surge el concepto de Alexitimia (*sin palabras para los sentimientos* en griego). Es importante comprender que Sifneos no es el primero en distinguir esta serie de síntomas en pacientes sino el primero en agrupar todos estos trastornos en un concepto.

En los años posteriores este concepto va cobrando importancia y Graeme J. Taylor y sus colaboradores presentan en 1997 la primera caracterización de la Alexitimia (Taylor: 1997):

- Problemas para identificar y comunicar sus sentimientos.
- Dificultad para distinguir sentimientos y sensaciones corporales propias de la activación emocional.
- Imaginación muy limitada visible en las escasas fantasías.
- Estilo cognitivo orientado a lo externo y concreto. El pensamiento externo se caracteriza por su objetividad y pragmatismo. Es un estilo de pensamiento en el que se contemplan y enuncian hechos que suceden pero no se reflexiona sobre las emociones y sentimientos que dichos hechos despiertan en el individuo. Por ejemplo, hablar sobre un funeral narrando hechos concretos (la iglesia donde tuvo lugar, su duración…) pero sin valorar aspecto como la tristeza que este pudo suponer en el individuo.

Estas características han sido secundadas por multitud de investigadores adeptos a distintas ramas de la psiquiatría y están indisolublemente ligadas a un déficit en la capacidad cognitiva para procesar y regular las emociones.

Es esencial comprender que este trastorno afecta a todos los sectores de la población en términos de edad, género o raza; por ello para el diseño de herramientas y modelos es imprescindible el empleo de bases de datos que garanticen la diversidad de los individuos estudiados en ellas con el fin de lograr una generalidad suficiente.

En los últimos años se han planteado distintas soluciones de Inteligencia Artificial en el ámbito de la Psicología. A continuación se presentan algunos ejemplos:

- Diversas empresas han creado prototipos de terapeutas virtuales capaces de comprender y comunicarse mediante lenguaje natural entre las que destaca X2A[3] con productos como Karim que es una plataforma para ayuda a refugiados íntegramente implementada en árabe o Emma destinada a ayudar a pacientes con ansiedad y fobia social.
- El MIT (Massachusetts Institute of Technology) ha desarrollado un modelo de deep learning capaz de detectar individuos con depresión únicamente mediante el análisis de textos y audios mediante la detección de ciertos patrones del lenguaje (Alhanai: 2018).

Es importante comprender los dos factores principales que hacen que los "terapeutas virtuales" hayan tenido una aceptación moderadamente buena y que indican un crecimiento exponencial en los años venideros. Por una parte, las aplicaciones antes mencionadas son gratuitas o tienen un coste ínfimo en comparación con el coste de una consulta con un especialista en la materia. En segundo lugar, se produce un suceso curioso, en ocasiones una persona siente más facilidad para compartir sus sentimientos y pensamientos más íntimos con una máquina que con una persona por vergüenza (a día de hoy, las consultas relacionadas con salud mental siguen muy estigmatizadas) o por miedo a ser juzgadas. Estas aplicaciones generan seguridad en el usuario que conserva su anonimato en todo momento.

En esta misma línea la investigación llevada a cabo hasta la fecha indica que no existen herramientas basadas en Inteligencia Artificial para el tratamiento de la Alexitimia. Este proyecto plantea una primera solución y abre camino en este campo que tanto podría beneficiar a las personas afectadas por esta condición más común de lo que se presupone. La Alexitimia tiene una prevalencia del 10% en población no clínica y llega a más del 45% en población clínica (trastorno por estrés postraumático, trastornos del espectro del autismo, dolor crónico…).

Debe entenderse que la herramienta presentada en este trabajo no busca sustituir la labor del terapeuta si no complementarla de una forma eficiente, no debe concebirse como una cura si no como un apoyo complementario a las sesiones y tratamientos propuestos por el especialista (de igual manera que una muleta no sustituye al traumatólogo). La herramienta busca reforzar la confianza de los usuarios (más facilidad para comunicarse y comprender a los seres que le rodean) y automatizar ejercicios monótonos permitiendo el mejor aprovechamiento del tiempo de los especialistas en la consulta.

El planteamiento que se realiza parte de que el entrenamiento de expresiones faciales relacionadas con emociones (sonreír con felicidad, fruncir el ceño con ira…) puede tener un impacto positivo en áreas como las habilidades sociales, la introspección o la identificación de la propia emoción.

[3] https://www.x2ai.com/

Así se plantea un protocolo de entrenamiento en el que paciente diagnosticados con Alexitimia realizan durante un periodo inicial de cuatro semanas un entrenamiento de media hora al día con la aplicación. Mediante la aplicación del TAS20 y el TAT (presentados a continuación) se evalúa el grado de Aleximitia en el paciente antes y después de dicho periodo comprobando si se perciben mejorías significativas.

3.2. DETECCIÓN

En este proyecto se emplean para la detección de la Alexitimia dos herramientas: el TAS-20 (Toronto Alexithymia Scale) y el TAT (Thematic apperception test).

3.2.1. TAS-20 (Toronto Alexithymia Scale)

El TAS-20 es un cuestionario planteado por (Parker, Bagby, Taylor, Enlder, Schmitz) en 1993. El test se puntúa en una escala de tipo Likert de cinco puntos. El cuestionario se encuentra compuesto por 20 afirmaciones de las que el sujeto debe indicar el grado de acuerdo pudiendo estar “muy en desacuerdo”, “en desacuerdo”, “neutro”, “de acuerdo” o “muy de acuerdo”. Cada respuesta lleva una puntuación asociada obteniendo el sujeto una puntuación final comprendida entre 20 y 100. El umbral de la alexitimia incluye todos aquellos sujetos con puntuación igual o superior a 61.

Un análisis factorial realizado sobre las respuestas de multitud de individuos muestran la detección de tres factores principales mediante este test:

- Problemática para identificar sentimientos y separarlos de las sensaciones corporales o fisiológicas que acompañan a la actividad emocional.
- Dificultad para expresar los sentimientos a los demás.
- Estilo cognitivo orientado hacia el exterior.

Varios estudios llevados a cabo por G. J. Taylor muestran que el componente cultural no influye en los resultados. En ocasiones los tests de este tipo se pueden ver fuertemente influenciados por la cultura de la persona a estudiar (en muchas culturas la expresión de los sentimientos es percibida como una muestra de debilidad y algunas culturas la censuran hasta puntos más extremos como la cultura nipona que considera la expresión de las emociones como algo impuro) pero este no es el caso.

Con el fin de comprobar la validez del test, Taylor realiza en 1993 un experimento en el que toma tres muestras compuestas por estudiantes alemanes, estadounidenses y canadienses (Parker: 1993). Los resultados muestran una estructura factorial idéntica a la anterior en todos los casos. En su artículo “Validez psicométrica de la escala de Alexitimia de Toronto (TAS-20): Un estudio transcultural” Darío Páez y su equipo muestran que estas mismas estructuras se replican en la población española (en concreto el estudio toma su muestra en Murcia).

En este proyecto se aborda la detección y valoración de todos los factores mediante el TAS-20 y el TAT y tras ello se trabaja la dificultad para identificar y comunicar sentimientos con el espejo virtual que será explicado más adelante.

En el Anexo III se encuentra el código en el que se implementa la plataforma online para la respuesta del test y el almacenamiento de sus resultados que se puede observar y realizar en el siguiente enlace[4].

3.2.2. TAT (Test de Apercepción Temática)

El Test de Apercepción Temática (TAT) fue diseñado por el psicólogo estadounidense Henry A. Murray y la psicoanalista Christiana D. Morgan en los años 30 en la Universidad de Harvard.

La idea surge de una de las estudiantes de Murray que le contó en clase que cuando su hijo estaba enfermo pasaba el tiempo inventando historias a partir de las imágenes que veía en las revistas y que eso le llevaba a preguntarse si las imágenes se podían emplear para explorar la personalidad del individuo según las historias elaboradas.

El TAT está compuesta por 31 láminas sobre las cuales se pueden construir narraciones. A medida que el sujeto va describiendo las imágenes se puede observar el punto común que encuentra en todas ellas aunque las historias vayan variando. El TAT es actualmente uno de los tres instrumentos más empleados en la exploración clínica junto con el test Rorschach y el Inventario Multifásico de Personalidad de Minesotta (MMPI).

A la hora de aplicar el test las imágenes se presentan en grupos uniformes (habitualmente de diez en diez en dos sesiones). Algunas imágenes se presentan a todos los sujetos mientras que otras son específicas para mayores o menores de edad o de un género u otro. De esta manera, del total de láminas solo veinte se aplican a cada sujeto debido a las exigencias de cada lámina. Las láminas son dibujos, fotografías, reproducciones de cuadros o grabados.

A la hora de pasar la prueba el sujeto debe ir visualizando las láminas de una en una y para cada una inventar una historia que contenga pasado, presente y futuro con especial atención a lo que los personajes puedan estar sintiendo o pensando. El examinador debe tomar nota textual de todo lo que dice el paciente sin intervenir en el relato.

[4] http://www.serendeepia.com/prolexitim_tas20/

Para terminar esta sección se presentan dos imágenes del TAT con sus posibles interpretaciones:

Figura 2. Lámina 1 del TAT.

Figura 3 . Lámina 12M del TAT.

La *Figura 2* muestra un niño contemplando un violín que está sobre la mesa. Esta lámina es siempre el comienzo de la prueba. Poner de relieve las actitudes del sujeto hacia el rendimiento, sus metas, aspiraciones, dificultades, esperanzas… Estas metas pueden ser propias o impuestas.

La *Figura 3* es una lámina destinada a individuos masculinos de cualquier edad y permite al sujeto mostrar sus sentimientos y esperanzas respecto a la terapia. Puede señalar dependencias pasivas: homosexualidad, deferencia, obediencia en la relación entre varones…

Además de la aportación social que supone el desarrollo de esta herramienta, este proyecto abre un nuevo campo en las aplicaciones de la Inteligencia Artificial. Hasta la fecha no se han publicado resultados de ninguna herramienta destinada al tratamiento de la Alexitimia.

4. CONCEPTOS TÉCNICOS

4.1. INTELIGENCIA ARTIFICIAL Y APRENDIZAJE AUTOMÁTICO

En esta sección se explican de manera superficial los principales conceptos técnicos implicados en la creación del modelo matemático subyacente a la aplicación.

Definir de manera exacta el concepto Inteligencia Artificial es un problema aun a resolver por la comunidad científica pero la definición más aceptada hasta la fecha considera la IA como la rama de las Ciencias Computacionales que busca simular en un ordenador comportamiento inteligente. Dentro de este área destaca en los últimos años el crecimiento del Machine Learning o Aprendizaje Automático.

El concepto de Machine Learning (englobado en la Inteligencia Artificial) supone un cambio en el paradigma de programación tradicional vigente hasta mediados del siglo XX. En su artículo "Computing Machinery and Intelligence" Alan Turing (conocido como el padre de la Inteligencia Artificial) rescata una cita de Ada Lovelace sobre un dispositivo diseñado por su contemporáneo Charles Babbage destinado a automatizar cálculos relacionados con el análisis matemáticos: "*The Analytical Engine has no pretensions whatever to originate anything. It can do whatever we know how to order it to perform.... Its province is to assist us in making available what we're already acquainted with*" en la que indica que dicho aparato no tiene capacidad de creación, simplemente realiza las tareas que se le ordenan. Esta cita, que pasará a la literatura de la Ciencia Computacional como la Objeción de Lady Lovelace, origina el concepto de Machine Learning o Aprendizaje Automático. Algunos años más tarde E. Tom Mitchell (1997) presenta una definición algo más técnica de este concepto definiendo el aprendizaje de un programa a partir de una experiencia E con respecto a una tarea T y una medida de acierto P a su desarrollo si su éxito en la realización de T medido por P mejora con la experiencia E.

Hasta la aparición de este concepto la programación se concebía como la introducción en un ordenador de una serie de reglas y un conjunto de datos obteniéndose como resultado unas determinadas respuestas. El Machine Learning invierte este proceso; el ordenador recibe un conjunto de datos a los que se les ha asignado la respuesta correcta y devuelve tras procesarlos las reglas que originan estas respuestas a partir de los datos, para ello busca la estructura estadística subyacente al modelo descubriendo así reglas que permiten automatizarlo por eso se dice que los modelos de aprendizaje automático son entrenados en lugar de programados. El siguiente diagrama (*Figura 4*) representa el contraste entre los dos paradigmas programáticos:

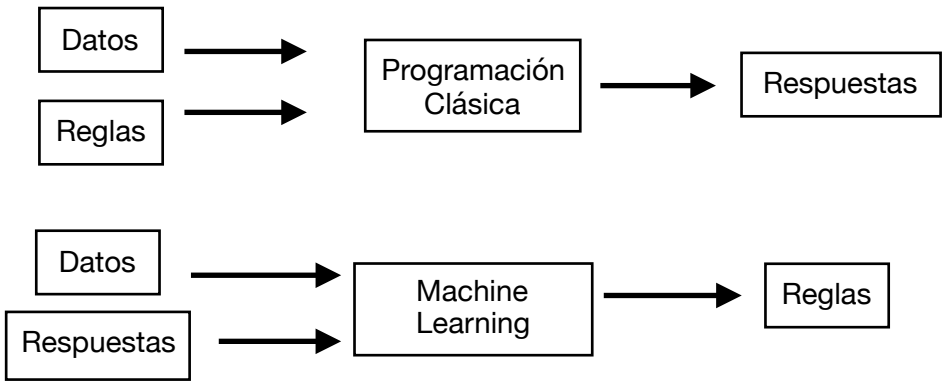


Figura 4. Esquema del cambio de paradigma programacional.

El Aprendizaje Automático tiene su origen en la década de los 90 y gracias a las grandes cantidades de datos disponibles y las mejoras en el hardware computacional a nivel de potencia del procesador y abaratamiento del almacenamiento, se ha impuesto como el área de la Inteligencia Artificial más demandada y aplicada actualmente.

Algunos campos en los que el Machine Learning ha cobrado vital importancia en las últimas décadas son el reconocimiento de voz (permitiendo la transcripción de voz a texto), los asistentes virtuales como Siri o Cortana entrenado para responder de manera correcta a partir de miles de conversaciones o los modelos de puntuación para la concesión de créditos que permiten detectar futuras moras así como lograr una distribución del crédito más justa favoreciendo el acceso al crédito a distintos perfiles.

En este sentido cabe distinguir dos grandes ramas dentro del Machine Learning; el aprendizaje supervisado que aprende a partir de datos etiquetados y el aprendizaje no supervisado que procesa los datos agrupándolos según diversos criterios. El mejor ejemplo para comparar estas ideas son la clasificación (supervisado) y el método clúster (no supervisado). Partiendo de un conjunto de flores de las que se dispone de una serie de características (color, tamaño, número de hojas…); para emplear una clasificación se debería disponer de un conjunto de flores clasificadas, por ejemplo, rosas, claveles, peonías… y a partir de estos datos el modelo trataría de inferir las especies para nuevas flores de las que se conocieran sus características pero siempre devolviendo como resultado alguna de las clases con las que fue entregado. El método clúster por el contrario recibiría solo las características de las flores y realizaría su propia clasificación según la similitud entre ellas y quizá en esta clasificación se mezclarían flores de distintas especies pero muy próximas en apariencia como pueden ser rosas y claveles. La siguiente figura (*Figura 5*) muestra las principales áreas del Machine Learning y algunas de sus aplicaciones:

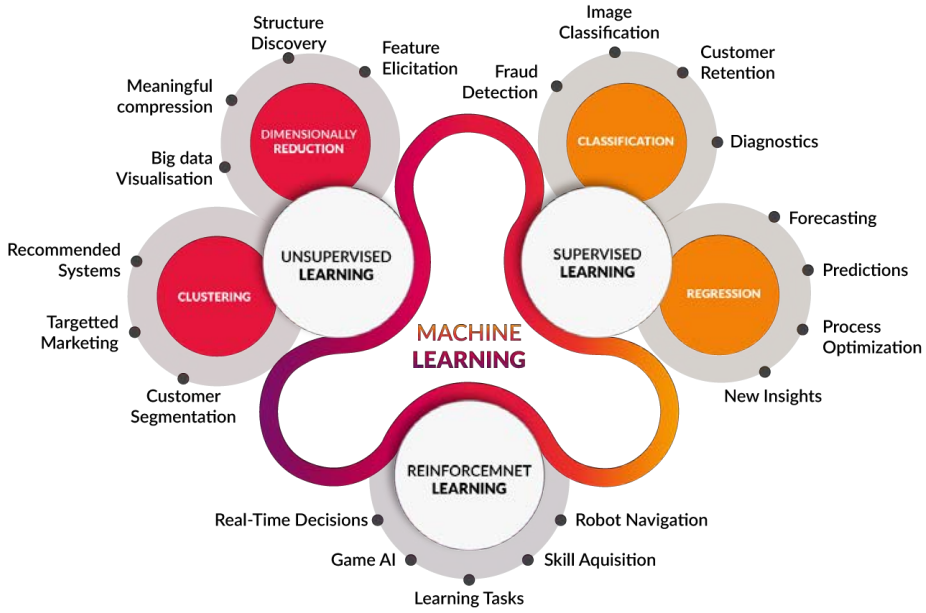


Figura 5. Esquema del Aprendizaje Automático.

El modelo construido en este trabajo es un modelo de clasificación, es decir, un modelo supervisado en el que la variable a predecir es categórica (toma un número finito de valores; en este caso tantos como emociones se consideren). Más en concreto es una clasificación muticlase, es decir, las imágenes se clasificarán en siete emociones distintas como se mencionó antes. La clasificación no es multietiqueta (multilabel) ya que las clases se consideran mutuamente excluyentes no se consideran imágenes en las que el usuario proyecte más de una emoción a la vez.

4.2. APRENDIZAJE PROFUNDO

El Deep Learning o Aprendizaje Profundo es una subárea específica del Machine Learning. En el Machine Learning el modelo absorbe datos y los transforma de manera que generen salidas significativas, una representación que "*aprende*" a partir de su exposición a una gran cantidad de ejemplos. En el Deep Learning este proceso de aprendizaje se produce a partir de capas sucesivas a través de las cuales la representación de los datos se va volviendo más y más significativa. En ningún caso el adjetivo profundo hace referencia a un conocimiento más minucioso de los datos si no a esta idea de sucesivas capas que conforman un modelo, de hecho, al número de capas que forman un modelo se le denomina profundidad del modelo.

Cada capa de un modelo dado está compuesta por varios elementos denominados neuronas. Se distinguen tres tipos de capas:

- Las capas de entrada son aquellas que reciben las enormes cantidades de datos a procesar.
- Las capas ocultas llevan a cabo las distintas operaciones matemáticas que permiten la extracción de características y patrones a partir de los datos. Los modelos con una cierta complejidad se encuentran compuestos por varias capas ocultas.
- La capa de salida es la encargada de generar el output deseado a partir de las transformaciones logradas en las distintas capas ocultas.

Cada nodo o neurona de cada capa tiene asociados unos determinados valores denominados pesos que condicionan la transformación de los datos. Dichos pesos contienen propiamente la información aprendida por el modelo. El proceso por el que se ajustan estos pesos es el siguiente:

1. Se toma aleatoriamente una parte de las muestras de entrenamiento correspondientes a una serie de variables objetivo.
2. Se ejecuta la red sobre dichas variables obteniendo una salida.
3. Se calcula la pérdida comparando la variable objetivo asociada a dichas variables de entrenamiento con la salida del modelo.
4. Los pesos se actualizan de manera que la pérdida se reduzca en dicha submuestra.

Las partes 1 al 3 resultan evidentes en el modelo en el que se trabajará. Se eligen una serie de fotografías etiquetadas, se pasan al modelo se obtiene un resultado y se compara el resultado con la etiqueta. La complejidad resulta en el ajuste de los pesos que se lleva a cabo mediante un método denominado backpropagation.

Backpropagation consiste en afrontar el problema de optimización que supone reducir la pérdida generada en el tercer paso del algoritmo. Para eso se podría plantear un problema básico de optimización del gradiente buscando el mínimo de la función de pérdida cuando su gradiente fuera 0, sin embargo, la complejidad matemática de dicha ecuación se dispara al plantearla en redes en las que se emplean millones y millones de parámetros. Para optimizar el proceso se calcula el gradiente de la función de pérdida para los parámetros dados y se ajustan ligeramente desplazándolas en dirección contraria a la que indica el gradiente logrando así reducir aunque sea mínimamente el valor de la función de pérdida.

Este proceso se denomina descenso del gradiente estocástico por minibatches. Se emplea el término estocástico (sinónimo de aleatorio en matemáticas) porque la submuestra elegida en el primer paso es aleatoria; minibatches es el término inglés empleado para hacer referencia a la submuestra seleccionada.

En el razonamiento anterior se asume que porque una función sea diferenciable se puede calcular explícitamente su gradiente. A la hora de calcular dicha función dentro de una red neuronal dicha función consiste en una sucesión de operaciones tensoriales. Sobre dicha sucesión es posible emplear la regla de la cadena que permite calcular el gradiente como el producto de las derivadas de cada operación. A esta idea se le denomina backpropagation. En términos sencillos, la backpropagation parte de la pérdida total final y va retrocediendo desde las capas superiores a las inferiores aplicando la regla de la cadena para calcular el aporte de cada parámetro a la pérdida total.

Para el caso concreto de la Visión Artificial hay una transformación que cobra especial relevancia: la convolución. Una convolución es en el sentido puramente matemático una operación entre dos funciones que produce una tercera función que expresa como la forma de una función es modificada por la otra. Entrando en materia, una convolución de una imagen es una operación por la cual se filtra una imagen obteniendo una nueva imagen. El objetivo principal de estas técnicas es la condensación de las imágenes a sus características más importantes, modificando los valores asociados a cada píxel de manera que se enfaticen algunas de las características de la imagen. Para cada píxel se trabaja modificándolo en función de sus vecinos. La siguiente figura muestra una imagen antes y después de sufrir una convolución:

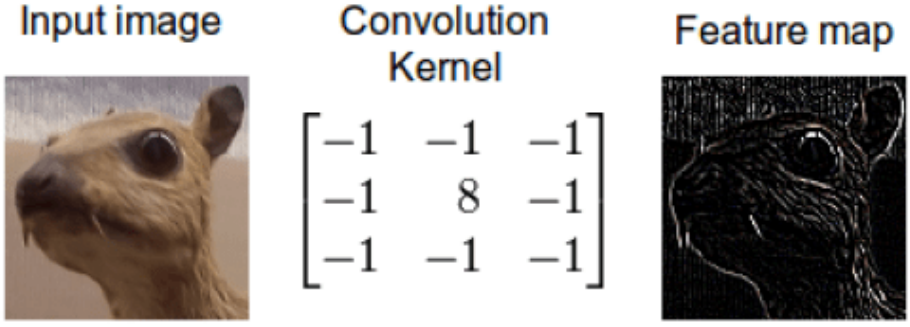


Figura 6. Efecto de una convolución sobre una imagen.

Como se puede observar la convolución desprecia algunas características (se pierden los cambios de tonalidad en el cuello) y sin embargo, remarca algunas otras como el perfil de la cabeza del animal.

Una vez comprendido lo que es una red neuronal el siguiente diagrama (*Figura 7*) muestra el flujo que sigue una imagen en un modelo de Visión Artificial que busca transcribir número escritos a mano. Este diagrama refuerza la idea antes explicada de que el Aprendizaje Profundo hace referencia a la transformación de los datos a través de una serie de capas sucesivas.

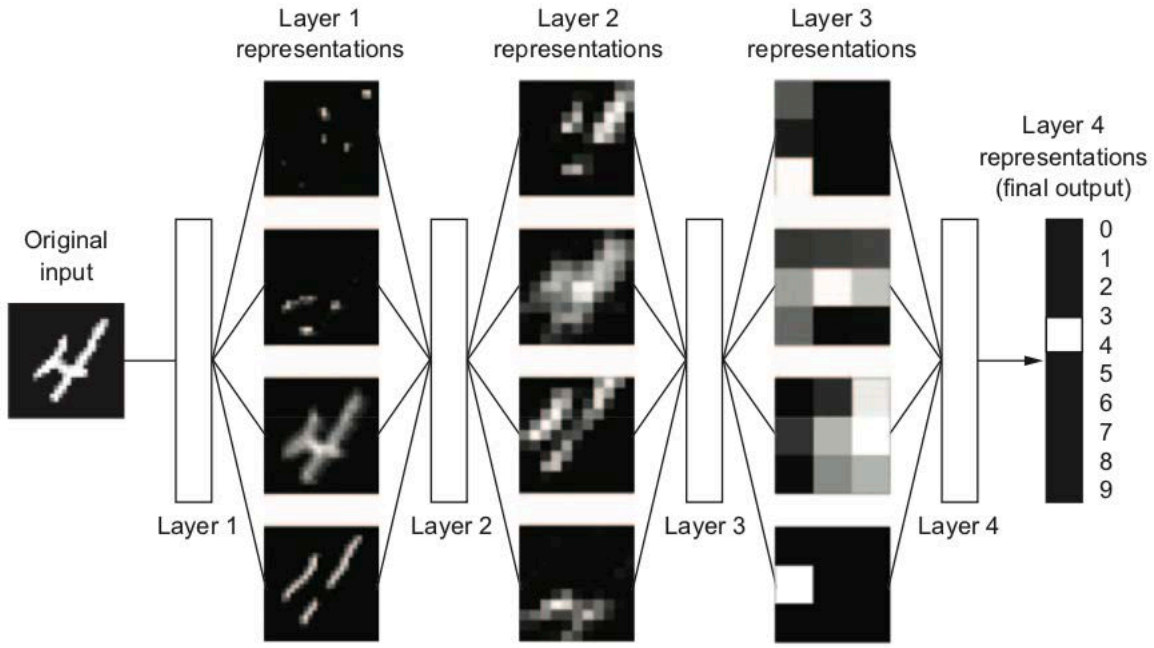


Figura 7. Representación de una red de Aprendizaje Profundo que representa un modelo de clasificación de imágenes.

En el ejemplo, la imagen que el modelo recibe como entrada es un cuatro, en cada una de las capas las convoluciones extraen la información que consideran relevante y modifican los píxeles, detectando aquellos que son útiles (aquellos en los que hay algo escrito) y aquellos que no aportan información (píxeles en fondo negro). A través de las cuatro capas el modelo logra construir representaciones simples, pero que no se pueden interpretar, de las características que representan en la imagen al número cuatro.

La explicación de conceptos como los tensores o las diferentes funciones de pérdidas exceden el alcance de este trabajo por ser demasiado técnicos. En los siguientes recursos bibliográficos (Chollet: 2017, LeCun: 2015, Schmidhuber: 2015) dichos conceptos vienen explicados con lujo de detalle para aquellos lectores ávidos de profundizar en la materia.

Para concluir es importante señalar los factores que han sido decisivo para optar por un modelo de Deep Learning a la hora de abordar este problema. En primer lugar, la aplicación perseguida no requiere una justificación del proceso de obtención del resultado, es decir, el único requerimiento es que el modelo sea capaz de identificar de manera correcta la emoción reflejada en la cara del usuario, sin importar el método o las variables tomadas para ello. Además, en este caso se trabaja con imágenes, es decir, datos no estructurados, formato en el que el Deep Learning suele reportar resultados notoriamente mejores que los obtenidos por otras técnicas englobadas en el Aprendizaje Automático. La superioridad del Deep Learning respecto a otras técnicas de Machine Learning para el trabajo con imágenes se debe a la capacidad de este para hallar relaciones no lineales y para establecer relaciones jerárquicas entre conceptos. El modelo es capaz de generar conceptos en las capas intermedias u ocultas descubriendo características que resultan esenciales para la detección de una determinada emoción.

Además es importante notar que en los últimos años se han hecho públicas distintas bases de datos en las que se recogen imágenes de personas reflejando una cierta emoción. Estas bases ya etiquetadas suponen el combustible necesario para la creación de estos modelos capaces de inferir patrones complejos a partir de enormes cantidades de datos.

Por último y en aras de ajustar el modelo de manera adecuada (el sobreajuste es uno de los mayores problemas cuando se trabaja con modelos de redes neuronales) las bases de datos se dividen dos subconjuntos de entrenamiento y test que permitirán controlar este sobreajuste que privaría al modelo de su capacidad de generalización. Una vez se haya terminado de ajustar el modelo y se ponga en producción se podrá proceder a la validación definitiva del modelo al introducirse en él imágenes de usuarios que no se encuentran recogidos en las fotos de la base de datos.

5. HERRAMIENTAS

En esta sección se realiza una brevísima introducción a las distintas herramientas empleadas a la hora de realizar este TFM.

5.1. PYCHARM[5]

Durante el máster todo el trabajo realizado se desarrolla en notebooks de Python por su comodidad y sencillez a la hora de llevar a cabo labores didácticas sin embargo a la hora de afrontar códigos de una mayor complejidad y envergadura es necesario un editor que permita la generación de distintos scripts en formato .py. En este sentido la herramienta PyCharm publicada bajo licencia Apache y desarrollada por JetBrains presenta grandes ventajas a la hora de procesar código Python entre las que destacan la posibilidad de creación y edición de entornos virtuales de manera intuitiva, la sincronización automática con GitHub desde el propio IDE, las opciones de autocompletado y su funcionamiento transversal a través de los distintos sistemas operativos (Windows, MasOS, Lynux…).

5.2. TENSORFLOW

TensorFlow es en la actualidad el marco de trabajo más empleado para la construcción de modelos de Deep Learning. Es un framework de código libre para computación numérica y aprendizaje automático a gran escala. Aunque trabaja sobre Python los algoritmos son procesados en C++ de alto nivel lo que permite una mayor velocidad y eficiencia a nivel recursos. Además el trabajo sobre Python facilita la puesta en producción de los modelos a gran escala una vez estos se encuentran entrenados.

El trabajo en TensorFlow se estructura mediante una estructura de grafo. Este grafo define como los datos se desplazan experimentando una serie de operaciones (localizadas en los nodos del grafo). Cada nodo en el grafo representa una operación matemática siendo cada arista un array multidimensional o tensor (en el caso unidimensional sería un vector, en el bidimensional una matriz…).

Todo esto se presenta al programador en lenguaje Python siendo los nodos y tensores objetos de Python sin embargo, tal y como se mencionó anteriormente, las operaciones y transformaciones matemáticas en sí han sido programadas en C++. Python se ocupa simplemente de regular el tránsito de la información entre dichas operaciones.

[5] https://www.jetbrains.com/pycharm/

La mayor ventaja de TensorFlow es la gran capacidad de abstracción que este framework supone permitiendo el diseño desde una visión superior del grafo de operaciones sin tener que prestar demasiada atención en la implementación de algoritmos básicos (por ejemplo, los optimizadores en las redes neuronales) y pudiendo abordarse el diseño de los modelos desde una lógica general mientras TensorFlow trata de manera interna algunos de los detalles más pesados (a nivel técnico).

5.3. GITHUB

GitHub es una plataforma que permite el almacenamiento de código para la colaboración y el control de versiones permitiendo el trabajo en equipo; distintas personas pueden estar trabajando a la vez sobre un mismo código. En este caso el GitHub es empleado para la colección y supervisión del código por parte de ambos tutores así como herramienta de control de versiones, una práctica recomendable y necesaria cuando se trabaja como es el caso con código cuya construcción es progresiva y se desarrolla durante periodos prolongados de tiempo (tres meses en este caso).

5.4. POLYAXON[6]

El objetivo de este software es permitir a empresas realizar y poner en producción de manera sencilla y organizada modelos de Aprendizaje Automático y Aprendizaje Profundo. Las distintas herramientas permiten la creación de flujos de trabajos que parten de experimentos sencillos y reproducibles para terminar convirtiéndose en modelo escalables puestos en producción. Polyaxon emplea Kubnernetes para permitir un desarrollo más rápido, seguro y eficientes de aplicaciones de Machine Learning y Deep Learning. Desde un enfoque poco técnico los Kubernetes son servicios en la nube que la conectan con el cluster o equipo local con el que se trabaja permitiendo al usuario especificar los requerimientos computacionales necesarios para la realización de los distintos experimentos gestionándose a la vez dichos servicios en el cluster garantizado su constante disponibilidad. Si alguno de estos recursos cayera Kubernetes trataría de arreglarlo y notificaría al usuario el problema.

La gran ventaja de Polyaxon a la hora de aplicarlo al presente trabajo es que soporta diversos marcos de trabajo para aprendizaje profundo y más en concreto TensorFlow. Polyaxon me ha permitido el entrenamiento de modelos de gran complejidad (la Mobilenet explicada más adelante presenta unos treinta millones de parámetros) sobre bases de datos de un tamaño significativo (RafD ocupa 3GB) que desde un ordenador convencional no habría sido posible.

[6] https://docs.polyaxon.com/concepts/introduction/

5.5. JEKYLL[7]

Jekyll fue el software empleado para la generación del sitio web en el que se localizará la aplicación web. Jekyll permite la simulación local de sitios web con componentes dinámicas. Jekyll permite servir sitios web a la manera en que lo haría un auténtico servidor web. Esto permite la realización de pruebas de manera local hasta el perfeccionamiento del sitio web evitando el consumo de recursos y permitiendo posteriormente la puesta en producción de manera directa y sin ninguna dificultad en un servidor que generará la página en la que se localizará la aplicación presentada en este trabajo.

Por último, a un nivel más administrativo, se ha empleado Trello para la coordinación de tareas y organización de las distintas fases del proyecto y Sublime para la creación de código en html, javascript y css.

[7] https://jekyllrb.com/

6. BASES DE DATOS

En esta sección se presentan las tres bases de datos con las que se trabajará; en este proyecto se emplean dos bases de datos no estructuradas y una base de datos estructurada. La bases de datos no estructuradas están compuestas por imágenes de rostros etiquetadas según la emoción transmitida por los retratados, la base de datos estructurada es de fabricación propia y almacena información sobre los individuos encuestados y sus respuestas a los distintos tests.

6.1. FER 2013

El conjunto Facial Expression Recognition 2013 (FER 2013) fue construido por Pierre Luc Carrier y Aaron Courville. Se creo a partir de imágenes de Google. A partir de las imágenes de Google se recortan las caras, se uniformiza el tamaño de las imágenes (28x28) y se transforman a escala de grises. La base de datos contiene 35887 imágenes en dicho formato etiquetadas según las siete emociones básicas a saber: ira, asco, miedo, alegría, tristeza, sorpresa y neutralidad. La *Figura 9* presenta ejemplos de imágenes contenidas en esta base de datos:

Figura 9. Ejemplo de imágenes del Facial Expresión Recognition Dataset

La base de datos final no está compuesta por las imágenes perse, si no que es una tabla en la que aparece un índice, la emoción codificada como un número de enteros y el array de arrays (28 x 28) en el que se almacenan los valores que toma cada píxel. Además la base de datos al provenir de un reto de Kaggle ya viene dividida en dos submuestras de entrenamiento y validación lo que facilita el proceso presentándose una columna extra en la base con la etiqueta training o test. Dicha información se encuentra almacenada en un archivo csv siendo posible la construcción de un dataframe a partir del mismo.

La tabla (*Tabla 1*) y la gráfica (*Figura 10*) presentadas a continuación muestran la distribución de las emociones en esta base de datos indicando cuántas imágenes asociadas a cada emoción se encuentran en ella:

Emoción	Número de apariciones
Ira	4953
Asco	547
Miedo	5121
Felicidad	8989
Tristeza	6077
Sorpresa	4002
Neutralidad	6198

Tabla 1. Distribución de las emociones en la base FER 2013.

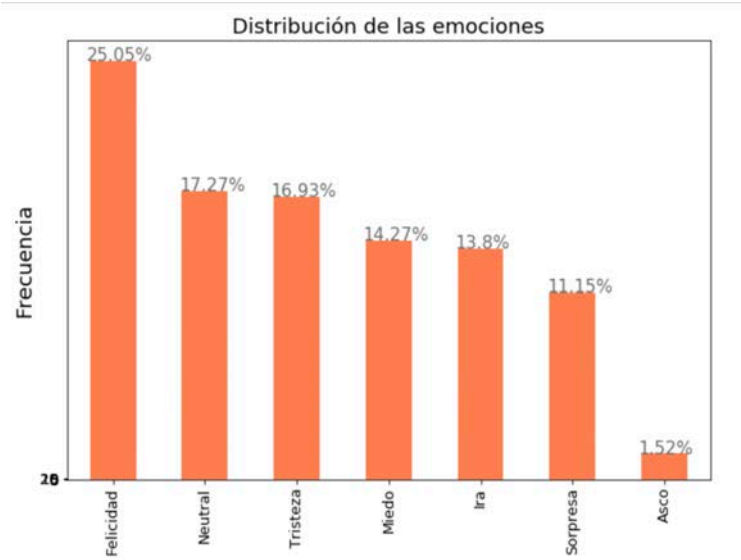


Figura 10. Distribución de las emociones en la base FER 2013.

Las emociones se encuentran más o menos balanceadas a excepción del asco y la felicidad. Esto se debe a que al no ser una base prediseñada o hecha a priori para este proyecto sufre las consecuencias de la disponibilidad online donde no es tan sencillo encontrar una imagen en la que la persona tenga una expresión marcada de asco. Por otra parte las fotos de gente sonriendo son las más habituales en la red.

6.2. RAFD

La base de datos Radboud (Langner: 2010) (RafD por sus siglas en inglés) está conformada por 8050 imágenes tomadas de 68 modelos (hombres, mujeres y niños caucásicos, y hombre holandeses-marroquíes) mostrando ocho emociones diferentes. Fue una iniciativa del Behavioural Science Institute de Raboud adscrito a la universidad de Nijmegen (Países Bajos). Las fotos aparecen con la emoción que transmiten en su título. Las emociones mostradas son ira, asco, miedo, alegría, tristeza, sorpresa, desprecio y neutralidad. Las seis primeras emociones de esta lista constituyen las denominadas por los expertos en la materia las seis emociones básicas (Ekman: 1992). A estas se les añade la neutralidad algo de gran utilidad pues las personas no estamos constantemente expresando emociones y el desprecio. Cada emoción es además mostrada con la mirada en distintas direcciones y fotografiada desde cinco ángulos de manera simultánea. En este caso las imágenes se ajustan más a la realidad por no estar restringido su formato y por combinar sujetos muy diferentes. Esto aunque supone una mejora en los resultados finales exige un mayor preprocesamiento y un aumento considerable en la complejidad del modelo.

Esta base de datos fue elegida por su gran tamaño y su relativa variedad. Si bien sería interesante una base de datos con mayor diversidad étnica (los géneros y la edad son bastante variados) hasta la fecha no existe ninguna base de datos con las emociones etiquetadas que combine más etnias, de hecho, la mayor parte de ellas usan una población más uniforme que el RafD en términos de origen racial. Por último se presenta un ejemplo (*Figura 11*) del tipo de imágenes presentes en esta base de datos:

Figura 11. Ejemplo de imágenes del RafD.

Por último la siguientes tablas (*Tabla 2 y 3*) y la gráfica (*Figura 12*) muestran brevemente la distribución en cuánto a edades (la base solo etiqueta niño o adulto), género y etnia por una parte y la distribución de emociones por otra:

Característica	**Frecuencia**
Hombre	5040
Mujer	3000
Niño	1200
Adulto	6840
Caucásico	5880
Marroquí	2160

Tabla 2. Distribución de género, raza y edad en la base de datos RafD.

Emoción	**Frecuencia**
Neutralidad	1005
Alegría	1005
Tristeza	1005
Ira	1005
Sorpresa	1005
Asco	1005
Desprecio	1005
Miedo	1005

Tabla 3. Distribución de las emociones en la base de datos RafD.

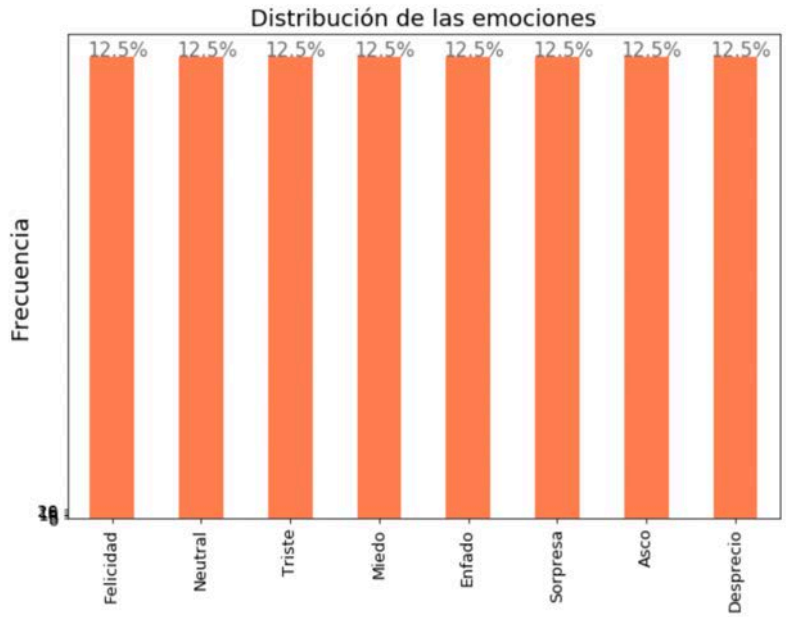


Figura 12. Distribución de emociones en la base de datos RafD.

En este caso se puede observar que las emociones se encuentran perfectamente equilibradas (ver *Figura 12*); esto se debe a que la base de datos fue creada de manera explícita, es decir, se reclutó a una serie de personas que recibieron instrucciones sobre cómo posar y siguieron un minucioso proceso para la toma de las fotografías que posteriormente serían etiquetadas por expertos en la materia.

6.3. BASE DE DATOS DE FABRICACIÓN PROPIA

Para el seguimiento de la evolución de los pacientes es necesaria la construcción de una base de datos a partir de los resultados recogidos por la aplicación web (presentada en el Capítulo 9) en la que el encuestado completa las preguntas del TAS-20 así como una serie de datos sociodemográficos. La realización de este test debe ser supervisada por un experto que además se ocupará de recoger el consentimiento explícito[8] para el tratamiento de los datos tal y como recoge la Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos y Garantía de los Derechos Digitales.

Dicha base de datos almacena de manera estructurada (todos los datos sean cuantitativos o cualitativos) la información de cada usuario en la que se recoge su resultado en cada uno de los factores de la Alexitimia (Taylor: 1997), su puntuación total, su edad, su sexo, su género, su edad, si es zurdo o diestro, su nivel de estudios, su clase social, su número de hermanos (y su posición entre ellos), su país de residencia, su origen étnico y su profesión. Resulta evidente que para el almacenamiento de estos datos especialmente sensibles es necesaria la seudonimización[9] de los mismos con objeto de cumplir la Ley de Protección de Datos. Para ello se pide al usuario el correo electrónico al principio del test y a partir de este mismo se genera una clave hash (cadena de caracteres de longitud finita) mediante el algoritmo md5 que permite la creación de un identificador totalmente anónimo a partir del cual no se puede volver a la secuencia original (en este caso el correo electrónico) que lo generó. Así finalmente se obtiene una base de datos estructurada que se puede ir actualizando a medidas que más y más personas responden el cuestionario. A continuación se presenta un ejemplo de cómo sería un registro:

ID	Sexo	Género	Edad	Mano dominante	Estudios	Origen
asdaA565bz	Mujer	Hombre	34	Diestro	Doctorado	España

Residencia	Tiempo en residencia	Puntuación total	F1	F2	F3
España	0	43	16	18	9

Clase social	Número de hermanos	Orden nacimiento	Origen étnico	Profesión
Media	3	3	Caucásico	Profesor

Tabla 4. Ejemplo de registro de la base de datos estructurada.

[8] De conformidad con lo dispuesto en el artículo 4.11 del Reglamento (UE) 2016/679, se entiende por consentimiento del afectado toda manifestación de voluntad libre, específica, informada e inequívoca por la que este acepta, ya sea mediante una declaración o una clara acción afirmativa, el tratamiento de datos personales que le conciernen.

[9] Se considera lícito el uso de datos personales seudonimizados con fines de investigación en salud. (Disposición adicional decimoséptima de la Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos y Garantía de los Derechos Digitales.

7. DESARROLLO DEL MODELO

A la hora de conducir la investigación en este trabajo se plantea un desarrollo por fases, construyéndose unas sobre otras buscando lograr lo antes posible un prototipo funcional que posteriormente se irá revisando y mejorando siempre adaptándose al tiempo disponible para el proyecto de Trabajo Fin de Máster y la exigencia del mismo.

7.1. CRITERIOS DE VALIDACIÓN

En primer lugar es necesario fijar una serie de criterios por los que medir la calidad del ajuste realizado por los modelos. Tal y como se explicó en el apartado de bases de datos las muestras se encuentran balanceadas luego bastará con emplear como criterio de calidad la proporción de aciertos (accuracy). En otros casos en los que la muestra se encuentre desbalanceada podría ser necesario el uso de coeficientes como el recall o la precisión. A la hora de validar los modelos se presentan situaciones diferentes para las dos bases de datos:

- La base de datos FER ya tiene sus muestras etiquetadas según sean para training o para test. 28.709 fotografía se dedican al entrenamiento dejando las 3589 restantes para validación.

- La base de datos RafD no indica qué individuos emplear para entrenamiento y cuáles para validación. En este trabajo se han elegido de manera arbitraria cinco individuos para la validación siendo las imágenes de estos eliminadas del conjunto de entrenamiento. El conjunto de validación se compone por las imágenes desde tres ángulos con tres miradas de todas las emociones de cada uno de los cinco individuos, es decir, de un total de 360 imágenes. A la hora de esta selección se ha buscado capturar la mayor variedad posible con el fin de inferir si realmente el modelo ha aprendido de manera transversal y no identifica emociones, por ejemplo, solo para varones adultos. Los individuos seleccionados para la validación han sido:

Figura 13. Individuos empleados para la validación en RafD.

Una vez definidos los criterios de validación que se emplearán para medir la calidad de los modelos se procede a la creación del primer modelo de Deep Learning. Para el entrenamiento de este primer modelo se emplea la base de datos FER 2013 (explicada en el capítulo anterior). En un primer lugar se plantea una arquitectura relativamente sencilla proponiendo una red neuronal con tres capas convolucionales (con 64, 128 y 256 neuronas respectivamente) y una capa densa final. Los resultados de esta primera red son bastante desoladores, logrando en el mejor de los casos una tasa de acierto del 22% tras un proceso de ajuste de parámetros.

Tras leer algunos artículos sobre la materia se plantea una nueva arquitectura denominada mobile net que presenta muy buenos resultados en distintas áreas. La información sobre esta red se encuentra recogida en el artículo “Learning Transferable Architectures for Scalable Image Recognition” (Barret Zoph, Vijay Vasudevan, Jonathon Shlens, Quoc V. Le). En esta memoria no se profundiza en la arquitectura de la red por su complejidad técnica. Esta red devuelve unos resultados mucho más halagüeños, tras un proceso de mejora de parámetros basada en los propuestos por Barret Zoph se llega a alcanzar un 99% de acierto en entrenamiento que se traslada a una tasa de acierto del 93% en test.

Una vez realizada esta primera construcción del modelo se abren dos líneas de trabajo. Por una parte se desarrolla una página web para la aplicación del TAS-20 que generará nuestra base de datos estructurada antes mencionada. Por otra se comienza a diseñar la aplicación web destinada a detectar las emociones de los usuarios; nuestro espejo virtual.

7.2 ESPEJO VIRTUAL

Para la construcción del espejo virtual se sigue la siguiente hoja de ruta:

- Construcción de un página web mediante html y Javascript que permita al usuario visualizar su imagen (previa concesión del acceso a la webcam)
- Valoración de distintas puestas en producción para introducir el modelo.
- Carga del modelo previmente implementado.
- Testeo en individuos y con entornos (fondo liso, fondo real, distintas intensidades de luz, sombras…)

Tras esto comienza el proceso de validación observando que los resultados obtenidos son bastante pobres; en la mayoría de los casos el modelo no clasifica correctamente la emoción. Esto se produce probablemente debido a la naturaleza de las fotos empleadas para el entrenamiento (fotos 28 x 28 en blanco y negro con la persona centrada). El modelo es entrenado trabajando con fotos de baja resolución en blanco y negro y a la hora de su puesta en producción se enfrenta a fotos de alta definición, en color, con fondo no uniforme… Es necesario reentrenar un modelo con una base más realista. Esta parte pone fin a la primera vuelta al proceso constructivo.

Hasta ahora se ha diseñado una aplicación capaz de emplear un modelo implementado en TensorFlow para la clasificación de emociones.

El siguiente paso consiste en buscar una nueva base de datos en la que las imágenes sean más cercanas a aquellas a las que la aplicación se va a enfrentar para ello se procede a investigar las bases de datos disponibles (es complicado encontrar este tipo de bases de datos por su peculiaridad a la hora del etiquetado, es necesario etiquetarlas una a una y por varios expertos pues en múltiples casos las emociones son mixtas y se entremezclan siendo complicado distinguirlas incluso para el observador humano). En el Anexo I se puede observar un listado de bases de datos asociadas a dicho tema.

Finalmente, se selecciona la base RaFD por contener estas imágenes más próximas en formato (mayor tamaño, en color, desde diferentes perspectivas…) y más diversas (contempla desde niños a ancianos, de distintos géneros, etnias y complexiones). El entrenamiento del modelo con esta nueva base de datos requiere un cierto preprocesado técnico debido a que mientras que la base inicial era un csv en el que se encontraban los valores asociados a cada píxel de la imagen en este caso la base de datos es en sí un conjunto de imágenes cuyos píxeles deben leerse y preprocesarse y de las cuales las emociones deben extraerse del título mediante el empleo de expresiones regulares para proceder a la construcción de las etiquetas. Además la red debe adaptarse pues en este caso se trabaja con ocho emociones en lugar de siete. La modularidad del código facilita algunos de los cambios como el del número de emociones regulado mediante un parámetro. A la hora de preprocesar las imágenes sin embargo en este caso se realiza un preprocesado mucho más exhaustivo que en el caso de la primera base. Las imágenes se transforman en primer lugar a escala de grises pues el color no aporta información a la hora de detectar una emoción, además se realiza un pequeño recorte para reducir el espacio vacío (fondo) que puede llevar a error al modelo. Por último se lleva a cabo un proceso de regularización modificando las fotos ligeramente para que no sean fotos de estudio sino algo un poco más realista (se introduce ruido blanco, se tuercen ligeramente, se aplican simetrías…). Todo este preprocesado se diseña de manera genérica y modular para que pueda ser adaptado a nuevas bases de datos en un futuro.

8. PUESTA EN PRODUCCIÓN

Una vez entrenado y validado el modelo se procede a la puesta en producción. Esta ha sido una de las etapas más complicadas de todo el proyecto debido a que TensorFlow es un framework aun en fase de crecimiento. Esto conlleva que algunos métodos resultan imposible de poner en producción. Entre estos método se encuentra la batchnormalization y el drop out presentes en la mobilenet motivo por el cual aunque estas redes presenten mejores resultados (en términos de accuracy) son descartadas para la puesta en producción por el momento eligiéndose finalmente para la puesta en producción la Alexnet. Los modelos son aun así presentados en esta memoria y a medida que TensorFlow avance e incluya los métodos antes citados se podría proceder a la puesta en producción con estos modelos que en un principio plantean resultados notoriamente mejores.

A la hora de la puesta en producción se han planteado dos vías: TensorFlow Serving y conversión a TensorFlow Javascript.

8.1 TENSORFLOW SERVING

A la hora de emplear el modelo usando TensorFlow Serving el proceso a seguir es el siguiente:

- Exportar el modelo generando un fichero en el que se almacenan entre otras cosas el grafo subyacente al modelo y los pesos asignados a cada nodo.
- Activar en el Docker los puertos necesarios para servir el modelo y proceder a su carga.
- Una vez hecho esto, el modelo será capaz de recibir fotografías nuevas, procesarlas y devolver un array de probabilidades indicando cada una de estas la asociación con cada emoción.

Un ejemplo de todo este proceso incluyendo las líneas de código necesarias para ello se encuentra en el notebook anejo (ver Anexo II). En el se presenta un proceso GRPC en el que se observa cómo se podría emplear este modelo para la predicción a partir de fotos dentro de un notebook.

Para proceder a la puesta en producción dentro de la aplicación web se emplean peticiones Ajax. Sin entrar en los detalles más técnicos, las peticiones Ajax son aquellas que permiten acceder a informaciones (en este caso el modelo) que se encuentran en un servidor web. Este método resulta algo más lento que el que se presentará a continuación por lo que se descarta para la página web aunque resulta interesante (sobre todo a nivel didáctico) su uso en el notebook.

8.2 TENSORFLOW JAVASCRIPT

Esta es actualmente la manera más eficiente para servir modelos en aplicaciones web. La gran dificultad de esta puesta en producción reside en los problemas para transformar un modelo en TensorFlow a un modelo en TensorFlow Javascript debido a la incompatibilidad entre versiones. Además esta técnica es principalmente usada por la comunidad de Deep Learning para transformación de modelos en Keras en lugar de modelos escrito sobre TensorFlow por lo que muchas de las dificultades que esta idea supone no han sido aun atacadas por el equipo de Google.

Una vez transformado el modelo de TensorFlow a TensorFlow Javascript este se sirve mediante código Javascript estando ya listo para ser usado en la web.

9. RESULTADOS

Una vez explicado todo el proceso llevado a cabo durante el proyecto es el momento de presentar los distintos resultados obtenidos. En este caso se entienden por resultados los modelos entrenados que permiten el reconocimiento de emociones como las dos aplicaciones resultantes: el TAS 20 y el *Prolexitim Mirror Web* (espejo virtual).

9.1. MODELOS

A continuación se presentan los resultados asociados a los modelos diseñados. Para el FER se presentan tan solo los resultados para la Mobilenet. (Recordemos que al pasar a producción una vez entrenado el Mobile con FER y RafD fue cuando se observaron las limitaciones de este modelo para la puesta en producción y fue sustituido por la Alexnet).

El modelo Mobile presenta para FER una tasa de acierto en la validación del 93%.

	Ira	Asco	Miedo	Alegría	Neutral	Tristeza	Sorpresa
Ira	4784	97	0	0	0	0	72
Asco	11	160	266	0	0	0	110
Miedo	0	223	4836	6	0	0	56
Alegría	0	0	0	8989	0	0	0
Neutral	0	52	59	321	5343	423	0
Tristeza	0	236	321	0	45	5475	0
Sorpresa	0	0	92	123	0	0	3787

Tabla 5. Matriz de confusión de Mobilenet sobre FER 2013.

En la siguiente tabla se presentan la precisión y el recall asociado a cada clase:

Clase	Precisión	Recall
Ira	0,998	0,966
Asco	0,208	0,293
Miedo	0,868	0,944
Alegría	0,953	1
Neutral	0,992	0,862
Tristeza	0,928	0,901
Sorpresa	0,941	0,946

El modelo Mobile devuelve resultados bastante malos cuando se entrena sobre la base de datos RafD esto se debe muy posiblemente a que esta base de datos es de un tamaño mucho menor. Esto conduce a que el modelo tienda a estancarse en una sola clase (en este caso la neutralidad). La matriz de confusión obtenida tras la obtención del modelo se presenta a continuación:

	Ira	Desprecio	Asco	Miedo	Alegría	Neutralidad	Tristeza	Sorpresa
Ira	0	0	0	0	0	45	0	0
Desprecio	0	0	0	0	0	45	0	0
Asco	0	0	0	0	0	45	0	0
Miedo	0	0	0	0	0	45	0	0
Alegría	0	0	0	0	0	45	0	0
Neutralidad	0	0	0	0	0	45	0	0
Tristeza	0	0	0	0	0	45	0	0
Sorpresa	0	0	0	0	0	45	0	0

Tabla 7. Matriz de confusión de Mobilenet sobre RafD.

Con una tasa de acierto de 12,5%. La neutralidad presenta un recall asociado de 1 pero la precisión es de 0.125. El resto de variables tienen el recall y la precisión a 0.

Por último el modelo Alex es el que mejores resultado ha devuelto sobre la base de datos RafD y además el que no plantea problemas para la puesta en producción. Tras ser entrenado reporta una tasa de acierto de 99% (un resultado increíble y reproducible). La matriz de confusión muestra que solo presenta un error en el que confunde tristeza con sorpresa:

	Ira	Desprecio	Asco	Miedo	Alegría	Neutralidad	Tristeza	Sorpresa
Ira	45	0	0	0	0	0	0	0
Desprecio	0	45	0	0	0	0	0	0
Asco	0	0	45	0	0	0	0	0
Miedo	0	0	0	45	0	0	0	0
Alegría	0	0	0	0	45	0	0	0
Neutralidad	0	0	0	0	0	45	0	0
Tristeza	0	0	0	0	0	0	44	1
Sorpresa	0	0	0	0	0	0	1	44

Tabla 8. Matriz de confusión de Alexnet sobre RafD.

La precisión y el recall asociados a la ira, el desprecio, el asco, el miedo, la alegría y la neutralidad es de 1. En cuanto a la tristeza y la sorpresa tienen una precisión asociada de 0.978 y un recall de 0.978.

9.2. TAS 20

Esta herramientas se encuentra disponible en el siguiente enlace (http:// www.serendeepia.com/prolexitim_tas20/). Se reproduce a continuación brevemente la experiencia de usuario:

La página ofrece un mensaje de bienvenida y pide el correo electrónico. La página está diseñada para no permitir acceder al cuestionario sin rellenar este campo pues es necesario para la generación del código identificativo:

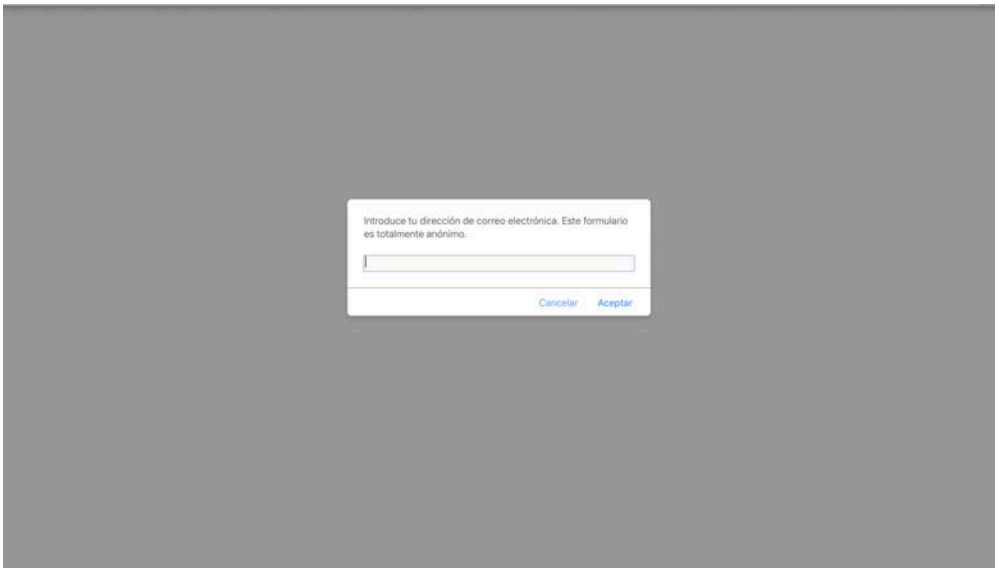


Figura 14. Petición de dirección de correo por parte de la página.

Una vez introducido el correo aparecen las preguntas (*Figura 15*). Tras rellenar las 20 preguntas el botón comprobar permite al usuario obtener sus resultados (*Figura 17*). Si alguna pregunta estuviera sin contestar se le avisaría por pantalla (*Figura 16*):

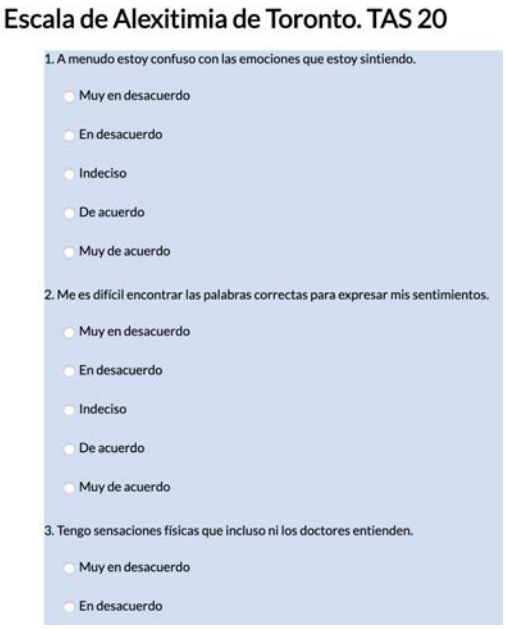


Figura 15. Ejemplo de preguntas del test.

Figura 16. El programa avisa si no se han respondido todas las preguntas.

Figura 17. Resultado presentado por pantalla al usuario.

Por último el programa pide al usuario que conteste una serie de preguntas sociodemográficas. De nuevo comprueba que todas estén respondidas y de ser así se activa el botón de enviar quedando registrada la información e incorporándose a la base de datos estructurada antes mencionada.

9.3. TENSORFLOW SERVING

El código necesario para cargar el modelo así como las instrucciones para hacerlo (configuración del docker, pasos para la exportación) se encuentran recogidas en un notebook de Jupyter en el Anexo X. Este cuaderno se encuentra igualmente presente en el repositorio de GitHub asociado a este trabajo.

9.4. PROLEXITIM VIRTUAL MIRROR

La aplicación se encuentra disponible en el siguiente enlace (http://www.serendeepia.com/prolexitim_mirror_web/)[10] y básicamente capta la expresión del usuario y la muestra por pantalla es totalmente automática e intuitiva:

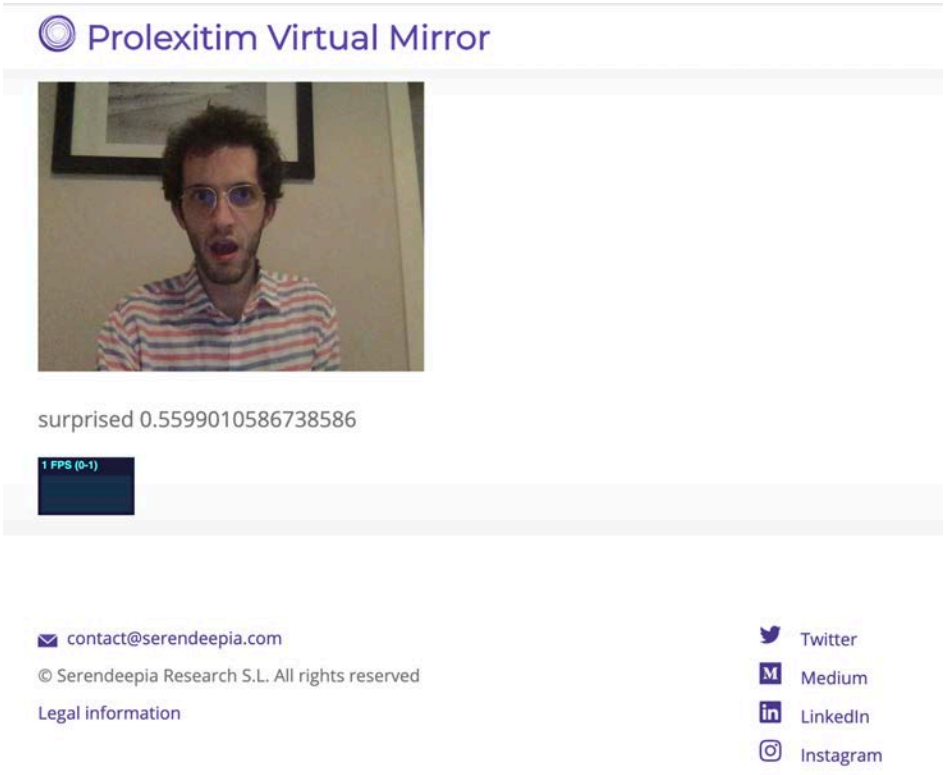


Figura 18. Ejemplo de la interfaz del espejo virtual.

[10] **Disclaimer.** El enlace y el servidor en el que corren pertenece a Serendeepia Research SL por lo que podría sufrir modificaciones o indisponibilidad temporal durante el tiempo. En el repositorio adjunto se encuentra todo el código de la web que puede ser ejecutado localmente mediante Jekyll. La aplicación en local funcionará y no sufrirá modificaciones. En cualquier caso en la defensa de este trabajo se realizará una demostración práctica de uso.

10. CONCLUSIONES

Esta memoria recoge todo el trabajo realizado en el desarrollo del proyecto. El principal objetivo de éste era el desarrollo de un modelo de clasificación de redes neuronales capaz de reconocer emociones en rostros. Este objetivo ha sido logrado con creces presentando resultados excelentes en la base de datos FER 2013 y de una calidad nada desdeñable en RafD.

A la hora de desarrollar la aplicación la puesta en producción en sus diferentes variables funciona pero el modelo no tiene suficiente capacidad generalizadora para ofrecer una performance adecuada. Es un buen resultado como prototipo pero debe mejorarse.

La gran dificultad para mejorar el modelo es la obtención de una base de datos suficientemente grande y transversal (distintas edades, géneros, etnias…) en la que se hayan etiquetado las emociones. En el Anexo I se presenta una lista de las principales bases de datos y se observa que aquellas de mayor tamaño suelen ser de pago. Una inversión en dichas bases de datos podría ser una solución para mejorar el modelo.

El trabajo se complementa con el diseño de la aplicación online para el test TAS-20 que permite por una parte realizar esta prueba de manera gratuita, rápida y cómoda así como construir una base de datos con la que estudiar más a fondo este trastorno. Dicho test será el elemento de control para medir la mejora que el entrenamiento empleando el *Prolexitim Mirror* pudiera suponer en los pacientes.

BIBLIOGRAFÍA

Alhanai, T., Ghassemi, M. Glass, J. (2018). Detecting Depression with Audio/Text Sequence Modeling of Interviews.

Bagby, R. M.; Parker, J. D.; Taylor, G. J.: The twenty-item Toronto Alexithymia Scale-I. Item selection and cross-validation of the factor structure. Journal of Psychosomatic Research 1994; 38(1):23–32.

Belmonte J. M. (2015). Comparación de dos métodos de escritura de historia clínica electrónica. Tesis Doctoral. Universidad Complutense de Madrid.

Chollet François. 2017. Deep Learning with Python (1st ed.). Manning Publications Co., Greenwich, CT, USA.

Dana H. Ballard; Christopher M. Brown (1982). Computer Vision. Prentice Hall.

Ekman P. (1992). An Argument for Basic Emotions. Cognition and Emotion. Pages 169-200

Esteva A, Robicquet A, et al. A guide to deep learning in healthcare. Nature 25, 24-29 (2019).

Feng H., Tang H., Tsema A., Zhu Z. (2018). Computer Vision for Sight: Computer Vision Techniques to Assist Visually Impaired People to Navigate in an Indoor Environment. Academic Press.

InfoWorld (2019) InfoWorld. What is TensorFlow? The Machine Learning library explained disponible en <https://www.infoworld.com/article/3278008/what-is-tensorflow-the-machine-learning-library-explained.html> [23 mayo 2019]

Jiang F, Jiang Y, Zhi H, et al. Artificial intelligence in healthcare: past, present and future Stroke and Vascular Neurology (2017).

Langner O., Dotsch R., Bijlstra G., Wigboldus D. H. J. (2010). Presentation and Validation of the Radboud Faces Database.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444.

Liebeskind DS. Artificial intelligence in stroke care: Deep learning or superficial insight? EBioMedicine. ;35:14–15 (2018).

Mitchell, T. (1997). Machine Learning, McGraw Hill.

Murray, H. (1973). The Analysis of Fantasy. Huntington, NY: Robert E. Krieger Publishing Company.

Nemiah J, C: Psychology and Psychosomatic Illness: Reflections on Theory and Research Methodology. Psychother Psychosom 1973;22:106-111.

Páez, D., Martínez-Sánchez, F. (1999). Validez psicométrica de la escala de Alexitimia de Toronto (TAS-20). Un estudio transcultural. Boletín de Psicología. 63.

Parker, J.D., Bagby, R.M., Taylor, G.J., Endler, N.S. & Schmitz, P. (1993). Factorial validity of the 20-item Toronto Alexithymia Scale. European Journal of Personality, 7, 221-232.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. Neural Networks, 61, 85–117.

Scimen, Bilal & Atasoy, Huseyin & Kutlu, Yakup & Yildirim, Serdar & Yildirim, Esen. Smart Robot Arm Motion Using Computer Vision. Elektronika ir Elektrotechnika. 21. 3-7. (2015).

Taylor G., Bagby R. and Parker J. (1997). Disorders of Affect Regulation: Alexithymia in Medical and Psychiatric Illness(Paperback edition 1999). Cambridge University Press.

Zoph B., Vasudevan V., Shlens J., Quoc V. (2017) Learning Transferable Architectures for Scalable Image Recognition. Cornell University.

ANEXO I. BASES DE DATOS RELACIONADAS CON DETECCIÓN DE EMOCIONES ETIQUETADA

A continuación se presentan una serie de bases de datos para reconocimiento de emociones mediante la expresión facial. En esta lista existen bases de datos en la que las imágenes se toman de manera espontánea (reflejo más natural de las emociones) y otras en las que se pide a los participantes que muestren una emoción (menos natural, más intensidad en los rasgos y mayor duración si se trata de vídeos). En negrita aparecen las bases de datos que he considerado más relevantes y en cursiva un pequeño resumen de los puntos de interés de cada base de datos.

- **Aff - Wild.** (Solicitud de descarga) Imágenes extraídas de vídeos de Youtube con un total de más de 3 millones de imágenes. Está anotado por experto respecto a distintas expresiones faciales y una parte incluye emociones etiquetadas por un experto.

(Muy interesante por su gran tamaño y por su espontaneidad. En mi opinión la mejor opción)

- **Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)** (Descargar) Esta base de datos contiene 7356 archivos (24.8 GB). En ellos 24 actores (12 actrices y 12 actores) leen distintas frases transmitiendo calma, felicidad, tristeza, enfado, miedo, sorpresa y asco. Así mismo se cantan canciones que transmiten calma, felicidad, tristeza, enfado o miedo. Cada expresión se produce a dos niveles (normal e intenso) incluyendo además una versión neutral.

  https://www.youtube.com/watch?v=XQkmH4oYZkg

  (*Las imágenes se pueden extraer a partir de fotogramas del vídeo. Cabe suponer que las emociones se reflejan más cuando cantan (tanto en sonido como en expresión facial) por lo que quizá sería un análisis interesante*).

- F-M FACS 3.0 (EDU, PRO & XYZ versions). (Hay que pagar en torno a 1000€ por la base de datos, ver enlace) Basado en distintos movimientos de la cara y la cabeza realizados por diez actores. Consta tanto de vídeos como de imágenes (más de 4877 vídeos). Está compuesta tanto por imágenes premeditadas (posado) como espontáneas. Las emociones recogidas son neutralidad, tristeza, sorpresa, felicidad, miedo, enfadado, desprecio y asco.

(En el siguiente enlace se muestran mediante GIFS las ideas recogidas. Parece interesante porque no solo entiende expresiones fijas si no también movimientos (giro de ojos, de cabeza, movimientos de ceja) que parecen esenciales para la comprensión del lenguaje facial. Además están etiquetadas tanto las emociones como el rasgo que se produce (ejemplo, encaramiento de ceja) y la intensidad con la que dicho rasgo se produce).

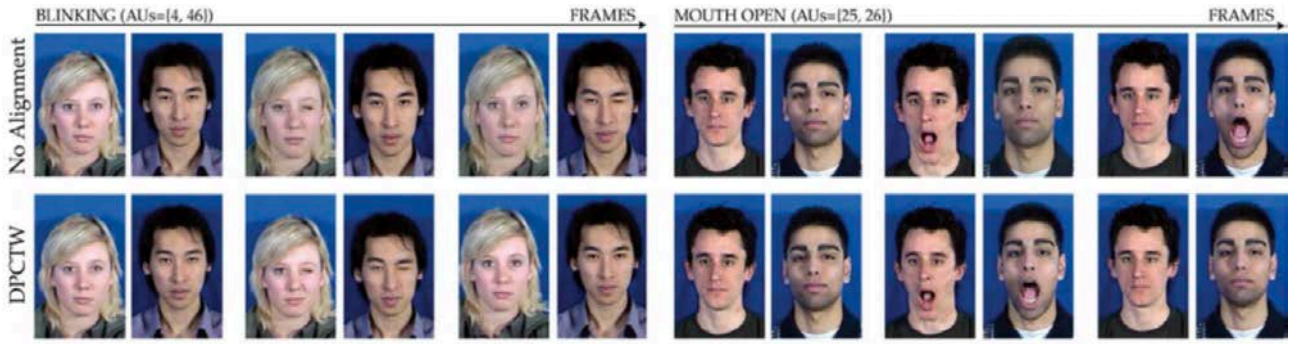
- **Extended Cohn-Kanade Dataset (CK+)**. (Enlace para descarga) Comportamiento de 210 adultos entre 18 y 50 años. 69% muejeres, 81%, caucásicos, 13% afroamericanos, and 6% otros. 700 secuencias. De las cuales 327 están etiquetadas como emociones discretas. La duración de la secuencia varía (de 10 a 60 frames) y las expresiones siempre acaban y terminan en neutralidad. Las emociones etiquetadas son neutralidad, tristeza, sorpresa, felicidad, miedo, enfado, desprecio y asco. Todas estas son forzadas (poses) pero además se registran 122 sonrisas de 66 sujetos que se consideran espontáneas (los sujetos no saben que hay una cámara).

(Parece una de las mejores opciones, tiene variedad en los sujetos y en sus razas. Las secuencias están etiquetadas de manera discreta y aunque en su mayoría son expresiones posadas también existen algunas espontáneas).

- Japanese Female Facial Expressions (JAFFE). (Enlace para la solicitud de descarga) 213 imágenes fijas (no de vídeo) de diez mujeres japonesas posando. Etiquetadas según transmiten tristeza, neutralidad, sorpresa, felicidad, miedo, enfado o asco.

(Esto puede resultar interesante como base de datos de comprobación para el prototipo, para descubrir si realmente los rasgos raciales de la persona influyen en la detección de la emoción).

- MMI Database. (Solicitud de descarga) 43 sujetos grabados etiquetándose los fotogramas según los diferentes gestos de su cara.



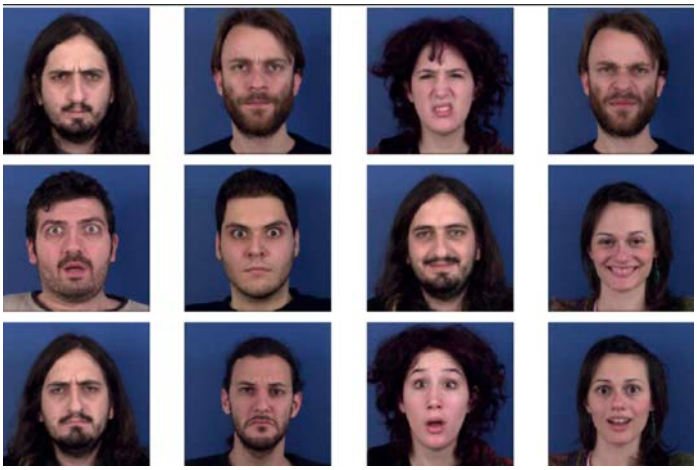
(Pega: No tiene las emociones per sé etiquetadas).

- **Belfast Database.** (Enlace para el acceso a la base de datos). Compuesta a su vez por tres bases de datos emplea en total 356 sujetos que se someten a distintos experimentos que les llevan a mostrar de manera "natural" asco, miedo, diversión, frustración, sorpresa, enfado y tristeza. Todo ello se encuentra recogido en más de 1300 vídeos cortos (5 a 6 segundos) que disponen de sonido.

  No he encontrado ninguna imagen de muestra.

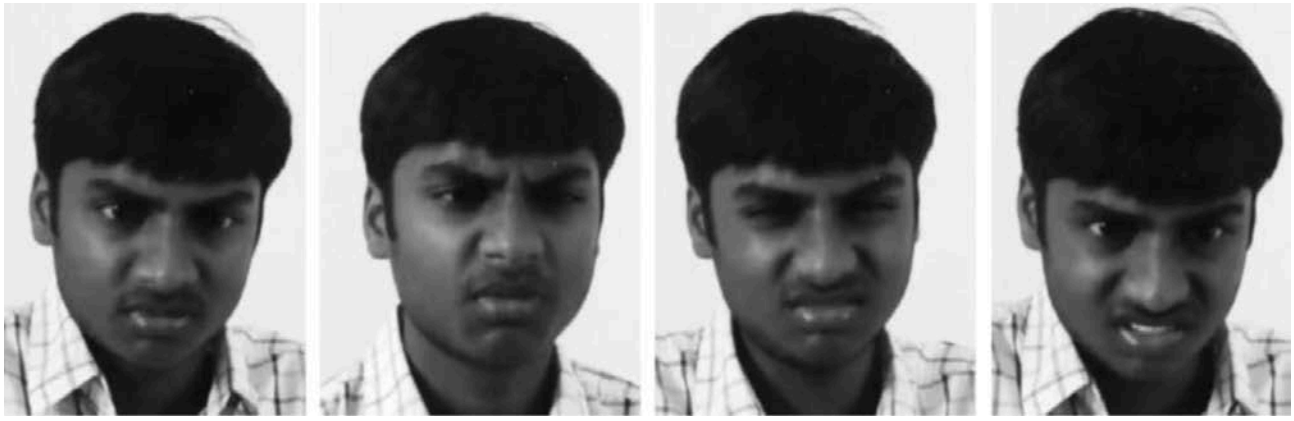
  (Parece de las más interesantes, son emociones naturales y se dispone tanto de vídeo como de sonido. La desventaja frente a otras es que los movimientos no están etiquetados aunque esto no parece muy relevante).

- La base de datos **MUG** (enlace para la solicitud de la descarga) busca solventar algunos problemas que presentan otras bases de datos empleando alta resolución, luz uniforme, muchos sujetos y muchas tomas de sujetos. Consiste en secuencias de imágenes realizando expresiones faciales (38 GB). En ella participican 35 mujeres y 51 hombres todo de origen caucásicos entre 20 y 45 años. Dataset en dos partes en el primero expresiones forzadas mostrando enfado, asco, miedo, felicidad, tristeza y sorpresa (está etiqueteda): La segunda son emociones inducidas (mediante un vídeo) pero no está etiquetada.



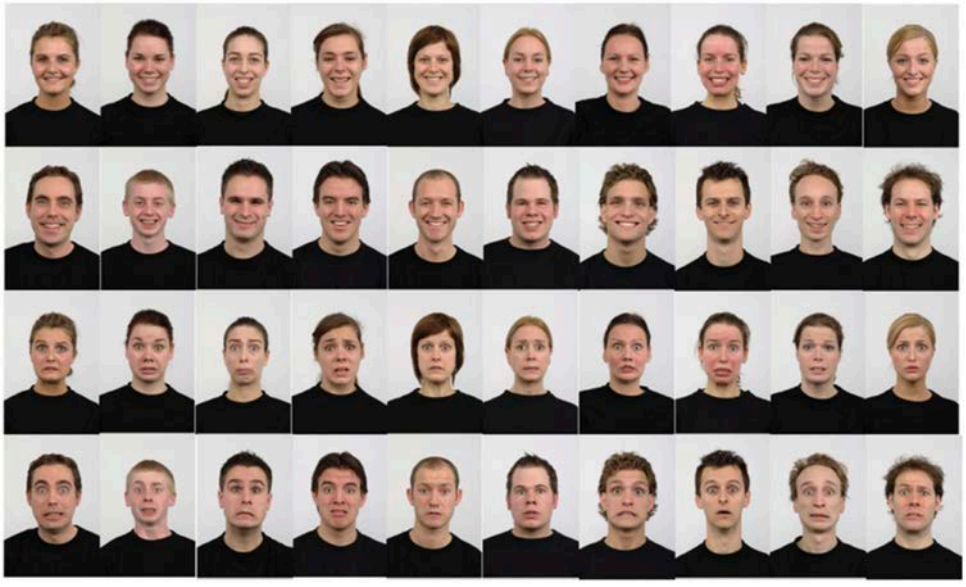
(*De nuevo tiene la pega de que solo recoge sujetos caucásicos, además la parte donde las emociones son inducidas no está etiquetada. A cambio tiene la ventaja de la nitidez en temas de resolución e iluminación.*)

- **Indian Spontaneous Expression Database**. (Enlace para la solicitud de la descarga) En este experimento participan 50 sujetos que graban más de 428 vídeos en los que se muestra su reacción a vídeos que buscan inducir distintas emociones. Posteriormente los vídeos son etiquetados según cinco emociones básicas: tristeza, sorpresa, felicidad y asco.



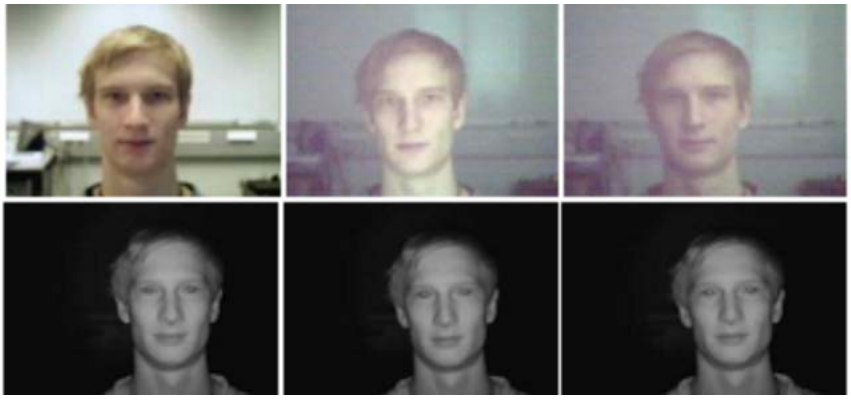
(*Parece bastante interesante porque son emociones relativamente naturales y están etiquetadas la mayor pega al respecto es la falta de variedad étnica*)

- The **Radboud Faces Database (RaFD)** (enlace para la solicitud de descarga) es un conjunto de fotografía de 67 modelos (incluyendo hombres, mujeres y niños caucásicos y hombre marroquí-holandeses) mostrando ocho emociones, miedo, asco, enfado, felicidad, tristeza, sorpresa, desprecio y neutral. Cada emoción se muestra mirando en tres direcciones y se fotografía desde cinco ángulos simultáneamente. Las emociones están etiquetadas.



(*Lo interesante de esta base de datos es que recoge las expresiones de todo el espectro de edad e induce una cierta variedad étnica. Además incluye la visión de los individuos desde distintos ángulos*)

- **Oulu-CASIA NIR&VIS** (solicitud de descarga) contiene vídeos con las seis expresiones faciales más típicas (felicidad, tristeza, sorpresa, miedo, enfado y asco) de 80 individuos entre 23 y 58 años siendo un 73.8% hombres. Los sujetos estudiados son de origen chino o finlandés y cada expresión facial (forzada) es capturada en tres condiciones de iluminación, normal, débil y oscuro.



(La ventaja de esta base de datos es que estudia la visibilidad según los tipos de luces quizá sería interesante para perfeccionar un modelo entrenado previamente en otras bases de datos)

- Facial Expression Research Group Database (solicitud de descarga) es una base de datos hecha con personajes en dibujo animado en 3D formada por 55767 imágenes frontales etiquetadas de los seis caracteres mostrando enfado, asco, miedo, alegría, neutralidad, tristeza y sorpresa.

(No sé hasta qué punto es útil trabajar con personajes no humanos si no humanoides, ficticios)

- **AffectNet** (enlace para la solicitud de descarga) es una base de datos de expresiones faciales tanto forzadas como espontáneas. Está formada por más de un millón de imágenes recogidas de Internet. 440000 fueron anotadas manualmente según las siete emociones típicas: enfado, asco, miedo, alegría, neutralidad, tristeza y sorpresa. También se valora la intensidad de la emoción (permitiendo desarrollar modelos continuos). Es la mayor base de datos del tema con diferencia.

(Esta base de datos resulta muy interesante por la enorme cantidad de datos de los que dispone así como por la posibilidad de plantear modelos en los que las emociones se modelan de manera continua en lugar de categórica. Además usa distintos anotadores por lo que en los casos en los que podría haber dudas las imágenes tienen más de una categoría asociada)

- IMPA-FACE3D. (Enlace para la descarga) La base de datos almacena imágenes de 38 individuos (22 hombres y 16 mujeres entre 22 y 50 años) con cara neutral y expresiones que muestran las seis emociones típicas: enfado, asco, miedo, alegría, tristeza y sorpresa En este dataset se considera la geometría y el color (geometría y texturas correlacionadas). Así mismo añade expresiones faciales etiquetadas como guiños o imágenes de perfil. Sobre cada individuo se toman 14 muestras habiendo en total 532 imágenes.

(En esta base de datos se trabaja también con color lo que puede resultar interesante)

- FEI Face Database (descarga directa). Imágenes tomadas sobre cien hombres y cien mujeres entre 19 y 40 años. De cada uno se toman 14 fotos (2800 en total). Son imágenes posadas y solo distingue entre sonrisas y neutralidad. Anota ciertos puntos de la cara para un mejor estudio.

  (Parece demasiado simple aunque puede ser útil si se decide construir un modelo binario)

Idea. No sé hasta qué punto sería posible mezclar distintas bases de datos lo que permitiría una mayor variedad a nivel étnico así como combinar emociones más “forzadas” con emociones naturales. Sería interesante saber si las expresiones forzadas son válidas para el entrenamiento (¿hasta qué punto se puede fingir una emoción?) o si es mejor ceñirse únicamente a las expresiones espontáneas.

Anexo II

Serving the model

Este notebook resume el proceso para servir el modelo mediante TensorServing y poder emplear un modelo ya entrenado para hacer predicciones.

El proceso a seguir es el siguiente:

1. Entrenamiento del modelo sobre el conjunto de trainig. Esto devuelve los pesos asociados al modelo.
2. Emplear el método `export` para exportar el modelo empleando como entrada la carpeta de checkpoints obtenida en el paso anterior.
3. Una vez obtenido el modelo se emplea docker para habilitar los puertos y se procede a la carga del modelo.
4. Una vez activado el modelo solo es necesario introducirle fotografías para obtener las respuestas deseadas.

Para activar docker:

export MODEL_NAME=emotions_mirror export MODEL_BASE_PATH=/models docker run -p 8500:8500 -p 8501:8501 -v "/Users/arturosanchezpalacio/Documents/Serendeepia/$MODEL_NAME:$MODEL_BASE_PATH/$MODEL_NAME" -e MODEL_NAME=$MODEL_NAME -e MODEL_BASE_PATH=$MODEL_BASE_PATH -t tensorflow/serving:latest

Una vez hecho esto se carga la imagen (o imágenes, pudiendo emplearse una lista de ser más de una) deseada:

In [3]:

```
images = ['./triste.jpg']
```

La imagen se pasa al modelo que devuelve un vector de probabilidades asociada a cada clase:

import sys import requests import json import base64 from PIL import Image import numpy as np data = [] for image in images: img = Image.open(image) img_data = np.array(img, dtype=np.uint8).tolist() data.append(img_data) payload = { 'inputs': img_data } host = 'localhost' port = 8501 model_name = 'emotions_mirror' url = 'http://{}:{}/v1/models/{}:predict'.format(host, port, model_name) print(url) r = requests.post(url, json=payload) print(r.content) result = json.loads(r.content)

Las emociones vienen codificadas por:

- 0 Enfadado
- 1 Asco
- 2 Miedo
- 3 Feliz
- 4 Triste
- 5 Sorpresa
- 6 Neutral

Por lo que mediante un diccionario podemos obtener la emoción más probables:

In [40]:

```
dic_emotions = {0:'Enfado', 1:'Asco', 2:'Miedo', 3:'Feliz', 4:'Triste', 5:'Sorpresa', 6:'Neutral'}
```

In [36]:

```
predictions = result['outputs']['prediction'][0]
```

In [41]:

```
dic_emotions[predictions.index(max(predictions))]
```

Out[41]:

```
'Triste'
```

ANEXO III

El código de este trabajo se encuentra en el siguiente repositorio de GitHub en la rama 31 de Mayo (donde no se realizarán cambios a partir de la entrega de este trabajo).

La carpeta *TAS20* contiene el código asociado a la aplicación para la realización de dicho test online.

La carpeta *emotions_mirror* contiene los distintos modelos de Deep Learning.

La carpeta *prolexitim_mirror_web* contiene el código para la generación de la aplicación web.

Por último se presenta el notebook recogido en el Anexo II.