UNIVERSIDAD DE SANTIAGO DE CHILE
FACULTAD DE INGENIERÍA
Departamento de Ingeniería Informática

Modelo de reconocimiento automático de rasgos de personalidad para organizaciones que trabajan con voluntarios

Felipe Maturana Guerra

Profesor guía: Carolina Bonacic Castro

Trabajo de titulación en conformidad a los requisitos para obtener el título de Ingeniero Civil en Informática, y Magíster en Ingeniería Informática

Santiago – Chile
2020

RESUMEN

En el presente proyecto se desarrolla un modelo de reconocimiento de rasgos de personalidad en perfil voluntarios de manera automática considerando la clasificación *BigFive* mediante un modelo de aprendizaje de máquinas que analiza textos libres de los voluntarios, de esta forma se obtendrá mayor información acerca de los voluntarios postulantes y se apoyará la toma de decisiones de organizaciones que trabajen con voluntarios.

De acuerdo al estado del arte, la detección automática de rasgos de la personalidad ha sido un tema de interés durante los últimos años por lo que se han desarrollado modelos que buscan resolver esta tarea de manera automática y han disponibilizado los datos de entrenamiento y etiquetado utilizados lo cual sirve para entrenar, enriquecer y decidir cuál es el mejor modelo a utilizar para el contexto del voluntario que asiste a actividades de trabajos voluntarios en instituciones chilenas.

Para ello se desarrolla una metodología separada en dos partes compuestas por una primera etapa de preparación, donde se define el modelo de red neuronal a trabajar junto a recopilación de la información de las encuestas y la puesta en marcha de la red neuronal con el entrenamiento con los conjuntos de datos existentes.

En la segunda parte de la metodología se encuentra centrada en la red construida anteriormente con la inclusión de los datos para entrenar, probar y validar el modelo a utilizar con los datos obtenidos en la encuesta para la detección automática de los rasgos de personalidad junto al análisis y comparación de los resultados obtenidos.

No se encuentran diferencias significativas en los resultados obtenidos utilizando esta técnica para distintos idiomas como lo son el español e inglés utilizando una correcta representación numérica de las palabras a través de la técnica de incrustación de palabras y un buen porcentaje de cobertura.

Existe un gran impacto en el desempeño de la detección automática de personalidad de acuerdo al tamaño del lote escogido ya que este determina la cantidad de veces que los parámetros de la red son actualizados por cada época y la cantidad de memoria utilizada. A mayor tamaño del lote: menor cantidad de actualizaciones por época y mayor utilización de memoria, sin embargo el desempeño mejora con tamaños de lotes grandes otorgando mayor estabilidad a la red.

Cada rasgo de la personalidad posee su propio grado de error en la detección automática siendo la extroversión el grado con mayor error absoluto medio de 0.27817 para el subconjunto de voluntarios, mientras que el grado de la apertura al conocimiento el de menor error asociado con un error absoluto medio de 0.23612.

Los resultados obtenidos se encuentran dentro del rango de los trabajos presentes en el estado del arte, sin embargo, la técnica utilizada no muestra un aumento en la precisión de los métodos de reconocimiento automático de rasgos de la personalidad en comparación de la línea base, aún así, conocer más acerca de la personalidad de los participantes mejora la información que se posee acerca de ellos para así apoyar la toma de decisiones de las entidades que trabajen con voluntarios.

Palabras Claves: Voluntario; *Big Five*; Personalidades; Detección Automática;Texto Libre; Análisis; Modelo;

ABSTRACT

In this project, a model of automatic recognition of personality traits in volunteer profiles is developed considering BigFive classification through a machine learning model that analyzes free texts of the volunteers, in this way more information will be obtained about the volunteer applicants and the decision-making of organizations that work with volunteers will be supported.

According to the state of the art, the automatic detection of personality traits has been a topic of interest in recent years, which is why models have been developed that seek to solve this task automatically and have made the training and labeling data used available. which serves to train, enrich and decide which is the best model to use for the context of the volunteer who attends volunteer work activities in Chilean institutions.

For this, a separate methodology is developed in two parts, composed of a first stage of preparation where the neural network model is defined to work together with the collection of information from the surveys and the training of the network with the existing data sets.

In the second part of the methodology, it is focused on the previously built network with the inclusion of data to train, test and validate the model to be used with the data obtained in the survey for the automatic detection of personality traits together with the analysis. and comparison of the results obtained.

No significant differences were found in the results obtained using this technique for different languages such as Spanish and English using a correct numerical representation of the words through the word embedding technique and a good percentage of coverage.

There is a great impact on the performance of the automatic personality detection according to the chosen batch size since it determines the number of times that the network parameters are updated for each epoch and the amount of memory used. The larger the batch size: fewer updates per epoch and higher memory utilization, however performance improves with large batch sizes, providing greater stability to the network.

Each personality trait has its own degree of error in automatic detection, with extraversion being the degree with the highest mean absolute error of 0.27817 for the subset of volunteers, while the degree of openness to knowledge has the lowest error associated with an error. mean absolute 0.23612.

The results obtained are within the range of the works present in the state of the art, however, the technique used does not show an increase in the precision of the methods of automatic recognition of personality traits compared to the baseline, even Thus, knowing more about the personality of the participants improves the information that is available about them in order to support the decision-making of the entities that work with volunteers.

Keywords: Volunteer; *Big Five*; Personalities; Automatic Detection; Free Text; Analysis; Model;

TABLA DE CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE ILUSTRACIONES

CAPÍTULO 1. INTRODUCCIÓN

1.1 CONTEXTO

Las Organizaciones de la Sociedad Civil (OSC) conocen la importancia de acompañar, crear o facilitar estrategias comunitarias para enfrentar las crisis. En Chile, estas crisis suelen ser habituales y relacionadas con los distintos ámbitos del bienestar local. Estas organizaciones, así como las propias comunidades en contexto de desastre, juegan un papel crítico en el impacto a nivel local de las estrategias de desarrollo (Lavell, 2009), se apoyan fuertemente con la comunidad, generando actividades locales con diferentes grupos e instituciones, consolidando capacidades en el ámbito local, entre otras acciones (Munro, 2015).

Dentro de las dificultades que deben sobrellevar las *OSC* en el contexto de desastres, está la articulación entre diferentes iniciativas y el Estado. Además de lograr un apoyo efectivo en armonía con las necesidades y procesos de las comunidades afectadas (Lillo, 2016).

En una situación de emergencia, gestionar la ayuda de los voluntarios es tan importante como la gestión de recursos materiales que llegan a la zona afectada. Una efectiva coordinación de los recursos humanos, que son finitos, favorece una mejor ejecución de las tareas.

Coordinar todo este personal humano resulta complicado para la Organización debido al tiempo que esto conlleva.

En primer lugar se debe leer las postulaciones, posteriormente categorizar y filtrar la gran cantidad de postulantes a los voluntariados, junto a esto es necesario distinguir ciertas personalidades que sean más afines a ciertas tareas o incluso prevenir factores de riesgo psicológico que se puedan presentar. La información de la personalidad amplía el conocimiento que se posee acerca del voluntario, de modo que puede favorecer la toma de decisiones por parte de la organización.

Existen tareas que de acuerdo a su naturaleza y en conjunto a los rasgos propios del voluntario, ven potenciadas de forma positiva su ejecución. (Barrick, 1991).

1.2 MOTIVACIÓN

Un sondeo realizado por el Instituto Nacional de la Juventud (INJUV) el año 2018 con la participación de más de 1000 jóvenes entre 18 y 25 años, reveló que 1 de cada 4 jóvenes ha participado en alguna actividad de voluntariado (INJUV, 2018). Además, el estudio destaca que

casi el 100% de los voluntarios acordó que es importante estar bien informados antes y durante su participación para apoyar en las tareas que se debe cumplir. Esto requiere un nivel avanzado de gestión por parte de la organización, que generalmente cuenta con pocos recursos tecnológicos, como sistemas de gestión territorial o sistemas de coordinación de recursos.

Por otro lado, las organizaciones establecen protocolos de selección de voluntarios mediante algún instrumento para conocer al participante, como encuestas y formularios a través de herramientas tecnológicas (o no) donde muchas veces debido a la alta convocatoria que estas tienen no es posible verificar y procesar todas las postulaciones teniendo que recurrir a procesos de selección sin considerar atributos importantes del voluntario.

Se tiende a aceptar a los primeros que llenaron el formulario, quedando así postulaciones incluso sin leer o procesar debido al enorme gasto de recursos humanos que conlleva actualmente para una organización realizar completamente esta labor.

Es por ello que mediante un modelo de aprendizaje de máquinas que identifique la personalidad del voluntario se puede ayudar a la organización tanto a conocer al voluntario como a gestionar de mejor medida la distribución de ellos a las funciones o tareas disponibles. Además, esta herramienta puede incentivar la participación del voluntario en otra organización o evento, mejorando la percepción de su participación en la misma.

1.3 IMPACTO ESPERADO

Se espera que el presente proyecto permita fortalecer la toma de decisiones por parte de las organizaciones respecto a la asignación de tareas en terreno, además de ampliar la información que se posee acerca de los voluntarios considerando rasgos de la personalidad.

La utilización de herramientas automáticas para detectar personalidad permite optimizar la cantidad de recursos humanos y de tiempo utilizado en comparación a los métodos tradicionales como BFI-40 (John, 1999) el cual consiste en un formulario estandarizado de 40 preguntas lo cual resulta sumamente extenso en contexto de voluntariados y puede afectar negativamente en los formularios, aumentando la cantidad de formularios incompletos de las postulaciones. Se utilizará BFI-40 para poder etiquetar de manera correcta a los voluntarios participantes y luego poder compararlos con los resultados obtenidos por la red neuronal.

1.4 PLANTEAMIENTO DEL PROBLEMA CIENTÍFICO

No considerar factores psicológicos al momento de la selección de postulantes en contexto de voluntariados ni en la asignación de tareas puede conllevar a problemas respecto a la permanencia y compromiso con el voluntario (González, 2004), así como afectar en el desempeño de ciertas tareas considerando la personalidad (Millette, 2008). En consecuencia considerar factores psicológicos permite optimizar la ejecución de tareas en tiempo y efectividad mediante una asignación inteligente de recursos considerando además otras dimensiones de la personalidad de los voluntarios.

La realización de la tarea de asignación y selección de forma completamente manual aparece como la gran limitante del recurso humano del que disponen las instituciones a cargo de los voluntarios ya que si se consideran que una campaña exitosa supera con facilidad los 1.000 voluntarios (seleccionados) (UACh, 2010), y en una situación de emergencia como el gran incendio de Valparaíso la cantidad de voluntarios considerados superan los 15.000 (Mercurio, 2014) esto hace impensable procesar esa cantidad a través de un sistema completamente manual mediante de los recursos humanos que disponen las instituciones.

Hoy muchos investigadores creen que son cinco los rasgos principales de la personalidad (Power & Pluess, 2015). La evidencia de esta teoría está continuamente en crecimiento, comenzando con la investigación de Fiske (1949) y luego ampliada por otros investigadores, incluidos Norman (1963), Goldberg (1993) continuando con estudios más recientes enfocados al reconocimiento de personalidad con técnicas vanguardistas utilizando con éxito técnicas de Inteligencia Artificial.

Recientemente el campo de de la predicción automática de rasgos de personalidad ha obtenido especial atención específicamente a través de modelos de aprendizaje profundo (Mehta et al., 2019) con sus distintas aplicaciones que van desde personalización de atención a través de asistentes hasta sistemas de recomendación.

Es importante tener en cuenta que cada uno de los factores descritos de personalidad por el modelo *Big Five* representa un rango entre dos extremos. Por ejemplo, el concepto descrito por el autor acerca de la extraversión representa un continuo entre la extraversión extrema y la introversión extrema. Sin embargo, en el mundo real, la mayoría de las personas se encuentran en algún lugar entre el espectro de los dos polos de cada dimensión.

1.5 BRECHA DEL CONOCIMIENTO

De acuerdo al estudio del estado del arte anteriormente descrito existen factores psicológicos asociados a la permanencia y compromiso del voluntario en sus actividades de voluntariado (Millette, 2008), además de un impacto de la motivación y compromiso en el desempeño de las tareas asignadas (González, 2004). Para detectar estos rasgos de personalidad existe en la literatura el modelo *BigFive* el cual es posible detectar de manera automática con técnicas de *deeplearning* (Salminen et al., 2020), especificamente, en el perfil de voluntarios a través de texto libro en español.

1.6 PREGUNTAS DE INVESTIGACIÓN

¿Con qué precisión es posible detectar la personalidad o rasgos de personalidad de los voluntarios a partir del análisis de respuestas de texto libre?

¿Existe una diferencia de más de un 10% en el promedio de precisión utilizando el modelo de la red neuronal al predecir los rasgos de la personalidad respecto al idioma inglés y español?

1.7 HIPÓTESIS DE TRABAJO

La precisión de la detección automática de rasgos de la personalidad a través de textos libres de los voluntarios espontáneos en un contexto nacional es superior a un 70%.

1.8 OBJETIVO GENERAL

Desarrollar un modelo de reconocimiento de rasgos de la personalidad automático para voluntarios detectados mediante el análisis de texto libre con una precisión similar a la encontrada en el estado del arte.

1.9 OBJETIVO ESPECÍFICOS

A continuación, se describen los siguientes objetivos específicos para la investigación:

1. Construir *dataset* de voluntarios con la información requerida, en especial la sección de texto libre para determinar su personalidad mediante el etiquetado de profesionales del área de psicología y posterior entrenamiento supervisado de la red.
2. Implementar modelo de inteligencia computacional que mediante el análisis de un texto libre permita detectar la personalidad del voluntario de manera automática.
3. Comparar resultados obtenidos de la clasificación automática de los voluntarios con la línea base.

1.10 MÉTODO DE INVESTIGACIÓN CIENTÍFICA

Debido a la naturaleza investigativa que posee el proyecto en la cual se obtiene mediante encuestas tanto la información de los rasgos de la personalidad como los textos libres a analizar y además se construye la red neuronal que permita el procesamiento de lenguaje natural a través del análisis de los textos libres de los voluntarios junto a la posterior etapa experimental en la cual se procesan tanto los datos obtenidos en el estado del arte como los obtenidos por la aplicación de la encuesta es que el desarrollo de la investigación se realiza en base al método empírico-analítico donde el conocimiento se presenta de manera lógica, autocorrectiva y progresiva. Con énfasis en el método experimental ya que en este se interviene sobre el objeto de estudio para obtener información acerca del mismo. Respecto al enfoque de la investigación será predominantemente de carácter de laboratorio ya que se busca aislar las variables de interés, permitiendo así un mayor control de la situación bajo estudio (Kelly, 2007).

1.11 IMPLICACIONES DEL ESTUDIO

Por una parte, la componente cualitativa es fuerte durante el principio del proyecto debido a que se estudian los texto libres de los voluntarios para elaborar y enriquecer el *dataset*.

Sin embargo, a medida que avanza el proyecto la componente cuantitativa toma más peso debido a que se busca probar una hipótesis.

1.11.1 Alcance

Dentro de los alcances principales del presente proyecto se destacan los siguientes:

- La participación en la recolección de textos libres será dirigido de manera excluyente a personas que hayan participado en actividades de voluntariados.
- La determinación de las personalidades a trabajar se basan en estudios realizados por Goldberg (1993). Como se ha detallado en la revisión del estado del arte, este modelo ha sido exitosamente relacionado con la manera en que los individuos se expresan de forma escrita.

1.11.2 Limitaciones

Dentro de las principales limitaciones se encuentra la edad del individuo donde se dispondrá a trabajar solo con personas mayores de 19 años ya que durante la adolescencia el estado mental de un adolescente [1] cambia con frecuencia debido a cambios hormonales durante este período de edad (Ranjith R, 2019).

1.11.3 Consideraciones éticas

El presente documento de tesis está dentro del marco del proyecto FONDEF código ID15I20560 “Plataforma de apoyo a la gestión de emergencia y aplicaciones” (Etapa 2), que ya pasó por comité de ética debido a la naturaleza del mismo, sin embargo, el presente proyecto también será sometido a revisión debido a la naturaleza y sensibilidad de los datos con los cuales se trabaja.

Debido a la naturaleza de los datos se mantendrán bajo el anonimato de los voluntarios frente a la publicación de los resultados y al momento de disponibilizar los datos se mantiene la confidencialidad de los datos personales de los voluntarios.

[1] Se define la adolescencia como el periodo de crecimiento y desarrollo humano que se produce después de la niñez y antes de la edad adulta, entre los 10 y los 19 años (OMS, n.d.)

Los voluntarios que participen en el experimento serán informados en todo momento acerca de la experiencia a realizarse y cómo se tratarán los datos, de modo que solo se realiza la experimentación con quienes estén de acuerdo con este acuerdo para evitar conflictos de consentimiento como pasó en el caso de *Cambridge Analyticia* [2] donde se recolectaban datos sin el consentimiento de los usuarios además de utilizar estos datos para fines de los cuales los usuarios desconocían.

1.12 PLAN DE TRABAJO

Para el desarrollo del plan de trabajo se determina como día de inicio el día lunes 5 de octubre del año 2020, la duración de cada una de las actividades está determinado en la figura B.3 señalando las actividades predecesoras de cada una cuando corresponde, esto quiere decir que para efectuar la tarea en particular debe llevarse a cabo como requisito la tarea predecesora. La etapa de documentación es transversal al proyecto.

1.12.1 Definición de actividades

Para el desarrollo del presente proyecto se han definido 4 etapas:

1. Investigación: Contempla el proceso de investigación del estado del arte, si bien contempla 21 días de trabajo éste se desarrollará de manera constante durante todo el proyecto vigilando el avance y aportes que realizan otros investigadores en el campo del reconocimiento automático de personalidad y la utilización de éste en tareas.

2. Preparación: Contempla desde la configuración del ambiente de desarrollo hasta la recolección de datos, incluyendo, la convocatoria de voluntarios.

3. Experimentación: En esta etapa se realiza el entrenamiento, validación, prueba, retroalimentación y análisis de la red construida junto al enriquecimiento del conjunto de datos obtenidos con los voluntarios.

4. Documentación: Esta etapa es transversal al proyecto y consta de registrar los componentes construidos, describir los experimentos realizados y exponer los resultados obtenidos, entre otros.

[2] Así funcionaba la recolección de datos de *Cambrige Analytica* https://www.nytimes.com/es/2018/04/10/espanol/facebook-cambridge-analytica.html

El resumen de los tiempos comprendidos entre cada una de las etapas está registrado en la tabla 1.1, mientras que el listado de cada una de las actividades se encuentra presente en la tabla 1.2.

Etapa	**Fecha de Inicio**	**Fecha de término**
Investigación	05/10/2020	11/02/2020
Preparación y desarrollo	03/11/2020	13/01/2021
Experimentación	14/01/2021	22/02/2021
Documentación	06/10/2020	01/03/2021

Tabla 1.1: Resumen de etapas del proyecto. Fuente: Elaboración propia.

Nº	**Actividad**	**Duración**	**Desde**	**Hasta**
Investigación				
1	Revisión del estado del arte	21	05/10/2020	02/11/2020
Preparación y desarrollo				
2	Configurar ambiente de desarrollo	5	03/11/2020	09/11/2020
3	Definir arquitectura del modelo de red neuronal	5	10/11/2020	16/11/2020
4	Construir modelo de la red neuronal	28	17/11/2020	24/12/2020
5	Entrenamiento del modelo	7	25/12/2020	04/01/2021
6	Validación del modelo	3	05/01/2021	07/01/2021
7	Revisión de formulario junto a institución	5	23/11/2020	27/11/2020
8	Convocatoria de voluntarios junto a la institución para responder formulario	14	30/11/2020	17/12/2020
10	Obtención de formularios de los voluntarios	14	18/12/2020	06/01/2021
11	Tratamiento de los datos de los formularios	5	07/01/2021	13/01/2021
Experimentación				
12	Enriquecimiento del dataset	3	14/01/2021	18/01/2021
13	Entrenamiento supervisado del modelo	7	19/01/2021	27/01/2021
14	Validación cruzada del modelo	3	28/01/2021	01/02/2021
15	Retroalimentación de la red	10	09/02/2021	22/02/2021
Documentación				
16	Documentación	105	06/10/2020	1/03/2021

Tabla 1.2: Lista de actividades con su etapa correspondiente, duración y fechas. Fuente: Elaboración propia.

CAPÍTULO 2. ESTADO DEL ARTE

La publicación central *The Big Five Personality Dimensions And Job Performance: A META-ANALYSIS* de Barrick (1991) enumera 5 grandes personalidades a través del modelo denominado *BigFive*. En el año 2004 se realiza una publicación en la revista de psicología donde se investiga la relación de factores psicológicos asociados a la permanencia y compromiso en los voluntarios (González, 2004). En el año 2008 en la publicación de *Extraction and Use of Personality Traits from Written Commentary* se examinan a 123 voluntarios para calificar su desempeño en tareas voluntarias, los resultados indican que las características del trabajo se relaciona directamente con su motivación impactando así en la satisfacción y desempeño (Millette, 2008). En el 2013 se publica *Private traits and attributes are predictable from digital records of human behavior* donde se genera un modelo de redes neuronales capaz de predecir de forma automática y precisa atributos personales como orientación sexual, sexo, rasgos de personalidad a través de estados de Facebook y sus interacciones ("me gusta") (Kosinski et al., 2013). En la publicación de 2013 de *Workshop on Computational Personality Recognition: Shared Task* se liberan 2 *datasets*: *Essays* y *MyPersonality*. *Essays* es un gran conjunto de datos de textos (alrededor de 2400, uno para cada autor / usuario), recopilados entre 1997 y 2004 etiquetados con clases de personalidad. Los textos han sido producidos por estudiantes que tomaron la prueba Big5. *MyPersonality* es un *dataset* de puntuaciones de personalidad y datos de perfil de Facebook de 250 usuarios y alrededor de 9900 actualizaciones de estado (Celli et al., 2013). En los últimos años se ha buscado continuamente clasificar automáticamente los rasgos de personalidad del texto a través análisis de campos de texto e incluso desde redes sociales. Para el 2017 en la publicación de *Personality Prediction System from Facebook Users* en una conferencia internacional se utilizan modelos de *Super Vector Machine* (SVM), *Multilayer Perceptron* (MLP), junto con algoritmos de *Deep Learning* como *Long short-term memory* (LSTM) y *Convoluntional Neural Network* (CNN) comparando los distintos métodos, destacando los modelos de *Deep Learning*. En el mismo año se publica *Deep Learning-Based Document Modeling for Personality Detection from Text* donde utilizan Redes Neuronales Convolucionales *(CNN)* para predecir personalidades de acuerdo al modelo *BigFive* Majumder et al. (2017). El 2018 de igual manera se utilizan los estados de Facebook convertidos en vectores para ser procesados y predecir la probabilidad de un usuario para pertenecer a *BigFive* predomina *Multi Naive Bayes* (Yuan et al., 2018). Más tarde en el mismo año utilizan a modo de estudio clasificación mediante Perceptrón simple de dos capas utilizando una base de datos con 250 usuarios y casi 10.000 estados. En el presente año 2020 se publica *Private traits and attributes are predictable from digital records of human behavior* donde a través de un entrenamiento de 3 conjuntos de datos distintos: *MyPersonality (Facebook)*, *Essays* y *Youtube*; Resultando un aumento de rendimiento

de un 4.84% de los métodos referidos en la publicación.

2.1 *THE BIG FIVE PERSONALITY DIMENSIONS AND JOB PERFORMANCE: A META-ANALYSIS*

En la publicación (Barrick, 1991) se dedica a investigar las dimensiones de la personalidad del modelo donde se enumeran 5 grandes personalidades: Extraversión, estabilidad emocional, amabilidad, conciencia y apertura a la experiencia con 3 criterios de desempeño laboral: Competencia laboral, competencia de capacitación y datos de personal respecto a cinco grupos ocupacionales: Profesionales, policía, gerentes, ventas y calificados / semi-calificados. Dentro de los resultados generales de la investigación se encuentra que existe una gran relación respecto a una dimensión en particular de la personalidad: la conciencia, desde la perspectiva de los 5 grandes significa una tendencia a ser responsable, organizado, trabajador, orientado a objetivos y a adherirse a normas y reglas. Esta mostró una relación consistente respecto al criterio de desempeño laboral para todos los grupos mencionados. Por otro lado, la apertura a la experiencia como la extraversión permite predecir de forma válida el criterio de competencia en la capacitación (en todas las ocupaciones). Los resultados ilustran los beneficios de usar el modelo de personalidad de 5 factores para acumular y comunicar hallazgos empíricos. Los hallazgos tienen numerosas implicaciones para la investigación y la práctica en psicología del personal, especialmente en los sub-campos de selección de personal, capacitación y desarrollo, y evaluación del desempeño.

2.2 FACTORES PSICOLÓGICOS ASOCIADOS A LA PERMANENCIA Y COMPROMISO DEL VOLUNTARIADO

Esta publicación de la revista de psicología investiga motivaciones asociadas a la permanencia de los voluntarios en la Institución de Hogar de Cristo. Relaciona la motivación con la permanencia en esta institución obteniendo que los desertores obtienen puntuaciones de motivación más baja que quienes pertenecen en el voluntariado en ítems como motivación total, motivaciones dirigidas a valores y motivaciones orientadas a autoestima (González, 2004). Por último, no se encontró que el grado de satisfacción en la organización para las motivaciones de ingreso tuviera un rol determinante en el tiempo de permanencia.

2.3 EXTRACTION AND USE OF PERSONALITY TRAITS FROM WRITTEN COMMENTARY

En esta publicación del 2008 se realiza un estudio para evaluar la aplicabilidad del modelo de características del trabajo (*JCM* por sus siglas en inglés) en organizaciones voluntarias para así poder examinar el impacto del modelo tanto en la motivación como en la satisfacción y la intención de renunciar, con la intención de además evaluar el desempeño de los voluntarios. Participaron 124 voluntarios quienes completaron las mediciones. Los supervisores calificaron el desempeño de las tareas voluntarias y su comportamiento. Los resultados señalan que las características del trabajo se relacionan directamente con la motivación autónoma, la satisfacción y el desempeño que obtuvieron los voluntarios. La motivación autónoma se sitúa como un mediador entre la relación de *JCM* y la satisfacción (Millette, 2008).

2.4 WORKSHOP ON COMPUTATIONAL PERSONALITY RECOGNITION: SHARED TASK

En esta versión se liberan dos *datasets* para que los asistentes e interesados puedan evaluar técnicas de aprendizajes y de caracterización e incluso comparar el rendimiento de las distintas propuestas, en la publicación (Celli et al., 2013) se comparan a 8 sistemas participantes con distintas técnicas dentro de las cuales destacan: *Super Vector Machine (SVM)*, *Bayesian Logistic Regression (BLR)*, *Multinomial Näive Bayes (mNB)*, entre otros. El mejor sistema del estudio muestra cómo la selección de características, en un espacio de características muy grande, puede aumentar el rendimiento de un clasificador, superando al resto (estado del arte), para lograr el alto rendimiento se utilzó algoritmos de clasificación para la selección de funciones *Super Vector Machine* y *Boosting* como algoritmos de aprendizaje. La base de liberada de *MyPersonality* cuenta con más de 10.000 entradas y más de 250 usuarios clasificados por el modelo *BigFive*.

2.5 PRIVATE TRAITS AND ATTRIBUTES ARE PREDICTABLE FROM DIGITAL RECORDS OF HUMAN BEHAVIOR

En esta publicación se demuestra que los registros digitales de fácil acceso como los "me gusta" de Facebook e interacciones se pueden utilizar para predecir de forma automática y precisa una variedad de atributos personales como: orientación sexual, etnia, rasgos de personalidad, entre otros. (Kosinski et al., 2013).

Se utiliza una base de datos de más de 58.000 voluntarios que proporcionaron sus datos en Facebook como interacciones, y perfiles demográficos detallados. El modelo utilizado discrimina correctamente en un 88% de los casos entre hombres homosexuales y heterosexuales. Diferencia en un 95% de los casos entre personas afroamericanas y estadounidenses caucásicos, un 95% el sexo de la persona y para el caso de la detección de personalidad de "apertura" del modelo *BigFive* es similar a la de un *test-retest* de una prueba de personalidad estándar.

Cabe mencionar que converge en la puesta en producción de la plataforma *ApplyMagicSauce* [1] la cual se ha seguido entrenando y ha evolucionado a tal punto que se encuentra en producción e incluso dispone de una *API*[2] la cual puede ser consumida por otra aplicación.

2.6 OVERVIEW OF THE PAN/CLEF 2015 EVALUATION LAB

PAN se ha establecido como el principal foro de investigación en minería de datos/textos que se centran en la identificación de rasgos de personalidad de los autores a través los textos (Stamatatos et al., 2015). En esta edición hay 3 tareas principales y una de ellas es el reconocimiento de rasgos de la personalidad mediante *BigFive* a través de estados de Twitter en distintos idiomas (inglés, español y alemán), cada usuario engloba un conjunto de tweets (promedio n = 100) con etiquetas de personalidad, las etiquetas han sido calculadas siguiendo las respuestas de autoevaluación del autor a la prueba corta de los 5 grandes, BFI-10[3] (Rammstedt & John, 2007) que tiene la base más sólida en el lenguaje y se considera el esquema de reconocimiento de personalidad más ampliamente aceptado y explotado (Poria et al., 2013). La solicitud del conjunto de datos en español se encuentra en curso para poder acceder y ser utilizados en el presente proyecto para la etapa de entrenamiento.

[1] https://applymagicsauce.com/demo

[2] *Application Programming Interface* o API: es un conjunto de funciones y procedimientos que cumplen una o muchas funciones con el fin de ser utilizadas por otro software.

[3] El BFI-10 se ha traducido a varios otros idiomas, por ejemplo, en el Programa de Encuestas Sociales Internacionales (ISSP), incluyendo el español

2.7 PERSONALITY PREDICTION SYSTEM FROM FACEBOOK USERS

En el contexto de la realización de *2nd International Conference on Computer Science and Computational Intelligence 2017* la publicación señala la relación entre la personalidad de una persona mediante el modelo *Big Five* y su actividad en *Facebook* y demuestra que es posible establecer una relación entre texto escrito por el usuario y su personalidad (Tandera et al., 2017). En este articulo se utilizan dos datasets: *myPersonality* y un recolectado manualmente, de tamaño 250 y 150 respectivamente, para etiquetar los estados recogidos manualmente se utiliza *Apply Magic Sauce*. En los modelos utilizados están *Support Vector Machine* (SVM), *Multilayer Perceptron* (MLP) y algoritmos de *Deep learning* como *Long short-term memory* (LSTM) o *Convolutional Neural Network* (CNN). Los resultados obtenidos demuestran que los algoritmos de *deep learning* logran mejorar la precisión de la clasificación.

Los mejores resultados respecto al promedio de precisión con los 5 rasgos para el conjunto de datos *MyPersonality* se obtienen utilizando *SVM* con regresión logística obteniendo 70.40% para el rasgo *OPN*, mientras que con algoritmo *LDA* se obtiene 79.33% para el rasgo *EXT*. Para los métodos de aprendizaje profundo se obtiene que *MLP* obtiene un 79.49% para el rasgo *OPN*

2.8 DEEP LEARNING-BASED DOCUMENT MODELING FOR PERSONALITY DETECTION FROM TEXT

Un estudio realizado por (Majumder et al., 2017), muestra una predicción de la personalidad definida por el modelo de *Big Five*. La predicción la realizan mediante una CNN, con un *dataset* de aproximadamente 2.500 ensayos libres. Las conclusiones obtenidas fueron que necesitaban más datos de entrenamiento para que la CNN sin modificaciones obtuviera mejores resultados.

2.9 PERSONALITY RECOGNITION BASED ON USER GENERATED CONTENT

En (Yuan et al., 2018), se analiza texto desde una base de datos de estados de usuarios de *Facebook*. Si bien el proyecto no trabaja con redes sociales, el análisis se realiza

mediante una CNN, convirtiendo los estados de los usuarios en vectores que luego de una serie de procesamiento sí logran predecir la probabilidad del usuario de pertenecer a una categoría de los *Big Five*. Si bien se obtuvo una mejora en predecir la categoría de *openness*, en el resto de categoría lideró el modelo de *Multinomial Naïve Bayes*.

2.10 EXTRACTION AND USE OF PERSONALITY TRAITS FROM WRITTEN COMMENTARY

En el estudio (Bawa et al., 2018) se utilizan los estados de *Facebook* para definir la personalidad del individuo mediante el modelo de *Big Five*. Esta vez, unen los estados y los tratan como un solo ensayo realizado por el autor. La base de datos utilizada contiene alrededor de 250 usuarios y 10.000 estados. La clasificación es realizada mediante un Perceptrón simple de dos capas y no se detallan los resultados obtenidos.

2.11 ENRICHING SOCIAL MEDIA PERSONAS WITH PERSONALITY TRAITS: A DEEP LEARNING APPROACH USING THE BIG FIVE CLASSES

Se desarrolla a profundidad el estudio de *Automatic Personality Prediction (APD)*: Predicción Automática de personalidad; junto con señalar 3 componentes que fomentan el estudio de este campo, en primer lugar la disponibilidad de los datos (publicaciones en redes sociales, comentarios, entre otros. En segundo lugar la información acerca de la naturaleza de los datos y por último el aumento creciente de la disposición de compartir pensamientos y sentimientos a través de redes sociales (Salminen et al., 2020).

En este estudio se utilizan 3 conjunto de datos para el proceso de entrenamiento: (1) *Essays* referentes a ensayos, (2) *MyPersonality* base de datos con estados de usuarios de Facebook y (3) *Youtube Personality Dataset* con transcripción de videoblogs de Youtube [4].

Se desarrolla un clasificador de aprendizaje profundo, especificamente un modelo compuesto de dos sub-arquitecturas: CNN + LSTM; y los resultados indican un aumento de rendimiento promedio del 4.84% en las puntuaciones de F1 en relación a los métodos referidos en la publicación.

[4]La base de datos de Youtube está disponible mediante solicitud en https://www.idiap.ch/ dataset/youtube-personality

2.12 CONJUNTO DE DATOS

Dentro de la literatura se encuentran distintos conjuntos de datos que han sido construido por distintos autores a lo largo del tiempo que han disponibilizado para fomentar y validar nuevas técnicas que contribuyen al campo de la detección automática de rasgos de la personalidad, en la tabla 2.1 se resume los conjuntos de datos utilizados en las distintas publicaciones de la revisión literaria. Cabe destacar que estos *datasets* están en inglés y serán ocupados exclusivamente para validar la red neuronal propuesta y luego se utilizará exclusivamente conjuntos de datos en idioma español para las etapas de entrenamiento y validación de la herramienta propuesta.

Nombre *Dataset*	**Descripción**
Essays I *(Pennebaker & King, 1999)*	2468 Ensayos anónimos etiquetados por el autor con los rasgos de personalidad definidos
Essays II *(Tausczik & Pennebaker, 2009)*	2400 Ensayos etiquetados de forma manual según diferentes rasgos de personalidad. Luego los puntajes eran convertidos a personalidad a través de una regresión
MyPersonality	myPersonality fue una aplicación de Facebook que a través de un cuestionario recolectaba información. En 2018 decide dejar de liberar sus datasets. Existe una versión disponible con 10.000 estados de 250 usuarios anónimos

Tabla 2.1: *Datasets* encontrados en la literatura para realizar detección de rasgos de personalidad. Fuente: Elaboración propia.

2.13 MODELOS Y SUS RESULTADOS

A continuación se resumen algunos modelos encontrados en las publicaciones revisadas en la sección anterior detallando el *dataset* utilizado de entrenamiento, junto a los resultados obtenidos de precisión.

Algoritmo	**Resultado (precisión)**	**Conjunto de datos**
SVM	72.00%	*ESSAYS+MP*
SVM, kNN, NB	58.60%	*MP+ESSAYS*
LR	63.00%	*MP*
SVM,BLR, mNB	58.60%	*MP*
SVM	57.00%	*ESSAYS*
NB	56.30%	*ESSAYS*

Tabla 2.2: Modelos y sus resultados de precisión publicados en *Workshop CPR*. (Celli et al., 2013).

Algoritmo	**Promedio**	**Conjunto de datos**
Naive Bayes	61.76%	MyPersonality
SVM	61.04%	MyPersonality
Logistic Regression	61.44%	MyPersonality
Gradient Boosting	62.00%	MyPersonality
LDA	63.04%	MyPersonality

Tabla 2.3: Resultados de precisión de algoritmos tradicionales comparativos obtenidos ICCSCI 2017, Indonesia. (Tandera et al., 2017).

Algoritmo	**Promedio**	**Conjunto de datos**
MLP	70.78%	MyPersonality
LSTM	58.63%	MyPersonality
GRU	63.44%	MyPersonality
CNN 1D	63.84%	MyPersonality
LSTM+CNN 1D	62.71%	MyPersonality

Tabla 2.4: Resultados de precisión para algoritmos de *machine learning* comparativos obtenidos ICCSCI 2017, Indonesia. (Tandera et al., 2017).

Modelo	**EXT**	**OPE**	**CON**	**AGR**	**NEU**	**Conjunto de datos**
CNN + LSTM*	0.541	0.529	0.538	0.553	0.484	*ESSAYS*
CNN + LSTM**	0.662	0.653	0.543	0.603	0.332	*ESSAYS+MPD+YT*

Tabla 2.5: Valores F1 obtenidos para cada uno de los rasgos de personalidad. (Salminen et al., 2020).

	EXT	NEU	AGR	CON	OPE
Inglés	0.125	0.195	0.131	0.110	0.120
Español	0.132	0.163	0.103	0.102	0.111

Tabla 2.6: Tabla resumen de los mejores resultados obtenidos de los distintos participantes de *PAN15*, resultados expresados en error promedio para cada rasgo de la personalidad por conjunto de datos inglés y español. (Rangel et al., 2015).

Los resultados expuestos en la tabla 2.6 serán utilizados para comparar los resultados obtenidos en la presente investigación ya que son los mejores resultados encontrados en el estado del arte.

CAPÍTULO 3. MARCO TEÓRICO

En el presente capítulo se describen aquellos fundamentos teóricos esenciales para el correcto entendimiento del proyecto, además, se describe el funcionamiento de éstos y cómo impactan en su desarrollo.

3.1 BIG FIVE

Big Five: Es un modelo que busca explicar la personalidad del individuo a través de cinco categorías. Este modelo ha evolucionando con el tiempo a través de distintos autores y trabajos en el área de la psicología, desde las rotaciones oblicuas de las 13 escalas de Guilford (Thurstone, 1951) y posteriormente el desarrollo de los siete factores en el Programa de temperamento de Thurstone (Thurstone, 1953), por nombrar algunos de los trabajos relevantes en el desarrollo del estudio de rasgos de la personalidad desde distintas perspectivas y factores, sin embargo, desde los estudios de la consistencia factorial de las estructuras de la personalidad desde distintas fuentes (Fiske, 1949) y de estudios de una taxonomía adecuada de los atributos de personalidad: estructura de factores replicada en las calificaciones de personalidad de nominación (Norman, 1963) se analiza que solo cinco factores demostraron ser replicables y dentro de los estudios que toman estos 5 factores destaca el modelo propuesto por Goldberg en su trabajo donde estudia las estructura de los rasgos de la personalidad fenotípicas (Goldberg, 1993) donde se describen los siguientes rasgos de la personalidad.

1. Extraversión o en inglés *Extraversion (EXT)*: Caracteriza a personas sociales, emocionales y asertivas. En una baja probabilidad describe a personas solitarias, que piensan antes de hablar.

2. Afabilidad o en inglés *Agreeableness (AGR)*: Incluye a personas que confiables, altruistas, amables y empáticas. Los individuos que presentan una baja probabilidad describen personas poco preocupadas por el resto, que no tienen interés en los problemas de los demás.

3. Conciencia o en inglés *Conscientiousness (CON)*: En una alta probabilidad describe personas meticulosas, que se preparan para nuevas tareas y ponen atención a los detalles. En una baja probalidad describe personas que no le gustan la planificación y estructura, fallan en completar tareas.

4. Neuroticismo o en inglés *Neuroticism (NEU)*: El neuroticismo es un rasgo caracterizado por la tristeza, el mal humor y la inestabilidad emocional.

5. Apertura o en inglés *Openness (OPE)*: En una alta probabilidad describe personas creativas, con imaginación, abierta a probar nuevas cosas. En una baja probabilidad, describe personas que no le gustan los cambios, no intentan cosas nuevas, resistentes a nuevas ideas.

Solo 5 rasgos pueden parecer poco, sin embargo, estos amplios dominios incorporan cientos, si no miles, de rasgos: el factor I que va desde la surgencia o extraversión contrasta rasgos como locuacidad, asertividad y nivel de actividad con rasgos como el silencio, la pasividad y la reserva, por otro lado, el factor II Afabilidad contrasta rasgos como la bondad, la confianza y la calidez con rasgos como la hostilidad, el egoísmo y la desconfianza, mientras que el factor III conciencia o confiabilidad contrasta rasgos tales como organización, minuciosidad y confiabilidad con rasgos tales como descuido, negligencia y falta de confiabilidad, el factor IV que va desde estabilidad emocional a neuroticismo incluye rasgos como nerviosismo, mal humor y temperamentalidad, por último, el factor V ya sea etiquetado como intelecto o apertura a la experiencia contrasta rasgos como la imaginación, la curiosidad y la creatividad con rasgos como la superficialidad y la imperceptibilidad, es decir, cada rasgo de la personalidad se puede dar desde un inicio del espectro hasta el final de éste pasando por grados intermedios entre estos extremos ya que lo normal es no situarse en ninguno de los extremos, pero sí decantarse más por un polo u otro, el resumen de los extremos puede verse en la figura 3.1.

Figura 3.1: Intérvalos de los rasgos de la personalidad del modelo BigFive. Elaboración propia.

Las siguientes definiciones de los 5 rasgos de la personalidad corresponden a una consulta psicológica española (Blázquez, 2019) y son adaptadas al contexto del presente estudio.

3.1.1 Extraversión

Las personas extrovertidas tienden a ser más sociables, a ser atrevidas socialmente, a buscar la compañía de los demás y a evitar la soledad. Se sienten comprometidas más con el mundo externo que con el interno y tienden a buscar nuevas sensaciones en compañía de otros. Además no sienten incómodos al atraer la atención sobre sí mismos, al contrario, tienen buenas sensaciones en este tipo de situaciones.

El polo opuesto sería la introversión, referida a personas reservadas, independientes, con menos tendencia a experimentar nuevas sensaciones con los demás, ya que se sienten más cómodas en la rutina, realizando cosas que no se salgan de lo habitual. Prefieren realizar actividades en solitario antes que hacer actividades animadas o muy estimulantes con otras personas. Esto no quiere decir que estas personas no sean sociales, sí disfrutan del contacto social, pero de forma diferente a la de las personas más extrovertidas. De hecho, en los momentos en los que se encuentran en un círculo de mucha confianza, pueden mostrarse tan animados y sociales como las personas extrovertidas.

3.1.2 Amabilidad

Una alta cordialidad refleja a personas con altos niveles de altruismo, solidaridad, personas confiadas y francas, con mayor capacidad para establecer relaciones interpersonales amistosas.

El polo opuesto es la hostilidad, también se denomina como “no agradabilidad”, lo que no es necesariamente algo negativo, ya que tienden a ser personas escépticas y con un pensamiento crítico, atributos necesarios para el desarrollo de diversos ámbitos, como el de la ciencia.

3.1.3 Conciencia

Este factor se basa en el autocontrol, la planificación, la organización , la persistencia hacia el logro de metas y la autodisciplina.

En su polo opuesto se encuentran personas más informales, descuidadas o espontáneas.

3.1.4 Neuroticismo o Inestabilidad Emocional

Las personas que obtienen una puntuación baja en este factor suelen presentar ansiedad, baja tolerancia a la frustración y al estrés y se centran en las posibles consecuencias negativas.

El opuesto es la estabilidad emocional, una persona estable emocionalmente es capaz de afrontar situaciones difíciles o estresantes sin experimentar gran variabilidad emocional, son personas más flexibles y calmadas.

3.1.5 Apertura a la experiencia

Esta dimensión mide la apertura al cambio. Las personas con una alta apertura a la experiencia son personas creativas, con una alta imaginación, curiosidad intelectual, con gusto por el arte y la estética. Son personas que están en contacto con sus emociones y con las de las demás, se interesan por conocer nuevas ideas y tener nuevas experiencias.

En el polo opuesto, cerrazón a la experiencia, se sitúan personas más convencionales con ideas más conservadoras, muestran interés por las tradiciones y prefieren las cosas familiares a las novedosas.

3.2 BIG FIVE INVENTORY

Existe un instrumento llamado *BigFiveInventory* o por sus siglas *BFI-44* la cual consta de 44 preguntas contestadas con una escala de *likert* y que permite una evaluación eficiente y flexible de las cinco dimensiones de la personalidad (John et al., 1991).

las preguntas BFI son cortas y evitan estructuras oracionales complejas, conservando las ventajas de la brevedad y simplicidad, mientras evitan significados ambiguos o múltiples. Los participantes califican cada ítem BFI en una escala de 5 puntos que va de 1 (muy en desacuerdo) a 5 (muy de acuerdo). Las puntuaciones de escala se calculan como la respuesta media del participante al ítem (es decir, sumando todos los ítems calificados en una escala y dividiendo por el número de ítems en la escala). A pesar de su brevedad, el BFI no sacrifica ni la cobertura de contenido ni las buenas propiedades psicométricas (PsycNetDirect, 2020).

El BFI a pesar de no ser de dominio público sí está disponible gratuitamente para que los investigadores lo utilicen con fines de investigación no comerciales.

Debido al gran impacto y relevancia que ha obtenido este instrumento ha sido traducido a lo largo de su historia en los siguientes idiomas: chino, holandés, alemán (solo BFI-10), inglés, hebreo, italiano, lituano, portugués. español, sueco.

Al ser una escala de *likert* de 1 a 5 existe la posibilidad que muchas respuestas sean "en medio", es decir, con puntaje igual a 3, sin embargo, según el autor su uso resulta apropiado y estos se agregan de todos modos al puntaje general ya que esta es realmente una escala de respuesta dimensional, no un cuestionario "verdadero o falso". De hecho, en algunos ítems, responder "3" es en realidad un buen diagnóstico; por ejemplo, responder "3" en el ítem inverso de Afabilidad "Empieza a pelearse con los demás" significa que el participante está admitiendo una considerable disconformidad (la mayoría de las personas responden 2 o incluso 1). Entonces, a menos que todas las respuestas sean 3 respuestas no es motivo de preocupación

Las 44 preguntas que componen el cuestionario son:

1. Es hablador/a
2. Tiende a encontrar fallas en los demás
3. Es cuidadoso/a en su trabajo
4. Es depresivo/a, triste
5. Es original, tiene ideas nuevas
6. Es reservado/a
7. Actúa desinteresadamente con los demás y les ayuda
8. Puede ser algo descuidado/a
9. Es relajado/a, maneja bien el estrés
10. Tiene curiosidad por cosas muy diferentes
11. Está lleno/a de energía

12. Suele reñir con los demás
13. Es un/a trabajador/a de fiar
14. Puede sentirse tenso/a
15. Es ingenioso/a, un/a pensador/a profundo/a
16. Genera mucho entusiasmo
17. Perdona fácilmente
18. Tiende a ser desorganizado/a
19. Se preocupa bastante
20. Tiene una imaginación activa
21. Tiende a ser tranquilo
22. Es generalmente confiado
23. Tiende a ser perezoso/a
24. Es emocionalmente estable, que no se altera con facilidad
25. Tiene Creatividad
26. Tiene una personalidad asertiva
27. Puede ser frío/a y distante
28. Persevera hasta que la tarea se haya terminado
29. Puede tener mal genio
30. Valora las experiencias artísticas
31. Es algo tímido/a, inhibido/a
32. Es considerado/a y amable con casi todo el mundo
33. Hace las cosas de manera eficiente
34. Permanece tranquilo/a en las situaciones tensas
35. Prefiere el trabajo rutinario
36. Es extrovertido/a, sociable
37. Es a veces rudo/a con los demás

38. Hace planes y los lleva adelante
39. Se pone nervioso/a fácilmente
40. Le gusta reflexionar, jugar con las ideas
41. Tiene pocos intereses artísticos
42. Le gusta cooperar con los demás
43. Se distrae fácilmente
44. Es refinado en arte, literatura, música

Donde cada una de las preguntas se responde en una escala de 1 a 5 y su representación es la siguiente:

1. Muy en desacuerdo
2. En desacuerdo un poco
3. Ni de acuerdo ni en desacuerdo
4. De acuerdo un poco
5. Totalmente de acuerdo

Para realizar el cálculo del puntaje total se debe realiza de la siguiente manera, cabe destacar que la R representa elementos con puntuación inversa, es decir, si el puntaje obtenido en una pregunta tipo R fue un 2, en la escala de *likert* [1, 2, 3, 4, 5], el puntaje obtenido de la pregunta 6 se convierte en 4, se puede generalizar como $valor\_revertido = 6 - valor\_real$:

- Extraversión: 1, 6R, 11, 16, 21R, 26, 31R, 36
- Amabilidad: 2R, 7, 12R, 17, 22, 27R, 32, 37R, 42
- Conciencia: 3, 8R, 13, 18R, 23R, 28, 33, 38, 43R
- Neuroticismo: 4, 9R, 14, 19, 24R, 29, 34R, 39
- Apertura a la experiencia: 5, 10, 15, 20, 25, 30, 35R, 40, 41R, 44

3.3 *WORD EMBEDDING*

Word Embedding o incrustaciones de palabras en su traducción literal al español, es una representación aprendida de un texto donde las palabras que tienen el mismo significado poseen una representación similar, este conjunto de técnicas de modelado y técnicas de aprendizaje poseen gran impacto en el campo del procesamiento de lenguaje natural (PLN) ya que son considerados como uno de los avances claves en el aprendizaje profundo en los problemas del procesamiento del lenguaje natural debido a los beneficios que su uso otorga.

Uno de los beneficios de usar vectores densos y de baja dimensión es computacional: la mayoría de los conjuntos de herramientas de redes neuronales no funcionan bien con vectores dispersos de muy alta dimensión. El principal beneficio de las representaciones densas es el poder de generalización: si creemos que algunas características pueden proporcionar pistas similares, vale la pena proporcionar una representación que sea capaz de capturar estas similitudes. (Goldberg, 2017).

Word Embedding son una clase de técnicas en las que las palabras individuales se representan como vectores con un valor real en un espacio vectorial predefinido por este. Cada palabra del texto se asigna a un vector y los valores del vector se aprenden de una manera que se asemeja a una red neuronal, es por esto, que resulta sumamente complementario al uso que está enfocado el estudio.

Cada palabra está representada por un vector de valor real, de múltiples dimensiones, decenas o cientos. Sin embargo, esto contrasta con los miles o millones de dimensiones requeridas para representaciones de palabras dispersas, como una codificación *one-hot*

Dado un vocabulario (conjunto de palabras) se asocia un vector de características de palabras distribuidas en él, este vector de características representa diferentes aspectos de la palabra, cada palabra está asociado con un punto en el espacio vectorial y la cantidad de funciones es menor que el tamaño del vocabulario (Bengio et al., 2003).

Estas representaciones son asignadas y aprendidas en base al uso de palabras, esto permite que las palabras que se usan de forma similar tengan representaciones similares lo que entenderse como que aprenden su significado natural, a diferencia de una bolsa de palabras ya que en esta representación las palabras distintas poseen representaciones diferentes no teniendo en cuenta su uso.

3.3.1 Algoritmos de WordEmbedding

A continuación se describen 2 de los algoritmos más utilizados en la literatura para los *WordEmbedding*, junto a la descripción se señala qué proyecto lo utiliza a modo de ejemplificar su uso y sus ventajas.

Word2Vec

Word2Vec es un método estadístico para aprender de manera eficiente una incrustación de palabras *(WordEmbedding)* independientes de un corpus de texto.

Fue desarrollado por Tomas Mikolov en Google en 2013 como una respuesta para hacer que el entrenamiento basado en redes neuronales de la incrustación sea más eficiente y desde entonces se ha convertido en el estándar para desarrollar *WordEmbedding* previamente entrenadas.

Además, el trabajo involucró el análisis de los vectores aprendidos y la exploración de la matemática vectorial en las representaciones de palabras. Por ejemplo, que restar la "masculinidad" de "Rey" y agregar "feminidad" da como resultado la palabra "Reina", capturando la analogía "el rey es la reina como el hombre es la mujer". (Mikolov et al., 2013b).

Dentro de Word2Vec existen dos modelos de aprendizaje diferentes que se pueden utilizar como parte del enfoque de aprender:

1. Bolsa de palabras continua o modelo *CBOW*: Se aprende la incrustación al predecir la palabra actual en función de su contexto.
2. Modelo continuo de *Skip-gram*: Se aprende prediciendo las palabras circundantes dada una palabra actual.

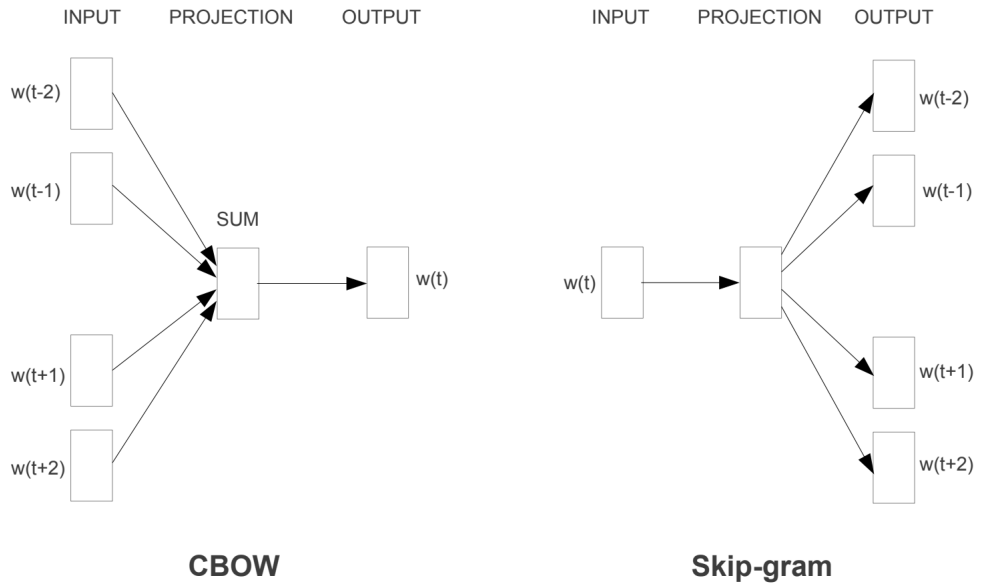


Figura 3.2: Modelos de entrenamiento para *Word2Vec*, tomado de (Mikolov et al., 2013a).

En la figura 3.2 se puede apreciar que ambos modelos están enfocados en aprender sobre palabras dado su contexto de uso local, donde el contexto se define por una ventana de palabras vecinas. Esta ventana es un parámetro configurable del modelo. El tamaño de la ventana deslizante tiene un fuerte efecto sobre las similitudes vectoriales resultantes. Las ventanas grandes tienden a producir más similitudes de actualidad, mientras que las ventanas más pequeñas tienden a producir similitudes más funcionales y sintácticas.

Un beneficio apreciable de este enfoque es que las *WordEmbedding* de alta calidad se pueden aprender de manera eficiente (poca complejidad de espacio y tiempo), lo que permite aprender *Embedding* más grandes (más dimensiones) de corpus de texto mucho más grandes (miles de millones de palabras) y luego distribuir estos modelos para realizar los distintos estudios.

3.3.2 FastText

FastText es esencialmente una extensión del modelo *Word2Vec* y la diferencia es que éste trata cada palabra como compuesta de n-gramas de caracteres. Entonces, el vector de una palabra está hecho de la suma de este carácter n-gramos.

FastText es una biblioteca creada por el equipo de investigación de Facebook para el aprendizaje eficiente de representaciones de palabras y clasificación de oraciones debido a la gran cantidad de datos que tenían acceso y debían procesar. La biblioteca ha ganado

mucha tracción en la comunidad de *NLP* y es una posible sustitución del paquete Gensim que proporciona la funcionalidad de *Word2Vec*.

Como se menciona anteriormente *Word2Vec* aprende vectores solo para palabras completas que se encuentran en el corpus de capacitación. *FastText*, por otro lado, aprende vectores para los n-gramas que se encuentran dentro de cada palabra, así como cada palabra completa. En cada paso de entrenamiento en *FastText*, la media del vector de palabra objetivo y sus vectores n-gramas componentes se utilizan para el entrenamiento. El ajuste que se calcula a partir del error se usa luego de manera uniforme para actualizar cada uno de los vectores que se combinaron para formar el objetivo. Esto agrega muchos cálculos adicionales al paso de entrenamiento. En cada punto, una palabra necesita sumar y promediar sus componentes de n-gramas. La compensación es un conjunto de vectores de palabras que contienen información de sub-palabras incrustadas. Se ha demostrado que estos vectores son más precisos que los vectores de *Word2Vec* mediante varias medidas diferentes.

La función N-gram es la mejora más significativa en *FastText*, está diseñada para resolver el problema de "fuera de vocabulario" o por sus siglas en inglés *OOV (Out of Vocabulary)*.

Por ejemplo: la palabra en inglés “aquarium” puede ser separada en los siguientes n-gramas “<aq/aqu/qua/uar/ari/riu/ium/um>”, donde "<" indica el inicio de la palabra y “>” el término de una palabra.

Bajo el enfoque tradicional si encuentra la palabra "Aquarius", es posible que no sea reconocida, sin embargo bajo el engoque de *FastText* puede adivinar debido a la parte de ambas palabras "Aquarium" y "Aquarius" comparten.

Las principales diferencias de ambos son:

1. Genera mejores representaciones para palabras raras (incluso si las palabras son raras, sus n-gramas de caracteres se comparten con otras palabras, por lo que las incrustaciones pueden ser buenas), estas palabras pueden ser propias de origenes de redes sociales donde el lenguaje se ve transformado

2. Esto se debe simplemente a que, en *Word2Vec*, una palabra rara (por ejemplo, 10 apariciones) tiene menos vecinos, en comparación con una palabra que aparece 100 veces; esta última tiene más palabras de contexto vecino y, por lo tanto, su mayor frecuencia se traduce en un mayor peso.

3. Sin palabras de vocabulario (OOV): pueden construir el vector de una palabra a partir de sus n-gramas, incluso si una palabra no aparece en el corpus de entrenamiento. Mientras que Word2vec como no puede.

4. A medida que aumenta el tamaño del corpus, también aumenta el requisito de memoria: aumentaría la cantidad de n-gramos que se cifran en el mismo depósito de n-gramos. Por lo

tanto, la elección del hiperparámetro que controla los depósitos de hash totales, por ejemplo, para crear trtar un texto con 50 millones de palabras con un mínimo y máximo de n-gramas igual a 3 y un recuento de palabras igual a 15 (quiere decir que se eliminan las palabras que su frecuencia es menor) una máquina de 256 GB de RAM sería insuficiente para generar estos vectores de palabras.

5. El uso de LSTM favorece a FastText en la generación de *WordEmbedding* debido al uso de memoria contextual que ofrece este tipo de redes.

3.3.3 WordEmbedding usados

Debido a que se estudia conjuntos de datos en inglés y español es necesario utilizar *WordEmbeddings* distintos para cada uno de los idiomas, en el caso del idioma español el repositorio[1] corresponde a la Universidad de Chile y está bajo licenciamiento MIT[2], por otro lado, para el idioma inglés los archivos[3] se encuentran bajo licenciamiento *Creative Commons Attribution-Share-Alike License 3.0*[4].

Cuerpo	Tamaño	Algoritmo	# vectores	dim	Créditos
Spanish Unannotated Corpora	2.6 GB	FastText	1.313.423	300	J. Cañete
Common Crawl	4.5 GB	FastText	2.000.000	300	T. Mikolov

Tabla 3.1: Resumen de *WordEmbeddings* utilizados para cada idioma. Fuente: Elaboración propia.

Español

Para el idioma español se ha utilizado un *WordEmbedding* donde a diferencia de los modelos populares que aprenden las representaciones tal y como son ignorando la morfología de las palabras asignando un vector distinto a cada palabra limitando así los idiomas extensos y con palabras "raras" como lo es el español donde existen palabras que en contextos similares se refieren a lo mismo o sus palabras son morfológicamente muy parecidas. Este enfoque se basa en el modelo de *SkipGram* donde cada palabra se representa como una bolsa de n-gramas de caracteres (Bojanowski et al., 2017) a cada n-grama se le asocia una representación vectorial y

[1] https://github.com/dccuchile/spanish-word-embeddings
[2] https://opensource.org/licenses/MIT
[3] https://fasttext.cc/docs/en/english-vectors.html
[4] https://creativecommons.org/licenses/by-sa/3.0/

las palabras se representan como la suma de estas representaciones, es un método rápido lo que permite entrenar modelos con grandes cuerpos y calcular así la representación de las palabras e incluso aquellas que no aparecen en los datos de entrenamiento.

El modelo de José Cañete está basado en el cuerpo *Spanish Unannotated Corpora (SUC)* y posee más de 1 millón 300 mil vectores y la dimensión de estos vectores es de 300, sin embargo existen otras versiones[5] de 10, 30 y 100 dimensiones, sus parámetros especiales en la implentación de n-grama son las siguientes:

- Tamaño mínimo de n-grama = 3
- Tamaño máximo de n-grama = 6
- Mínimo de frecuencia = 5
- Épocas = 20
- Dimensión = 300
- Resto de parámetros por defecto

Inglés

Para el caso de idioma inglés se utiliza un *WordEmbedding* basado en el algoritmo *FastText* el cual posee 2 millones de vectores de palabras entrenados en Common Crawl (Mikolov et al., 2018). Este modelo fue preentrenado con 300 dimensiones y es lanzado por desarrolladores de facebook.

La primera línea del archivo contiene el número de palabras del vocabulario y el tamaño de los vectores. Cada línea contiene una palabra seguida de sus vectores, como en el formato de texto *FastText* predeterminado. Cada valor está separado por espacios. Las palabras están ordenadas por frecuencia descendente.

[5]https://github.com/dccuchile/spanish-word-embeddings/blob/master/emb-from-suc.md

CAPÍTULO 4. ESCENARIO EXPERIMENTAL

En el presente capítulo se detalla en cada una de las secciones lo necesario para llevar a cabo la experimentación, las condiciones que se deben reunir, el procedimiento a realizar, la descripción de los datos a considerar, software, hardware requerido y cada uno de los parámetros que sean esenciales para el correcto funcionamiento del experimento.

La realización del experimento como tal se divide en dos: preparación y experimento. La etapa de preparación está dado en primera instancia por la recopilación de información a través de encuestas de textos libres de voluntarios junto a la construcción de la red neuronal. Por otro lado en la etapa de experimento donde se enriquece el *dataset* con los nuevos datos, se validan y se comparan los resultados obtenidos.

4.1 AMBIENTE EXPERIMENTAL

De acuerdo a lo anteriormente mencionado se disponen de ambientes experimentales para ambas fases del proyecto

4.1.1 Ambiente preparación

El ambiente experimental en la fase de preparación está dado de forma no presencial y para este sólo se requiere equipo personal donde los voluntarios realicen las encuestas para enriquecer el *dataset*.

4.1.2 Ambiente experimento

Para el caso de la experimentación será de carácter de laboratorio ya que se busca aislar el estudio de las variables de interés, se decide utilizar esta forma de experimentación ya que permite un mayor control de la situación bajo estudio (Kelly, 2007).

4.2 PROCEDIMIENTO EXPERIMENTAL

Para la ejecución del proyecto se utilizan dos fases: preparación y experimento. En la primera se realizan las recopilación de encuestas, construcción y enriquecimiento del conjunto de datos. Por otro lado en la segunda se compone de la experimentación y aplicación del modelo para validar la hipótesis del proyecto donde se construirá y aplicará el modelo de clasificación.

4.2.1 Preparación

En la primera parte del proyecto es necesario definir las dimensiones que serán abordadas para construir el formulario a utilizar para posteriormente recopilar la información de los voluntarios para construir el *dataset* a utilizar, posteriormente la construcción del modelo de la red neuronal a utilizar, de acuerdo a la arquitectura seleccionada y finalmente la etapa de entrenamiento supervisado del modelo junto a los profesionales. Cabe destacar que junto a la encuesta realizada a los voluntarios se aplicará el instrumento *BFI-40* (John, 1999) para medir los rasgos de la personalidad de *BigFive* y estos con estos datos poder comparar la precisión del modelo propuesto.

1. Implementación de modelo de inteligencia computacional de detección de rasgos de personalidad de manera automática.
2. Recopilar información necesaria a través de encuestas realizadas a través de un formulario en *Google Forms* para enriquecer el *dataset PAN* con entradas de por lo menos 80 nuevos voluntarios con sus respectivos campos incluyendo los textos libres.
3. Puesta en marcha de la arquitectura del modelo seleccionado en la tecnología Python 3.

4.2.2 Experimento

Para responder la pregunta de hipótesis central del proyecto se desarrolla un experimento el cual se puede separar en 3 etapas las cuales son:

1. Entrenamiento supervisado donde se realizará el análisis y posterior determinación de la personalidad de las nuevas entradas en el *dataset*, toda esta información será vital para la etapa de entrenamiento del modelo de la red neuronal, de esta forma a futuro la red

neuronal podrá discriminar la personalidad del voluntario, de acuerdo a lo que aprendió en este proceso.

2. Validación de ajuste del modelo de acuerdo al nuevo conjunto de datos de entrenamiento agregado con los datos de los voluntarios

3. Análisis y comparación de resultados obtenidos con la línea base.

4.3 SOFTWARE Y HARDWARE REQUERIDO

En la presente sección se detalla los *software* requeridos para la correcta ejecución del proyecto junto a los requerimientos recomendados de *hardware* para realizar la construcción y posterior validación del sistema.

4.3.1 Software

Para efectos la fase de preparación es necesario contar con una *suite* de software necesarios para poder realizar las repectivas tareas de puesta en marcha del modelo de arquitectura junto al entrenamiento del modelo, dichos *software* son: *Anaconda* versión 4.8.4, *Jupyter* en su versión 1.0, *TensorFlow* versión 2.3.0 y *Keras* en su versión estable 2.3.

4.3.2 Hardware

Para efectos la fase de preparación es necesario contar con un computador que cuente con los requisitos recomendadosde CPU [1] RAM [2] y GPU [3] para poder ejecutar de forma satisfactoria la *suite* de software mencionados anteriormente.

[1] Basados en recomendación de https://timdettmers.com/2018/12/16/deep-learning-hardware-guide/

[2] De acuerdo a https://towardsdatascience.com/setup-an-environment-for-machine-learning-and-deep-learning-with-anaconda-in-windows-5d7134a3db10

[3] Para más información acerca de GPUs NVIDIA para *deep learning* y su compatibilidad visitar https://developer.nvidia.com/cuda-gpus.

Central Processing Unit (CPU)	4 núcleos sobre 2GHz
RAM	mínimo 8GB y recomendado 16GB
Graphics Processing Unit (GPU)	Tarjeta GPU NVIDIA® con capacidad de procesamiento CUDA® 3.5 o versiones posteriores que sea compatible con cuda
Sistema Operativo	Ubuntu 18.04 o Windows 10

Tabla 4.1: Requisitos de Sistema recomendados para *Deep Learning*. Elaboración Propia.

4.3.3 Ejecución en la nube

Los resultados expuestos en la presente investigación se obtienen a través de la ejecución de una instancia de máquina virtual en el servicio de *Google Cloud* [4], específicamente consumiendo el servicio de AI-Platform en el cual permite incorporar tecnologías de vanguardia en el campo de la inteligencia artificial, permitiendo escoger características de la máquina a utilizar como *RAM*, *GPU* y *CPU* variando el valor del servicio de acuerdo a las especificaciones de nuestra máquina.

Central Processing Unit (CPU)	4 vCPUs *(n1-standard-4)*
RAM	16 GB
Graphics Processing Unit (GPU)	NVIDIA Tesla T4
Sistema Operativo	Ubuntu 20.04

Tabla 4.2: Especificaciones utilizadas en la instancia de máquian virtual en *Google Cloud*. Elaboración Propia.

4.4 DESCRIPCIÓN DE LOS DATOS

A continuación se presenta la estructura de los datos a trabajar, se proponen 2 estructuras de datos que son abordadas en detalle en la próxima sección. Junto a esto, se exponen las distintas bases de datos con las que se trabaja, el origen de estas, una descripción de los datos y los alcances de estas.

[4] https://console.cloud.google.com/ai-platform

4.4.1 Estructura general

La estructura general es aquella que predomina en el proyecto, de este modo los distintos conjuntos de datos a trabajar son homologados a esta estructura, leyendo los datos, transformándolos (en caso de ser necesario) y guardados de manera local.

Autores

Para el caso de los autores se dispone de un modelo con los siguientes atributos

AUTHOR_ID	Identificador del autor del texto
HASHNAME	Un texto codificado para asegurar el anonimato de los usuarios

Tabla 4.3: Atributos modelo *AUTHORS*. Fuente: Elaboración propia.

Texto libre y Rasgos de la Personalidad

Para el caso de los textos libres se dispone de un modelo con los siguientes atributos

#AUTHID	Identificador del autor del texto
STATUS	Texto en particular a analizar
sEXT	Grado de Extroversión
sNEU	Grado de Inestabilidad Emocional
sAGR	Grado de Amabilidad
sCON	Grado de Responsabilidad
sOPN	Grado de Apertura a Nuevas Experiencias

Tabla 4.4: Atributos tabla *Status*. Fuente: Elaboración propia.

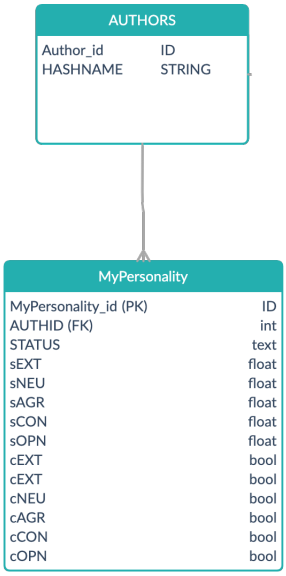


Figura 4.1: Diagrama de Entidad Relación. Fuente: Elaboración propia. Fuente: Elaboración propia.

A modo de desnormalizar la relación entre Autores y Rasgos de la personalidad estos son dispuestos en un único modelo que unifica ambos datos y cada entrada es una línea en un archivo *Comma Separated Value (CSV)*.

#AUTHID	Status	sEXT	sNEU	sAGR	sCON	sOPN
# Identificador	Texto	Flotante	Flotante	Flotante	Flotante	Flotante

Tabla 4.5: Línea de ejemplo para representación de archivo .CSV. Fuente: Elaboración propia.

4.4.2 *Dataset MyPersonality*

El conjunto de datos de MyPersonality (Celli et al., 2013) utilizado es la versión con 10.000 entradas de 250 usuarios.

El dataset de *MyPersonality* posee las siguientes columnas, se utilizarán aquellas en negrita:

#AUTHID	Identificador del autor del texto
STATUS	Texto en particular a analizar
sEXT	Grado de Extroversión
sNEU	Grado de Inestabilidad Emocional
sAGR	Grado de Amabilidad
sCON	Grado de Responsabilidad
sOPN	Grado de Apertura a Nuevas Experiencias
cEXT	Variable de clase para Extroversión
cNEU	Variable de clase para Inestabilidad Emocional
cAGR	Variable de clase para Amabilidad
cCON	Variable de Clase para Responsabilidad
cOPN	Variable de Clase para Apertura a Nuevas Experiencias

Tabla 4.6: Atributos dataset *MyPersonality*. Fuente: Elaboración propia.

4.4.3 *Dataset PAN15: Author Profiling*

En la competencia PAN15 (Rangel et al., 2015) realizada en el año 2015 en la categoría de perfil de autores, además de los datos demográficos habituales de los autores, como el género y la edad, se introducen cinco rasgos de personalidad correspondientes a *BigFive* (apertura, conciencia, extraversión, amabilidad y neuroticismo). En esta instancia se crean conjuntos de datos en 4 idiomas: inglés, español, alemán e italiano, siendo los dos primeros los relevantes para el presente estudio.

El conjunto de datos de *PAN15* es de carácter restringido a la cual se accede mediante solicitud expresa vía correo electrónico con motivo de investigación la cual fue otorgada de manera exitosa en primera instancia con fecha 9 de octubre de 2020.

#AUTHID	Identificador del autor del texto
STATUS	Texto en particular a analizar
sEXT	Grado de Extroversión
sNEU	Grado de Inestabilidad Emocional
sAGR	Grado de Amabilidad
sCON	Grado de Responsabilidad
sOPN	Grado de Apertura a Nuevas Experiencias
Age	Variable de clase para la edad en rango
Gender	Variable de clase para el sexo del autor del texto libre

Tabla 4.7: Atributos de dataset *Pan15: Author Profiling*. Fuente: Elaboración propia.

4.4.4 *Dataset* construido

En el marco del proyecto FONDEF IDEA código ID15I-10560: "Plataforma de Apoyo a la Gestión de Emergencia y Aplicaciones" se construye una encuesta para medir distintas dimensiones, en esta sección se abordan aquellos que son de interés para la investigación. Cabe destacar que los correos e información personal como nombre y correo son recogidos única y exclusivamente para asegurar la unicidad de las respuestas y estos no son ingresados ni procesados por el modelo expuesto, sino, se hace a través de un identificador de carácter no vinculante.

Datos socio-demográficos

De los datos recogidos en el instrumento son de interés para el experimento son los siguientes:

1. Sexo: Se pregunta de manera binaria hombre o mujer para ser utilizadas como variable de clase.
2. Edad: Se pregunta de manera cuantitativa, es decir, de manera numérica. Luego en el periodo de post-proceso estos datos son leídos y categorizados de acuerdo a los rangos del conjunto de datos de *PAN15* (Rangel et al., 2015): 18-24, 25-34, 35-49, 50+.

Rasgos de la personalidad

Para medir los rasgos de la personalidad de los encuestados se aplica *BigFiveInventory-44*, el cual consiste en 44 preguntas que son contestadas en una escala de *likert* de 1 a 5, luego estas respuestas son transformadas al rango correspondiente.

Respuestas de texto libre

Esta sección se agrega con motivo específico de esta investigación para poder verificar la hipótesis planteada relacionando la escritura con respecto a los rasgos de la

personalidad.

1. Describe desde tu perspectiva la importancia de la instancia de los voluntariados, además, de la participación de los voluntarios.
2. ¿Cuál es o fue tu principal motivación para participar como voluntario?
3. ¿Qué cualidades te gustaría que tuviesen las personas con las que trabajas en los voluntariados? Indique al menos tres y cuéntanos el porqué
4. ¿Cuál fue su rol en la última tarea que desempeñó en un voluntariado? Describa lo que más pueda
5. De acuerdo a las tareas que has desarrollado como voluntario ¿Cómo evalúas tus capacidades y cuáles son?
6. ¿Cómo describirías tu experiencia y actuar frente a situaciones que se presentan imprevistos o emergencias?
7. ¿Consideras a futuro realizar una nueva actividad de voluntariado? ¿Qué factores influyen en tu respuesta?

En el anexo B.2.2 se encuentra un extracto de respuestas a estas preguntas por parte de los voluntarios a los que se le aplicó el instrumento asegurando su anonimato.

4.5 REPRESENTACIÓN DE DATOS

En el campo del procesamiento de lenguaje formal se utiliza *embeddings* previamente entrenadas para las palabras en el idioma correspondiente, en este caso español (Bojanowski et al., 2017) e inglés (Mikolov et al., 2018). Los *Word Embeddings* proporcionan una representación numérica de la entrada de texto que se utiliza para modelar, sin embargo, asignar palabras a *embeddings* no siempre es sencillo: los datos pueden no estar ordenados o la representación puede no existir en la herramienta.

4.6 PREPROCESAMIENTO DE DATOS

Antes de entrenar la red es necesario verificar la integridad de los datos, que la representación de éstos sean la correctos, evitar duplicidad de entradas, identificar y eliminar

basura o palabras vacías (*stopwords*), para esto se realizan diversas técnicas que son detalladas a continuación, cabe destacar que cada técnica es especializada en algunos casos para el lenguaje correspondiente.

4.6.1 Medir cobertura

Para asegurar una correcta representación de los datos y que las palabras a utilizar posean su representación numérica de acuerdo al *WordEmbedding* utilizado se utilizan indicadores para el vocabulario y para el texto completo. En primera instancia se crea el vocabulario, para esto se recorre cada uno de los textos y posteriormente cada una de las palabras utilizadas en los textos creados los usuarios y se obtiene un *hash* donde la llave es la palabra utilizada y el valor es la frecuencia de uso de aquella palabra.

Para medir las coberturas se mapean todas las palabras existentes en el texto con su representación numérica que posee el *WordEmbedding* del lenguaje correspondiente y se verifica si estas palabras poseen representación o no en la estructura de datos, dando origen a 4 variables: palabras conocidas, número de palabras conocidas, palabras desconocidas, número de palabras desconocidas. Ver tabla 4.8.

Nombre	**Descripción**	**Tipo de Variable**	**Variable**
Vocabulario	Todas las palabras que se utilizan	Hash	Vocab
Palabras conocidas	Palabras que poseen representación en el WordEmbedding	Hash	known_words
Números de palabras conocidas	Frecuencia total de apariciones de palabras conocidas en todos los textos leídos	Entero	n_known_words
Palabras desconocidas	Palabras que no poseen representación en el WordEmbedding	Hash	unknown_words
Número de palabras desconocidas	Suma total de frecuencias de palabras que no poseen representación en el WordEmbedding	Entero	nb_unknown_words

Tabla 4.8: Variables utilizadas para la medición de cobertura de *WordEmbedding* en los textos. Fuente: Elaboración propia.

Cobertura Vocabulario

Para medir la cobertura del vocabulario se dividen la cantidad de palabras conocidas en la cantidad total del vocabulario que posee el texto, su valor representa el % de palabras que

poseen una representación numérica.

$$vocab\_coverage = \frac{len(known\_words)}{len(vocab)} \tag{4.1}$$

Los valores iniciales de cobertura de vocabulario son:

- Cobertura vocabulario en inglés: 40.983%
- Cobertura vocabulario en español: 40.271%

Cobertura Texto completo

Para medir la cobertura del texto completo se divide la cantidad de apariciones (frecuencia) de palabras conocidas en la suma de las frecuencias de palabras conocidas y frecuencia de palabras desconocidas

$$all\_text\_coverage = \frac{nb\_known\_words}{nb\_known\_words + nb\_unknown\_words} \tag{4.2}$$

Los valores iniciales de cobertura de texto son:

- Cobertura texto en inglés: 71.135%
- Cobertura texto en español: 72.829%

4.6.2 Contracciones

Para el algoritmo de FastText del *WordEmbedding* las contacciones son un problema, especialmente para el idioma inglés ya que existe una única representación para palabras y no así para sus contracciones, por ejemplo: *are not* y su contracción *aren't*, las cuales ambas se encuentran presentes en el texto pero solo la primer posee su respectiva representación, por lo tanto se realiza una conversión de estas contracciones mediante una lista creada con las contracciones más habituales ya que debido a la naturaleza de los datos y de la misma escritura humana se puede escribir de distintas formas, al igual que el uso de emoticones los cuales también son transformados debido a que éstos tampoco poseen representación en el *WordEmbedding*, por ejemplo, el emoticon ":)" es reemplazado por la palabra *"happy"* en inglés y por la palabra feliz en español. La lista completa de estas contracciones se encuentra en el código adjunto.

Los valores de cobertura después de procesar las contracciones:

- Cobertura vocabulario en inglés: 41.005%
- Cobertura de texto completo en inglés: 75.555%
- Cobertura vocabulario en español: 40.276%
- Cobertura de texto completo en español: 72.829%

4.6.3 Puntuaciones

La puntuación desconocida puede llegar a ser un problema, como lo son: "/-'?!., $%́()*+-/:;<=>@[]| ", aquellos caracteres especiales que no poseen representación, es decir, son reconocidos como desconocidos son limpiados. El listado completo de caracteres especiales se encuentra disponible en el código adjunto.

Los valores de cobertura después de limpiar las puntuaciones desconocidas:

- Cobertura vocabulario en inglés: 61.488%
- Cobertura de texto completo en inglés: 95.229%
- Cobertura vocabulario en español: 67.273%
- Cobertura de texto completo en español: 93.169%

4.6.4 Limpieza de menciones y direcciones web

Debido a la naturaleza de algunos conjuntos de datos como *MyPersonality* y los de *PAN15* provenientes de redes sociales como *Facebook* y *Twitter* es necesario limpiar sintaxis propia de estas plataformas como menciones y direcciones web (URLs), entre otros. Para esta tarea se utilizan expresiones regulares para identificar menciones (que posean @, en el caso de *PAN15* el nombre de los usuarios en las menciones es reemplazado por "username" para asegurar confidencialidad y para identificar las direcciones web se utiliza bajo la presencia de links que posean un protocolo (http, https, ftp). La implementación de este método se encuentra en el código anexo.

Los valores de cobertura después de limpiar menciones y direcciones web:

- Cobertura vocabulario en inglés: 61.521%

- Cobertura de texto completo en inglés: 94.935%
- Cobertura vocabulario en español: 64.394%
- Cobertura de texto completo en español: 91.682%

Hasta este punto se identifican palabras que simplemente no van a tener *embeddings*. Estas pueden ser debido a errores ortográficos, utilización de nombres propios u otros, sin embargo con las coberturas alcanzadas del texto es suficiente como para realizar la experimentación, ya que el subconjunto que no posee representación es referente a caracteres fuera de la codificación y errores de tipografía los cuales no inciden en el sentido de los párrafos analizados. Los resultados finales de cobertura demuestra que a pesar de tener solamente 6 de cada 10 elementos del vocabulario éstos representan aproximadamente un 95% de todo el texto, esto se puede interpretar como la aparición de caracteres desconocidos o de puntuación no reconocidos los cuales tan solo representan un 5% de todo el cuerpo del texto.

4.7 ARQUITECTURA DE MODELO

La arquitectura de la red consta de distintas capas basadas en la implementación de la herramienta de *deep learning* de Python: Keras, la cual corre sobre *TensorFlow*, la representación gráfica de la arquitectura a nivel de capas se encuentra en la figura 4.2 y las capas más relevantes se encuentran explicadas y resumidas en la presente sección de acuerdo a las entradas, salidas y parámetros que estos reciben en su implementación dando cuenta de la elección de estos en los casos que sea pertinente.

4.7.1 Capa de *Embeddings*

Esta capa posee un argumento de entrada de dos dimensiones: *(batch_size, input_length)* y de salida posee 3 dimensiones: *(batch_size, input_length, output_dim)*. Los parámetros de la instancia de la red utilizados son:

- *Input_dim*: Es un entero, corresponde al tamaño del vocabulario, es decir, índice de enteros máximo + 1. Para el caso de la red construida es: El largo del tokenizador + 1.
- *Output_dim*: Es un entero, corresponde a la dimensión del *embedding*.

- *Input_length*: Largo de las secuencias de entrada, cuando es constante. Este argumento es necesario si va a conectar *Flatten* y luego capas tipo *Dense* en sentido ascendente (sin él, no se puede calcular la forma de las salidas *Dense*). Para el caso de estudio el largo máximo de los textos es de 280, en caso de añadir nuevas entradas considerar este parámetro ya que para prevenir errores mediante una función los textos son truncados a este largo máximo.

4.7.2 Capa de abandono espacial 1D

La capa de abandono espacial de una dimensión o por su nombre en inglés *Spatial Dropout 1D* realiza la misma función que *Dropout*, sin embargo, elimina conjuntos de características 1D completos en lugar de elementos individuales. Si los elementos adyacentes dentro de los conjuntos de características están fuertemente correlacionados (como suele ser el caso en las primeras capas de convolución), el *dropout* regular no regularizará las activaciones y, de lo contrario, solo dará como resultado una disminución de la tasa de aprendizaje la cual es indicada como parámetro. En este caso, *SpatialDropout1D* favorece la independencia entre conjunto de características y por eso se decide incluirla en la arquitectura.

Esta capa posee un argumento de entrada de tres dimensiones: *samples, timesteps, channels* y su salida coincide con la entrada. Los parámetros de la instancia de la red utilizados son:

- *Rate*: Es de tipo flotante entre 0 y 1 y representa la proporción de elementos de entrada que serán "soltados"

4.7.3 Capa Bidireccional LSTM

Los *LSTM* bidireccionales son una extensión de los *LSTM* tradicionales que pueden mejorar el rendimiento del modelo en problemas de clasificación de secuencias.

En problemas donde todos los pasos de tiempo *(timestep)* de la secuencia de entrada están disponibles, los *LSTM* bidireccionales entrenan dos *LSTM* en lugar de uno en la secuencia de entrada. El primero en la secuencia de entrada tal cual y el segundo en una copia invertida de la secuencia de entrada. Esto puede proporcionar un contexto adicional a la red y dar como resultado un aprendizaje más rápido e incluso más completo sobre el problema.

En el caso del lenguaje resulta sumamente conveniente utilizar este tipo de redes ya que las palabras e incluso las oraciones completas, que al principio no significan nada, tienen sentido a la luz del contexto futuro. Lo que debemos recordar es la distinción entre las distintas palabras usadas y su eventual significado de acuerdo a las palabras que lo preceden y proceden, que requieren una salida después de cada entrada, y aquellas en las que las salidas solo se necesitan al final de algún segmento de entrada (Graves & Schmidhuber, 2005).

En el caso de la implementación en el modelo bidireccional admite una capa recurrente como argumento, en este caso se utiliza una capa *LSTM*, además se especifica el método de unión entre estas, que para este caso se utiliza el método de concatenación (por defecto, no es necesario explicitarlo), el cual proporciona el doble de salidas para la siguiente etapa.

Para el caso de la implementación de la capa *LSTM* que se encuentra envuelta por el modelo bidireccional se especifica con los siguientes argumentos:

- *Units*: Entero positivo, representa la dimensionalidad del espacio de salida, en este caso es de 64, Sin embargo al ser una capa bidireccional concatenada su salida es el doble, es decir 128. Ver figura 4.2.
- *Return_sequences*: Es de tipo booleano. En el caso que sea verdadero permite devolver la última salida.

4.7.4 Capa convolucional

Esta capa crea un núcleo *(kernel)* de convolución que se convoluciona con la entrada de la capa sobre una única dimensión espacial para producir un tensor [5] de salidas. En el caso de la implementación se utilizan los siguientes parámetros:

- *Filters*: Es un entero y representa la dimensionalidad del espacio de salida (es decir, el número de filtros de salida en la convolución). Para este caso es de 64.
- *Kernel_size*: Es un entero que especifica la longitud de la ventana de convolución 1D, se utiliza el valor recomendado = 2.
- *Padding*: Puede ser "válido", "mismo" o "causal". Para esta caso se utiliza el parámetro "válido" *(valid)* el cual no realiza un relleno a diferencia de "mismo" *(same)* que realiza un relleno uniforme para que la salida tenga la misma dimensión que la entrada.

[5] En matemáticas, un tensor es un objeto algebraico que describe una relación entre conjuntos de objetos algebraicos relacionados con un espacio vectorial

4.7.5 Capas de agrupación

Debido a que se utilizan capas LSTM, se podría usar el parámetro *return_sequences = false* en la última capa *LSTM*, esto es también una posibilidad en lugar de la capa de agrupación *(pooling layer)*. Sin embargo, esto mantiene solo el último paso de la secuencia. En este caso esto significaría una desventaja para nuestro modelo por que nos interesa mantener parte de la secuencia completa para analizar el contexto del uso de esta.

Debido a lo anterior se utilizan dos funciones que en su salida son concatenadas.

- *GlobalAveragePooling1D*: Es una operación de agrupación de promedio global para datos temporales. Resulta adecuado su uso si es importante la contribución de la secuencia completa para el resultado.
- *GlobalMaxPooling1D*: Es una operación de agrupación máxima que opera sobre datos temporales. Para la entrada toma el valor máximo sobre cada dimensión de tiempo. Resulta adecuado su uso cuando se desea detectar la presencia de "algo" en la secuencia

4.7.6 Capas de densidad

La capa de densidad implementa la operación: salida = activación (punto (entrada, kernel) + sesgo) donde la activación es la función de activación por elementos que se pasa como argumento de activación, el kernel es una matriz de pesos creada por la capa y el sesgo es un vector de sesgo creado por capa (solo se aplica si use_bias es *true*).

En este caso se poseen simplemente dos capas de densidad anexadas ambas con los siguientes parámetros:

- *Units*: Entero positivo, representa la dimensionalidad del espacio de salida, la primera capa es de 128 y la segunda es la mitad, es decir, 64.
- *Activations*: Función de activación a utilizar. Si no especifica nada, no se aplica ninguna activación (es decir, activación "lineal": a (x) = x), sin embargo, para este caso se ha utilizado una función de activación de Rectificación Lineal Unitaria o *ReLU* por sus siglas en inglés. Esta función se utiliza debido a que para utilizar descenso de gradiente estocástico con retropropagación *(backpropagation)* de errores para entrenar redes neuronales profundas, se necesita una función de activación que se vea y actúe como una función lineal, pero que sea, de hecho, una función no lineal que permita aprender relaciones complejas en los

datos. La activación lineal rectificada es la activación predeterminada cuando se desarrollan perceptrones multicapa y redes neuronales convolucionales.

4.7.7 Capa de salida

Los parámetros de e esta capa son sumamente importante ya que establecen la cantidad de elementos a predecir, para nuestra investigación corresponden a 5 (rasgos de la personalidad). Ver figura 4.2.

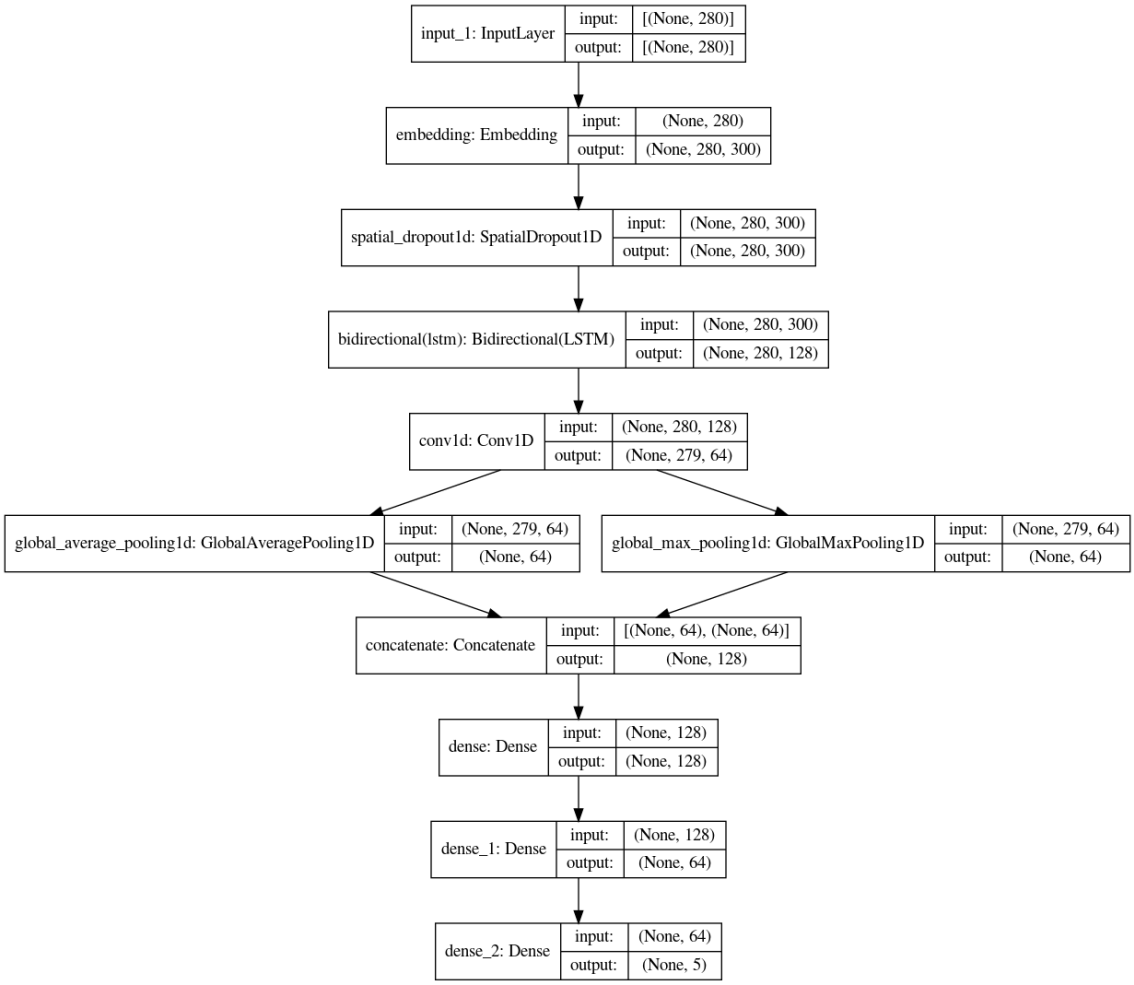


Figura 4.2: Arquitectura del modelo utilizado argumentos correspondientes a *embedding* en idioma inglés. Fuente: Elaboración propia.

4.8 ENTRENAMIENTO

La función de entrenamiento requiere definir un conjunto de parámetros, a continuación se explica el funcionamiento, las opciones disponibles y el parámetro utilizado en la presente investigación.

4.8.1 Tamaño de lote

Keras utiliza bibliotecas matemáticas en sus modelos como *TensorFlow* y *Theano* y una desventaja de utilizar estas las bibliotecas mencionadas es que la forma y el tamaño de sus datos deben definirse una vez a priori (por adelantado) y mantenerse constantes e independientemente de si está entrenando su red o haciendo predicciones.

El *batch size* o de ahora en adelante tamaño del lote, limita la cantidad de muestras (en nuestro caso texto) que se muestra a la red antes de que se pueda realizar una actualización de los pesos. Esta misma limitación se impone luego al hacer predicciones con el modelo de ajuste.

Existen distintos tipos de soluciones para abordar este desafío:

1. Aprendizaje en línea (tamaño de lote = 1): Aquí es donde el tamaño del lote se establece en un valor de 1 y los pesos de la red se actualizan después de cada ejemplo de entrenamiento.

   Esto puede tener el efecto de un aprendizaje más rápido, pero también agrega inestabilidad al proceso de aprendizaje, ya que los pesos varían ampliamente con cada lote.

2. Pronóstico por lotes *(Batch Forecasting)*: Otra solución es hacer todas las predicciones a la vez en un único lote. Esto significaría que podríamos estar muy limitados en la forma en que se usa el modelo.

   Tendríamos que usar todas las predicciones hechas a la vez, o solo mantener la primera predicción y descartar el resto.

   Podemos adaptar el ejemplo para la previsión por lotes prediciendo con un tamaño de lote igual al tamaño del lote de entrenamiento y luego enumerando el lote de predicciones.

3. Enfoque intermedio: Existe la posibilidad de lograr un parámetro intermedio que sea conveniente para el entrenamiento y la validación, es este último enfoque el que se utilizará para el estudio

El tamaño del lote define la cantidad de ejemplos que son propagadas en la red, por ejemplo: poseemos 1050 ejemplos de entrenamiento y utilizamos un tamaño de lote igual a 100, el algoritmo tomará los 100 primeros ejemplos (del 1° al 100°) del conjunto de datos de entrenamiento y la red será entrenada, los pesos serán ajustado y posteriormente la red tomará los siguientes 100 ejemplos (del 101° al 200°) y así sucesivamente sucederá el entrenamiento de la red, se puede realizar este procedimiento hasta haber propagado todos los ejemplos por la red, en el ejemplo expuesto 1050 no es divisible por 100 de manera exacta, la forma más sencilla es que los 50 ejemplos restantes entren directamente a la red y sea entrenada.

Una ventaja de utilizar un tamaño de lote < al número total de ejemplos es que requiere menos memoria. Dado que entrena la red con menos muestras, el procedimiento de entrenamiento general requiere menos memoria. Eso es especialmente importante si no puede colocar todo el conjunto de datos en la memoria de su máquina, ya que a mayor tamaño de lote, mayor memoria se requiere, además, esto permite que los pesos de la red se ajusten una mayor cantidad de veces.

Las redes se entrenan más rápido con lotes pequeños. Eso es porque actualizamos los pesos de la red después de cada propagación. En el ejemplo expuesto, se ha propagado 11 lotes (10 de ellos con 100 muestras y el último con 50 muestras) y, después de cada propagación se actualizan los parámetros de la red. Si usáramos todas las muestras durante la propagación, haríamos solo 1 actualización para el parámetro de la red.

4.8.2 Épocas

Según la definición en la documentación de *Keras* las épocas son un límite arbitrario, generalmente definido como "una pasada sobre todo el conjunto de datos", que se utiliza para separar el entrenamiento en distintas fases, que es útil para el registro y la evaluación periódica lo que resulta útil para el registro y la evaluación periódica.

Cuando se usa un conjunto de data de validación con un método de ajuste de los modelos en Keras, la evaluación se ejecuta al final de cada época. Dentro de Keras, existe la capacidad de agregar devoluciones *(callbacks)* de llamada diseñadas específicamente para ejecutarse al final de una época. Ejemplos de estos son los cambios en la tasa de aprendizaje, el modelo de puntos de control y métodos de parada anticipada.

Entonces, en otras palabras, varias épocas significa cuántas veces pasas por tu conjunto de entrenamiento.

El modelo se actualiza cada vez que se procesa un lote, lo que significa que se puede actualizar varias veces durante una época. Si tamaño del lote se establece igual a la longitud de

n, entonces el modelo se actualizará una vez por época.

4.8.3 Devolución de llamada

Una devolución de llamada o por su nombre en inglés *callback* es un objeto que puede realizar acciones en varias etapas del entrenamiento (por ejemplo, al comienzo o al final de una época, antes o después de un solo lote, etc.), estos son utilizados en el proyecto para mejorar los resultados y prevenir comportamientos indeseados.

ReduceLROnPlateau

Un *callback* de suma importancia utilizado es *ReduceLROnPlateau* el cual se añade para reducir la tasa de aprendizaje *(learning rate)* cuando una métrica ha dejado de mejorar.

Los modelos a menudo se benefician de la reducción de la tasa de aprendizaje en un factor de 2 a 10 una vez que el aprendizaje se estanca. Esta devolución de llamada monitorea una cantidad y si no se observa ninguna mejora en un número de épocas de "paciencia", la tasa de aprendizaje se reduce.

Para el caso del proyecto se utilizan los siguientes parámetros:

- Monitor: cantidad a monitorizar, *monitor='val_loss'*.
- Factor: factor por el cual se reducirá la tasa de aprendizaje. $new\_lr = lr * factor$, *factor = 0.1*.
- Paciencia: número de épocas sin mejora después de las cuales se reducirá la tasa de aprendizaje, *patience = 10*.
- Modo: Puede ser: *'auto', 'min', 'max'*. En el modo 'min', la tasa de aprendizaje se reduce cuando la cantidad monitoreada haya dejado de disminuir; en modo 'max' se reducirá cuando la cantidad monitoreada haya dejado de aumentar; en el modo 'automático', la dirección se infiere automáticamente del nombre de la cantidad monitoreada. Se utiliza 'auto'.
- Límite de taza de aprendizaje: límite inferior de la tasa de aprendizaje: $min\_lr = 1.1e^{-6}$.

ModelCheckpoint

Esta devolución de llamada o *callback* se utiliza para guardar el modelo de *Keras* y los pesos del modelo con cierta frecuencia.

La devolución de llamada *ModelCheckpoint* se usa junto con el entrenamiento usando el método *model.fit()* para guardar un modelo o pesos (en un archivo llamado best_model.hdf5) en algún intervalo, por lo que el modelo o los pesos se pueden cargar más tarde para continuar el entrenamiento desde el estado guardado o incluso para divulgarlo para su futura utilización.

Los parámetros utilizados para este *callback* son:

- Monitor: el nombre de la métrica que se va a supervisar. Normalmente, las métricas se establecen mediante el método Model.compile. Se utiliza *'val_loss'*
- Guardar el mejor: parámetro *save_best_only = true*, se encarga de solo guardar cuando el modelo se considera el mejor, de acuerdo a los parámetros definidos.
- Modo: puede ser *auto, min, max*. En el modo *min*, la tasa de aprendizaje se reducirá cuando la cantidad monitoreada haya dejado de disminuir, por otro lado, en modo *max* se reducirá cuando la cantidad monitoreada haya dejado de aumentar y en el modo *automático*, la dirección se infiere automáticamente del nombre de la cantidad monitoreada. Se utiliza *mode = min*.

4.8.4 Función de pérdida

El propósito de las funciones de pérdida es calcular la cantidad que un modelo debe buscar minimizar durante el entrenamiento. Para efectos de la investigación se utiliza pérdida por error absoluta media *(mean absolute error MAE)*

En algunos problemas de regresión, la distribución de la variable objetivo a pesar de ser principalmente gaussiana, pero puede tener valores atípicos o en nuestro caso valores tanto positivos como negativos (escala de personalidad desde -0,5 a 0,5) por lo cual resulta ventajoso obtener el valor absoluto de esas diferencias ya que esta diferencia corresponde al error que al sumarlo puede darse el caso que estos errores se vean disminuidos o incluso anulados.

$$loss = |y_{verdadero} - y_{predecido}| \tag{4.3}$$

Se calcula como el promedio de la diferencia absoluta entre los valores reales y pronosticados. En el modelo se utiliza el parámetro para la función de pérdida "mean_absolute_error".

4.8.5 Optimizador

Un optimizador es uno de los dos argumentos necesarios, junto con el de función de pérdida, para compilar cualquier modelo, para efectos de nuestro experimento se utiliza el optimizador que implementa el algoritmo de Adam.

La optimización de Adam es un método de descenso de gradiente estocástico que se basa en la estimación adaptativa de momentos de primer y segundo orden. de acuerdo a (Kingma & Ba, 2017) El método es "computacionalmente eficiente, requiere poca memoria, invariante al reajuste diagonal de gradientes y es muy adecuado para problemas que son grandes en términos de datos / parámetros", es por esto, que es el optimizador que mejor se adecúa a las necesidades de la experiencia.

CAPÍTULO 5. RESULTADOS Y CONCLUSIONES FINALES

Se trabaja con 3 conjuntos de datos de los cuales 2 pertenecen a la competencia de *PAN15* (Rangel et al., 2015) en los cuales se disponibilizan mediante solicitud expresa para investigación en 4 idiomas distintos inglés, español, alemán e italiano, para efectos de la presente investigación son de vital importancia los dos primeros. Además, se ha construido un conjunto de datos propio con el contexto nacional haciendo especial énfasis a el escenario de voluntariados. El resumen de los conjuntos de datos separados en conjunto de entrenamiento y de prueba se encuentra en la tabla 5.1 con la información más relevante de los datos que se poseen de cada conjunto. Los valores de los rasgos de la personalidad se encuentran normalizados entre 0.1 y 1.

	Entrenamiento		Prueba		Validación
	Inglés PAN15	Español PAN15	Inglés PAN15	Español PAN15	Español Elab. Propia
Usuarios	152	110	142	88	44
18-24	58	22	56	18	6
24-34	60	56	58	44	20
35-49	22	22	20	18	16
50+	12	10	8	8	2
Hombre	76	55	71	44	20
Mujer	76	55	71	44	24
Ext (prom)	0.66	0.68	0.67	0.66	0.56
STA (prom)	0.64	0.57	0.63	0.69	0.65
AGR (prom)	0.62	0.64	0.64	0.64	0.54
CON (prom)	0.67	0.74	0.67	0.71	0.67
OPN (prom)	0.74	0.68	0.76	0.69	0.60

Tabla 5.1: Resumen conjunto de datos utilizados para la etapa de experimentación de acuerdo a los distintos idiomas, distribución de edades, sexo y rasgos de la personalidad. Fuente: Elaboración propia.

5.1 CONJUNTO DE DATOS EN INGLÉS *PAN15*

El conjunto de datos obtenidos en (Rangel et al., 2015) posee un conjunto de datos de entrenamiento para el cual se posee información de 152 usuarios sumando un total de 14.166 entradas de texto obteniendo un promedio de 93.2 entradas de texto por cada usuario y otro conjunto de datos de prueba del cual se posee información de 142 usuarios sumando un total de 13.179 entradas de texto obteniendo un promedio de 92.8 entradas de texto por cada usuario.

De acuerdo a lo expuesto en el capítulo 4.8.1 el tamaño del lote o *batch size* es el parámetro de mayor interés para el estudio de *word embeddings* dado el impacto que éste produce al ajustar los pesos de la red de acuerdo y el impacto que posee en el aprendizaje. Se

utilizarán distintos valores de parámetros para el tamaño del lote y analizar el impacto para cada uno de los lenguajes utilizados. Se utilizará 128, 512, 4096 y el tamaño de lote máximo donde es igual a la cantidad de entradas de texto.

5.1.1 Tamaño del lote 128

El tiempo de ejecución del entrenamiento es de 512 segundos para la ejecución de 39 épocas ya que a pesar de tener como objetivo 300 épocas posee un llamado de parada anticipada si la tasa de aprendizaje no mejora lo suficiente luego de ciertas iteraciones para prevenir un posible sobre aprendizaje.

Dado que el tamaño del lote es de 128 y el tamaño total de entradas para el conjunto de entrenamiento es de 14166 por cada época se ajustan los parámetros de la red 111 veces (14166/128).

Al finalizar la época 39 el valor de la función de pérdida para el conjunto de entrenamiento es de 0.11604, mientras que el valor de la función de pérdida para el conjunto de prueba es de 0.18612, en el figura 5.1 se puede ver la evolución por época de ambos resultados.

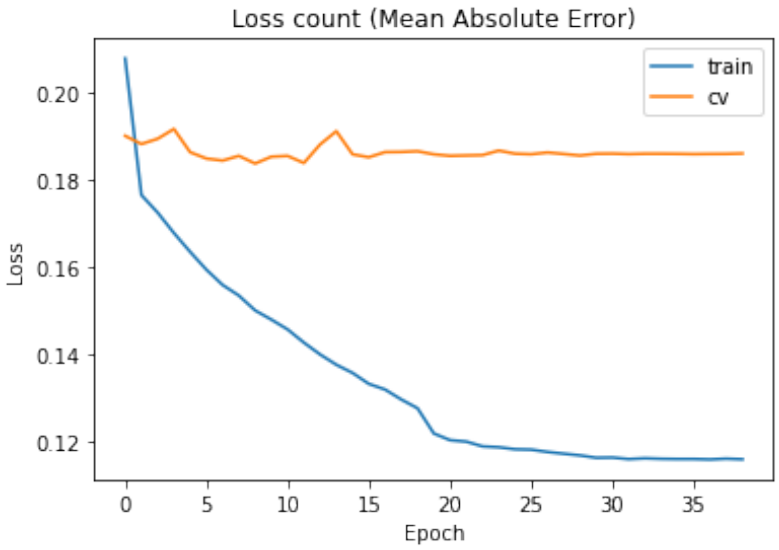


Figura 5.1: Evolución de error absoluto medio por época para el conjunto de entrenamiento y de validación para el conjunto de datos en inglés de *PAN15* con un tamaño de lote de 128. Fuente: Elaboración propia.

.

Por otro lado la tasa de aprendizaje comienza en 0.001 en las primeras iteraciones y finaliza en $1x10^{-6}$, al ser la tasa de aprendizaje lo suficientemente baja se ejecuta el llamado de parada anticipada. En la figura 5.2 se aprecia su evolución dadas las épocas correspondientes.

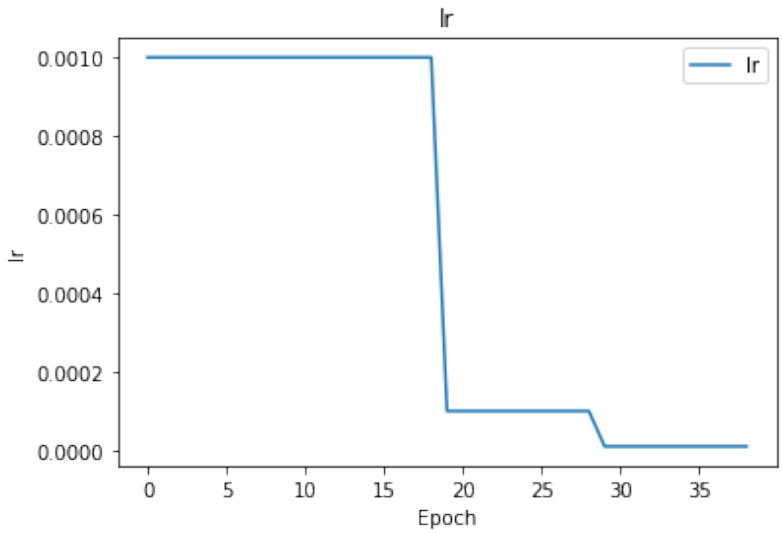


Figura 5.2: Evolución de la tasa de aprendizaje por época para el conjunto de datos en inglés de *PAN15* con un tamaño de lote de 128. Fuente: Elaboración propia.

.

Los valores de predicción varían para cada uno de los rasgos de personalidad, el promedio de la suma del error absoluto obtenido por el modelo para cada uno de los rasgos de la personalidad se encuentran en la tabla 5.2

	EXT	NEU	AGR	CON	OPN
Error absoluto medio	0.15983	0.24869	0.15983	0.19105	0.24794

Tabla 5.2: Resumen de los errores absolutos medios obtenidos para cada uno de los 5 rasgos de la personalidad para el conjunto de datos en inglés de *PAN15* con un tamaño de lote de 128. Fuente: Elaboración propia.

5.1.2 Tamaño del lote 512

El tiempo de ejecución del entrenamiento es de 485 segundos para la ejecución de 41 épocas ya que a pesar de tener como objetivo 300 épocas posee un llamado de parada anticipada si la tasa de aprendizaje no mejora lo suficiente luego de ciertas iteraciones para prevenir un posible sobre aprendizaje.

Dado que el tamaño del lote es de 512 y el tamaño total de entradas para el conjunto de entrenamiento es de 14166 por cada época se ajustan los parámetros de la red 28 veces (14166/512).

Al finalizar la época 41 el valor de la función de pérdida para el conjunto de entrenamiento es de 0.13785, mientras que el valor de la función de pérdida para el conjunto de

prueba es de 0.18546, en el figura 5.3 se puede ver la evolución por época de ambos resultados.

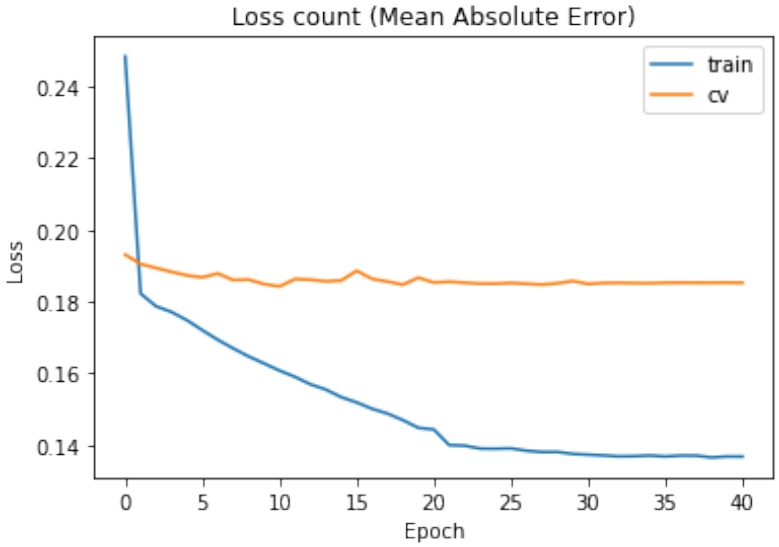


Figura 5.3: Evolución de error absoluto medio por época para el conjunto de entrenamiento y de validación para el conjunto de datos en inglés de *PAN15* con un tamaño de lote de 512. Fuente: Elaboración propia.

.

Por otro lado la tasa de aprendizaje comienza en 0.001 en las primeras iteraciones y finaliza en $1x10^{-6}$, al ser la tasa de aprendizaje lo suficientemente baja se ejecuta el llamado de parada anticipada. En la figura 5.4 se aprecia su evolución dadas las épocas correspondientes.

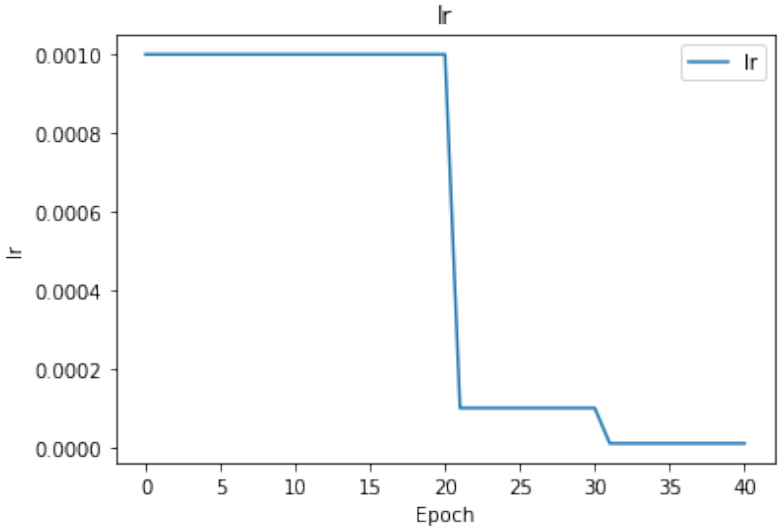


Figura 5.4: Evolución de la tasa de aprendizaje por época para el conjunto de datos en inglés de *PAN15* con un tamaño de lote de 512. Fuente: Elaboración propia.

.

Los valores de predicción varían para cada uno de los rasgos de personalidad, el

promedio de la suma del error absoluto obtenido por el modelo para cada uno de los rasgos de la personalidad se encuentran en la tabla 5.3

	EXT	NEU	AGR	CON	OPN
Error absoluto medio	0.15973	0.24425	0.15842	0.19030	0.24481

Tabla 5.3: Resumen de los errores absolutos medios obtenidos para cada uno de los 5 rasgos de la personalidad para el conjunto de datos en inglés de *PAN15* con un tamaño de lote de 512. Fuente: Elaboración propia.

5.1.3 Tamaño del lote 4096

El tiempo de ejecución del entrenamiento es de 404 segundos para la ejecución de 92 épocas ya que a pesar de tener como objetivo 300 épocas posee un llamado de parada anticipada si la tasa de aprendizaje no mejora lo suficiente luego de ciertas iteraciones para prevenir un posible sobre aprendizaje.

Dado que el tamaño del lote es de 4096 y el tamaño total de entradas para el conjunto de entrenamiento es de 14166 por cada época se ajustan los parámetros de la red 4 veces (14166/4096).

Al finalizar la época 92 el valor de la función de pérdida para el conjunto de entrenamiento es de 0.15358, mientras que el valor de la función de pérdida para el conjunto de prueba es de 0.17873, en el figura 5.5 se puede ver la evolución por época de ambos resultados.

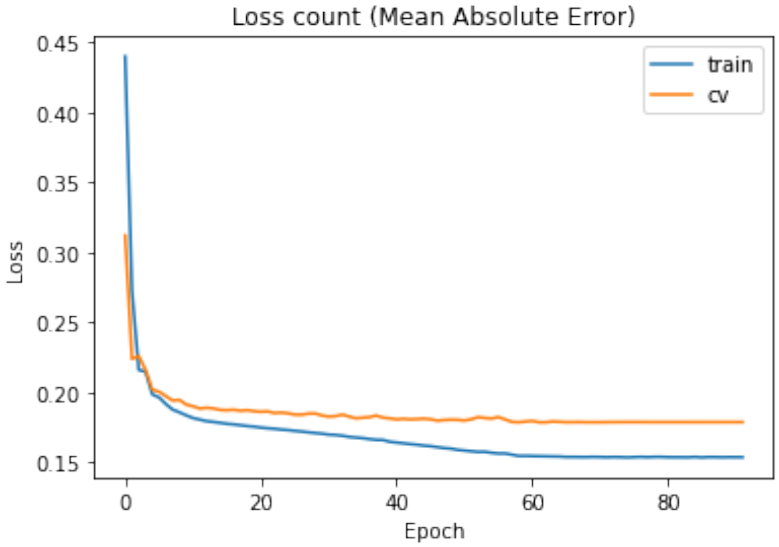


Figura 5.5: Evolución de error absoluto medio por época para el conjunto de entrenamiento y de validación para el conjunto de datos en inglés de *PAN15* con un tamaño de lote de 4096. Fuente: Elaboración propia.

.

Por otro lado la tasa de aprendizaje comienza en 0.001 en las primeras iteraciones y finaliza en $1x10^{-6}$, al ser la tasa de aprendizaje lo suficientemente baja se ejecuta el llamado de parada anticipada. En la figura 5.6 se aprecia su evolución dadas las épocas correspondientes.

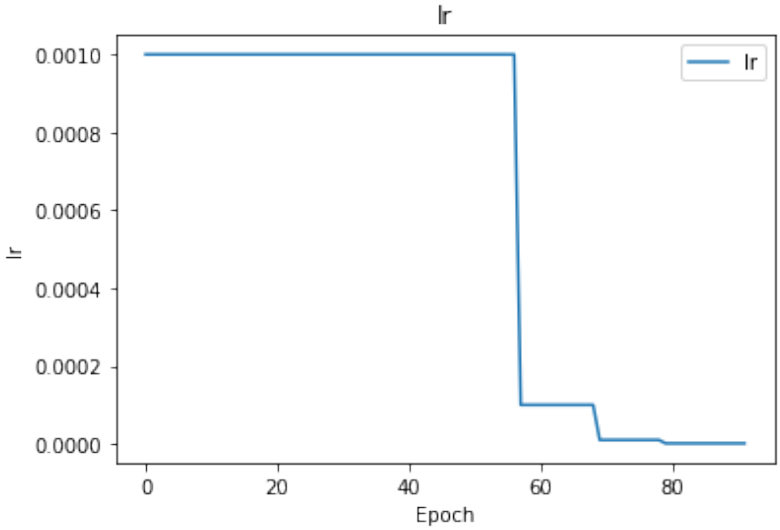


Figura 5.6: Evolución de la tasa de aprendizaje por época para el conjunto de datos en inglés de *PAN15* con un tamaño de lote de 4096. Fuente: Elaboración propia.

.

Los valores de predicción varían para cada uno de los rasgos de personalidad, el promedio de la suma del error absoluto obtenido por el modelo para cada uno de los rasgos de la personalidad se encuentran en la tabla 5.4

	EXT	NEU	AGR	CON	OPN
Error absoluto medio	0.16416	0.24490	0.15085	0.18265	0.24056

Tabla 5.4: Resumen de los errores absolutos medios obtenidos para cada uno de los 5 rasgos de la personalidad para el conjunto de datos en inglés de *PAN15* con un tamaño de lote de 4096. Fuente: Elaboración propia.

5.1.4 Tamaño del lote 14166

El tiempo de ejecución del entrenamiento es de 1504 segundos para la ejecución de 68 épocas ya que a pesar de tener como objetivo 300 épocas posee un llamado de parada anticipada si la tasa de aprendizaje no mejora lo suficiente luego de ciertas iteraciones para prevenir un posible sobre aprendizaje.

Dado que el tamaño del lote es de 14166 y el tamaño total de entradas para el conjunto de entrenamiento es de 14166 por cada época se ajustan los parámetros de la red 1 vez

(14166/14166).

Al finalizar la época 68 el valor de la función de pérdida para el conjunto de entrenamiento es de 0.18867, mientras que el valor de la función de pérdida para el conjunto de prueba es de 0.19261, en el figura 5.7 se puede ver la evolución por época de ambos resultados.

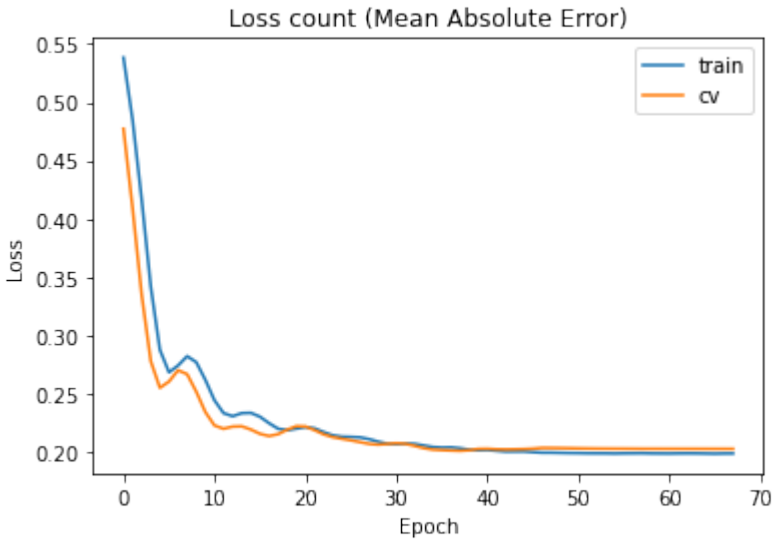


Figura 5.7: Evolución de error absoluto medio por época para el conjunto de entrenamiento y de validación para el conjunto de datos en inglés de *PAN15* con un tamaño de lote de 14166. Fuente: Elaboración propia.

.

Por otro lado la tasa de aprendizaje comienza en 0.001 en las primeras iteraciones y finaliza en $1x10^{-6}$, al ser la tasa de aprendizaje lo suficientemente baja se ejecuta el llamado de parada anticipada. En la figura 5.8 se aprecia su evolución dadas las épocas correspondientes.

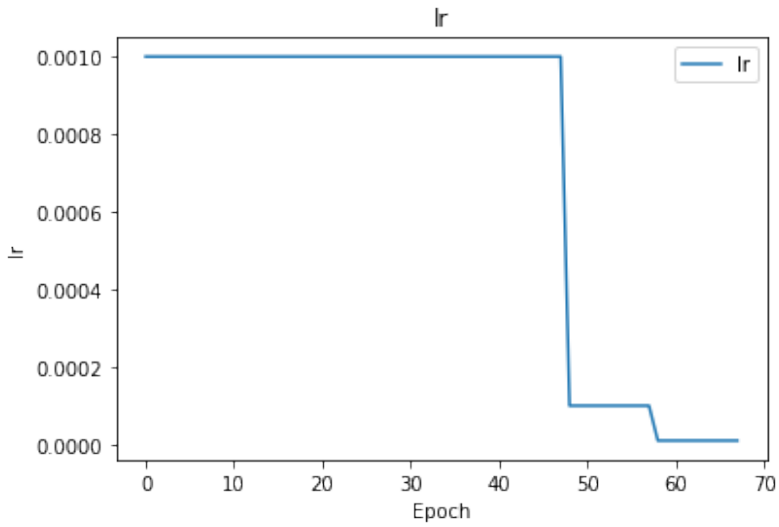


Figura 5.8: Evolución de la tasa de aprendizaje por época para el conjunto de datos en inglés de *PAN15* con un tamaño de lote de 14166. Fuente: Elaboración propia.

.

Los valores de predicción varían para cada uno de los rasgos de personalidad, el promedio de la suma del error absoluto obtenido por el modelo para cada uno de los rasgos de la personalidad se encuentran en la tabla 5.5

	EXT	NEU	AGR	CON	OPN
Error absoluto medio	0.15820	0.24247	0.14939	0.18099	0.23229

Tabla 5.5: Resumen de los errores absolutos medios obtenidos para cada uno de los 5 rasgos de la personalidad para el conjunto de datos en inglés de *PAN15* con un tamaño de lote de 14166. Fuente: Elaboración propia.

El resumen de los resultados por lote para cada rasgo de la personalidad, los resultados en conjuntos de entrenamiento y validación y tiempo obtenidos en los experimentos con el conjunto de datos en inglés se encuentra en la tabla 5.6.

	RASGOS DE LA PERSONALIDAD					Resultados		
LOTE	EXT	NEU	AGR	CON	OPN	Entren.	Validación	T(s)
128	0.15983	0.24869	0.15983	0.19105	0.24794	0.11604	0.18612	512
512	0.15973	0.24425	0.15842	0.19030	0.24481	0.13785	0.18546	485
4096	0.16416	0.24490	0.15085	0.18265	0.24056	0.15358	0.17873	404
14166	0.15820	0.24247	0.14939	0.18099	0.23229	0.18867	0.19261	1504

Tabla 5.6: Resumen de errores absolutos medios obtenidos para el conjunto de datos en inglés para los rasgos de la personalidad de manera general para cada uno de los tamaños de los lotes de los experimentos. Fuente: Elaboración propia.

5.2 CONJUNTO DE DATOS EN ESPAÑOL

El conjunto de datos obtenidos en (Rangel et al., 2015) posee un conjunto de datos de entrenamiento para el cual se posee información de 110 usuarios sumando un total de 9.879 entradas de texto obteniendo un promedio de 89.8 entradas de texto por cada usuario y otro conjunto de datos de prueba del cual se posee información de 88 usuarios sumando un total de 8.609 entradas de texto obteniendo un promedio de 97.8 entradas de texto por cada usuario.

De acuerdo a lo expuesto en el capítulo 4.8.1 el tamaño del lote o *batch size* es el parámetro de mayor interés para el estudio de *word embeddings* dado el impacto que éste produce al ajustar los pesos de la red de acuerdo y el impacto que posee en el aprendizaje. Se utilizarán distintos valores de parámetros para el tamaño del lote y analizar el impacto para cada uno de los lenguajes utilizados. Se utilizará 128, 512, 4096 y el tamaño de lote máximo donde es igual a la cantidad de entradas de texto.

5.2.1 Tamaño del lote 128

El tiempo de ejecución del entrenamiento es de 1.785 segundos para la ejecución de 35 épocas ya que a pesar de tener como objetivo 300 épocas posee un llamado de parada anticipada si la tasa de aprendizaje no mejora lo suficiente luego de ciertas iteraciones para prevenir un posible sobre aprendizaje.

Dado que el tamaño del lote es de 128 y el tamaño total de entradas para el conjunto de entrenamiento es de 9879 por cada época se ajustan los parámetros de la red 78 veces (9879/128).

Al finalizar la época 35 el valor de la función de pérdida para el conjunto de entrenamiento es de 0.15762, mientras que el valor de la función de pérdida para el conjunto de prueba es de 0.20043, en el figura 5.9 se puede ver la evolución por época de ambos resultados.

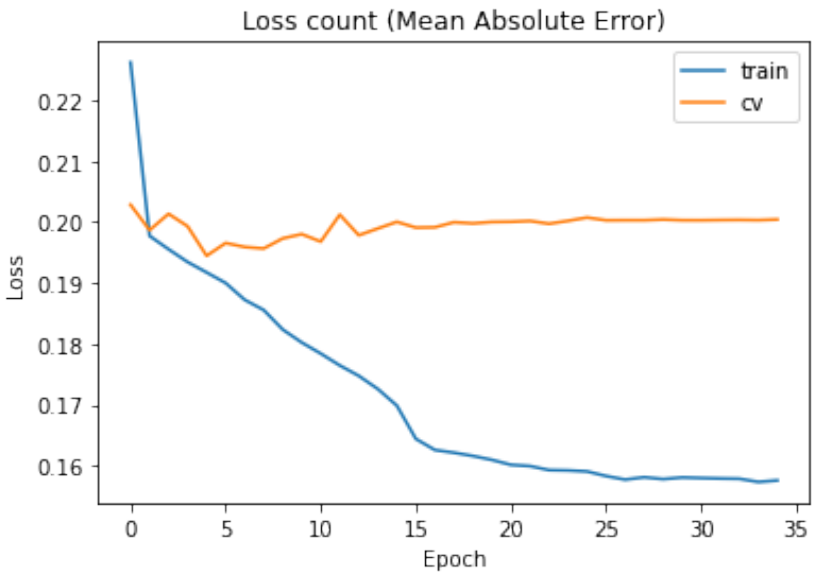


Figura 5.9: Evolución de error absoluto medio por época para el conjunto de entrenamiento y de validación para el conjunto de datos en español de *PAN15* con un tamaño de lote de 128. Fuente: Elaboración propia.

.

Por otro lado la tasa de aprendizaje comienza en 0.001 en las primeras iteraciones y finaliza en $1x10^{-6}$, al ser la tasa de aprendizaje lo suficientemente baja se ejecuta el llamado de parada anticipada. En la figura 5.10 se aprecia su evolución dadas las épocas correspondientes.

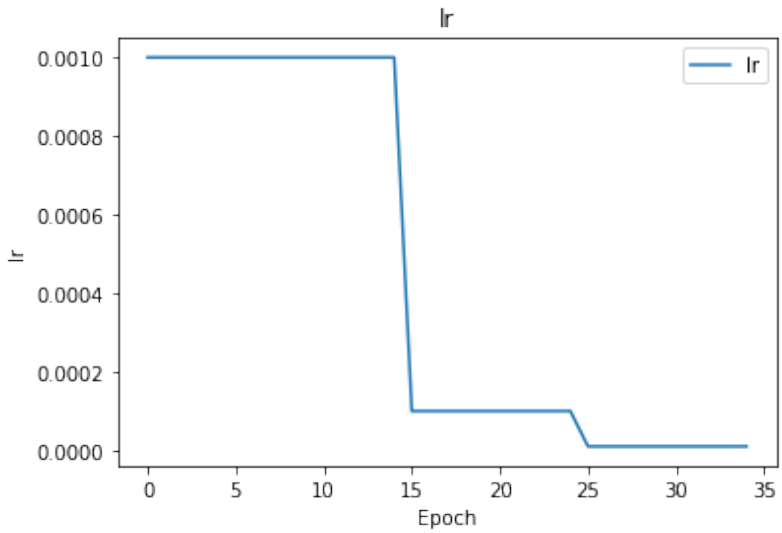


Figura 5.10: Evolución de la tasa de aprendizaje por época para el conjunto de datos en español de *PAN15* con un tamaño de lote de 128. Fuente: Elaboración propia.

.

Los valores de predicción varían para cada uno de los rasgos de personalidad, el promedio de la suma del error absoluto obtenido por el modelo para cada uno de los rasgos de la

personalidad se encuentran en la tabla 5.7

	EXT	NEU	AGR	CON	OPN
Error absoluto medio	0.18576	0.24005	0.21292	0.20270	0.22907

Tabla 5.7: Resumen de los errores absolutos medios obtenidos para cada uno de los 5 rasgos de la personalidad para el conjunto de datos en español de *PAN15* con un tamaño de lote de 128. Fuente: Elaboración propia.

5.2.2 Tamaño del lote 512

El tiempo de ejecución del entrenamiento es de 1.259 segundos para la ejecución de 43 épocas ya que a pesar de tener como objetivo 300 épocas posee un llamado de parada anticipada si la tasa de aprendizaje no mejora lo suficiente luego de ciertas iteraciones para prevenir un posible sobre aprendizaje.

Dado que el tamaño del lote es de 512 y el tamaño total de entradas para el conjunto de entrenamiento es de 9879 por cada época se ajustan los parámetros de la red 20 veces (9879/512).

Al finalizar la época 43 el valor de la función de pérdida para el conjunto de entrenamiento es de 0.17165, mientras que el valor de la función de pérdida para el conjunto de prueba es de 0.19901, en el figura 5.11 se puede ver la evolución por época de ambos resultados.

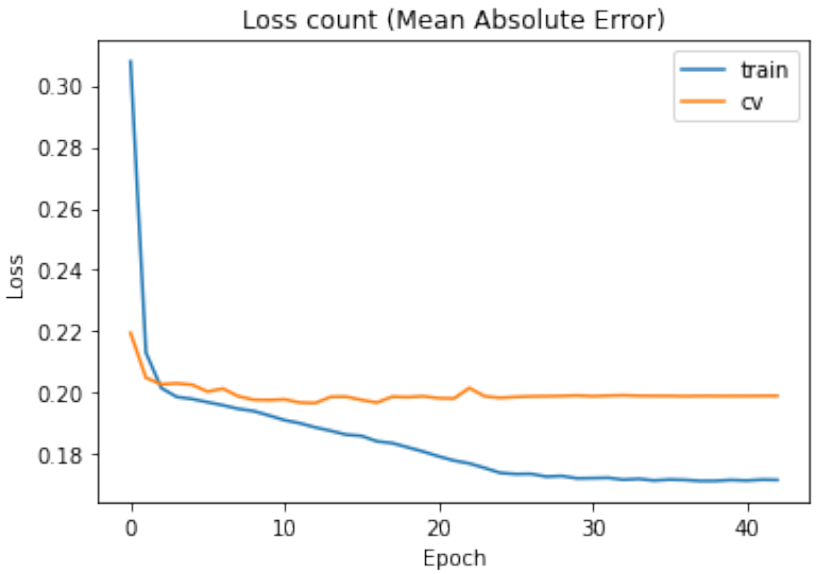


Figura 5.11: Evolución de error absoluto medio por época para el conjunto de entrenamiento y de validación para el conjunto de datos en español de *PAN15* con un tamaño de lote de 512. Fuente: Elaboración propia.

.

Por otro lado la tasa de aprendizaje comienza en 0.001 en las primeras iteraciones y finaliza en $1x10^{-6}$, al ser la tasa de aprendizaje lo suficientemente baja se ejecuta el llamado de parada anticipada. En la figura 5.12 se aprecia su evolución dadas las épocas correspondientes.

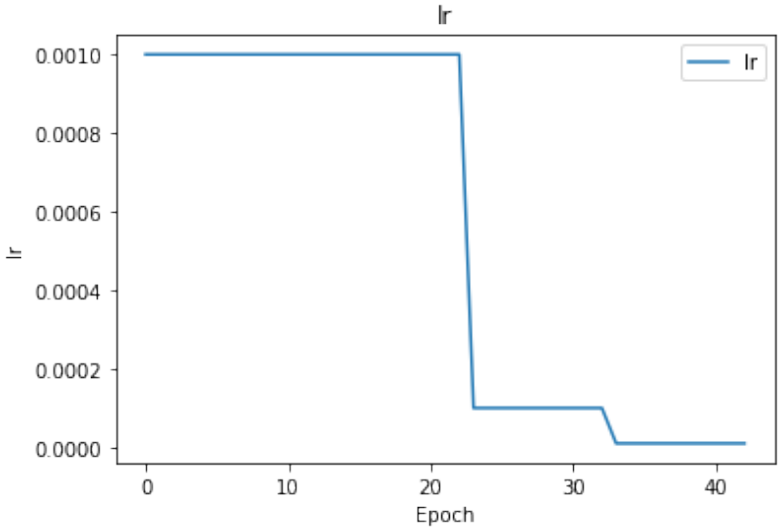


Figura 5.12: Evolución de la tasa de aprendizaje por época para el conjunto de datos en español de *PAN15* con un tamaño de lote de 512. Fuente: Elaboración propia.

.

Los valores de predicción varían para cada uno de los rasgos de personalidad, el promedio de la suma del error absoluto obtenido por el modelo para cada uno de los rasgos de la personalidad se encuentran en la tabla 5.8

	EXT	NEU	AGR	CON	OPN
Error absoluto medio	0.18719	0.23638	0.21080	0.20037	0.22769

Tabla 5.8: Resumen de los errores absolutos medios obtenidos para cada uno de los 5 rasgos de la personalidad para el conjunto de datos en español de *PAN15* con un tamaño de lote de 512. Fuente: Elaboración propia.

5.2.3 Tamaño del lote 4096

El tiempo de ejecución del entrenamiento es de 1598 segundos para la ejecución de 94 épocas ya que a pesar de tener como objetivo 300 épocas posee un llamado de parada anticipada si la tasa de aprendizaje no mejora lo suficiente luego de ciertas iteraciones para prevenir un posible sobre aprendizaje.

Dado que el tamaño del lote es de 4096 y el tamaño total de entradas para el conjunto de entrenamiento es de 9879 por cada época se ajustan los parámetros de la red 3

veces (9879/4096).

Al finalizar la época 94 el valor de la función de pérdida para el conjunto de entrenamiento es de 0.18005, mientras que el valor de la función de pérdida para el conjunto de prueba es de 0.19692, en el figura 5.13 se puede ver la evolución por época de ambos resultados.

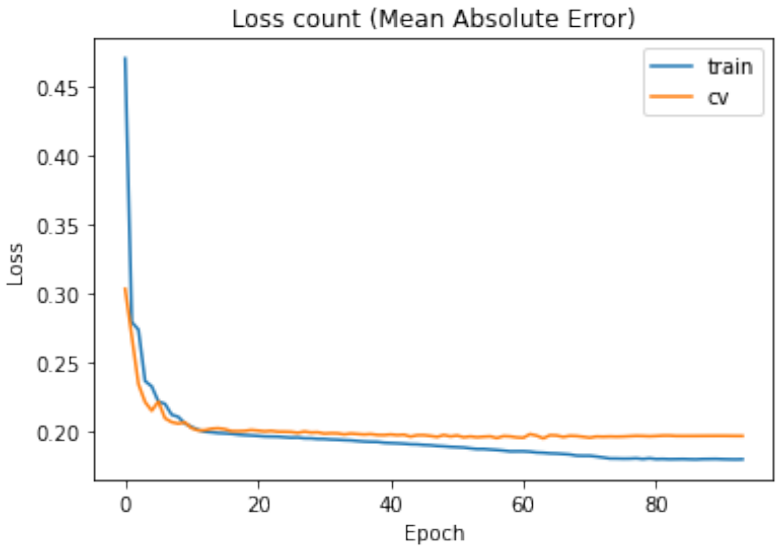


Figura 5.13: Evolución de error absoluto medio por época para el conjunto de entrenamiento y de validación para el conjunto de datos en español de *PAN15* con un tamaño de lote de 4096. Fuente: Elaboración propia.

.

Por otro lado la tasa de aprendizaje comienza en 0.001 en las primeras iteraciones y finaliza en $1x10^{-6}$, al ser la tasa de aprendizaje lo suficientemente baja se ejecuta el llamado de parada anticipada. En la figura 5.14 se aprecia su evolución dadas las épocas correspondientes.

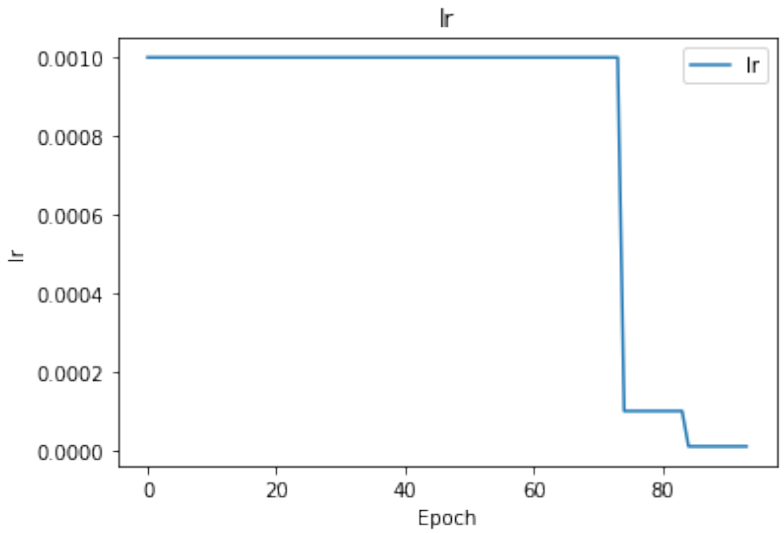


Figura 5.14: Evolución de la tasa de aprendizaje por época para el conjunto de datos en español de *PAN15* con un tamaño de lote de 4096. Fuente: Elaboración propia.

.

Los valores de predicción varían para cada uno de los rasgos de personalidad, el promedio de la suma del error absoluto obtenido por el modelo para cada uno de los rasgos de la personalidad se encuentran en la tabla 5.8

	EXT	NEU	AGR	CON	OPN
Error absoluto medio	0.18505	0.24002	0.21764	0.19579	0.23515

Tabla 5.9: Resumen de los errores absolutos medios obtenidos para cada uno de los 5 rasgos de la personalidad para el conjunto de datos en español de *PAN15* con un tamaño de lote de 4096. Fuente: Elaboración propia.

5.2.4 Tamaño del lote 9879

El tiempo de ejecución del entrenamiento es de 1116 segundos para la ejecución de 61 épocas ya que a pesar de tener como objetivo 300 épocas posee un llamado de parada anticipada si la tasa de aprendizaje no mejora lo suficiente luego de ciertas iteraciones para prevenir un posible sobre aprendizaje.

Dado que el tamaño del lote es de 9879 y el tamaño total de entradas para el conjunto de entrenamiento es de 9879 por cada época se ajustan los parámetros de la red 1 veces (9879/9879).

Al finalizar la época 94 el valor de la función de pérdida para el conjunto de entrenamiento es de 0.19874, mientras que el valor de la función de pérdida para el conjunto de

prueba es de 0.20172, en el figura 5.15 se puede ver la evolución por época de ambos resultados.

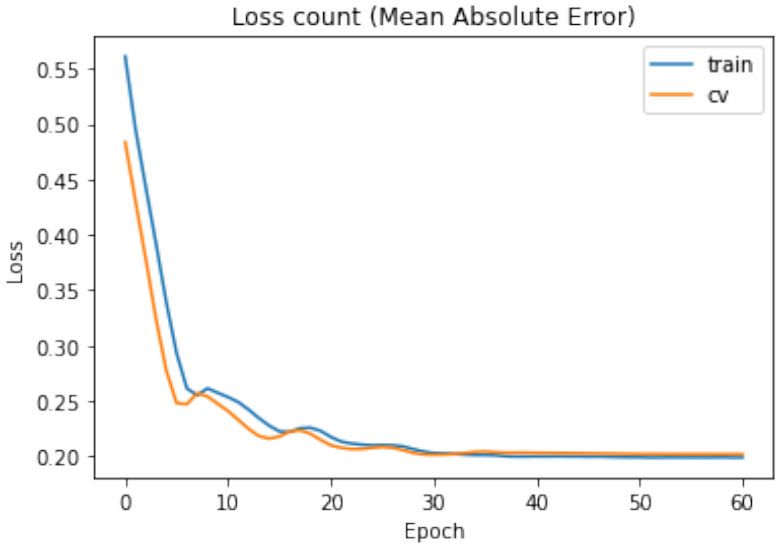


Figura 5.15: Evolución de error absoluto medio por época para el conjunto de entrenamiento y de validación para el conjunto de datos en español de *PAN15* con un tamaño de lote de 9879. Fuente: Elaboración propia.

.

Por otro lado la tasa de aprendizaje comienza en 0.001 en las primeras iteraciones y finaliza en $1x10^{-6}$, al ser la tasa de aprendizaje lo suficientemente baja se ejecuta el llamado de parada anticipada. En la figura 5.16 se aprecia su evolución dadas las épocas correspondientes.

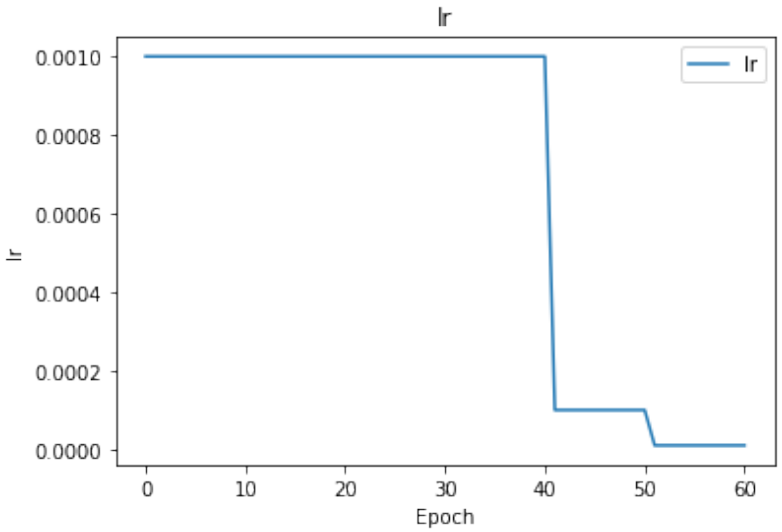


Figura 5.16: Evolución de la tasa de aprendizaje por época para el conjunto de datos en español de *PAN15* con un tamaño de lote de 9879. Fuente: Elaboración propia.

.

Los valores de predicción varían para cada uno de los rasgos de personalidad, el

promedio de la suma del error absoluto obtenido por el modelo para cada uno de los rasgos de la personalidad se encuentran en la tabla 5.10

	EXT	NEU	AGR	CON	OPN
Error absoluto medio	0.19083	0.23773	0.21230	0.19623	0.23794

Tabla 5.10: Resumen de los errores absolutos medios obtenidos para cada uno de los 5 rasgos de la personalidad para el conjunto de datos en español de *PAN15* con un tamaño de lote de 9879. Fuente: Elaboración propia.

El resumen de los resultados por lote para cada rasgo de la personalidad, los resultados en conjuntos de entrenamiento y validación y tiempo obtenidos en los experimentos con el conjunto de datos en español se encuentra en la tabla 5.11.

	RASGOS DE LA PERSONALIDAD					Resultados		
LOTE	EXT	NEU	AGR	CON	OPN	Entren.	Validación	T(s)
128	0.18576	0.24005	0.21292	0.20270	0.22907	0.15762	0.20043	1785
512	0.18719	0.23638	0.21080	0.20037	0.22769	0.17165	0.19901	1259
4096	0.18505	0.24002	0.21764	0.19579	0.23515	0.18005	0.19692	1598
9879	0.19083	0.23773	0.21230	0.19623	0.23794	0.19874	0.20172	1116

Tabla 5.11: Resumen de errores absolutos medios obtenidos para el conjunto de datos en español para los rasgos de la personalidad de manera general para cada uno de los tamaños de los lotes de los experimentos. Fuente: Elaboración propia.

5.3 CONJUNTO DE DATOS DE ELABORACIÓN PROPIA

El conjunto de datos obtenidos a través de las encuestas realizadas corresponde a las respuestas del *BFI-44* además de la información entregada con textos libres, estas 310 nuevas entradas son agregadas y utilizadas como conjunto de validación, es decir, no se utilizan dentro de la etapa de entrenamiento ni de pruebas, sino que son utilizados para validar los resultados obtenidos de acuerdo a los experimentos con distintos tamaños de lotes.

5.3.1 Tamaño del lote de 128

La tabla de diferencias entre los resultados obtenidos con el conjunto de prueba correspondiente a PAN15 en comparación a los resultados obtenidos con el conjunto de datos de elaboración propia se encuentran expresados mediante el error absoluto medio en la tabla 5.12, mientras que la diferencia de éstos de manera visual se puede ver en la figura 5.17.

	Elab. Propia	PAN15	Diferencia
EXT	0.28496	0.18576	0.09921
NEU	0.26386	0.24005	0.02381
AGR	0.24653	0.21292	0.03361
CON	0.25878	0.20270	0.05608
OPN	0.23847	0.22907	0.00940

Tabla 5.12: Diferencias en error medio absoluto entre conjunto de elaboración propio y conjunto de pruebas de *PAN15* en idioma español con un tamaño de lote de 128. Fuente: Elaboración propia.

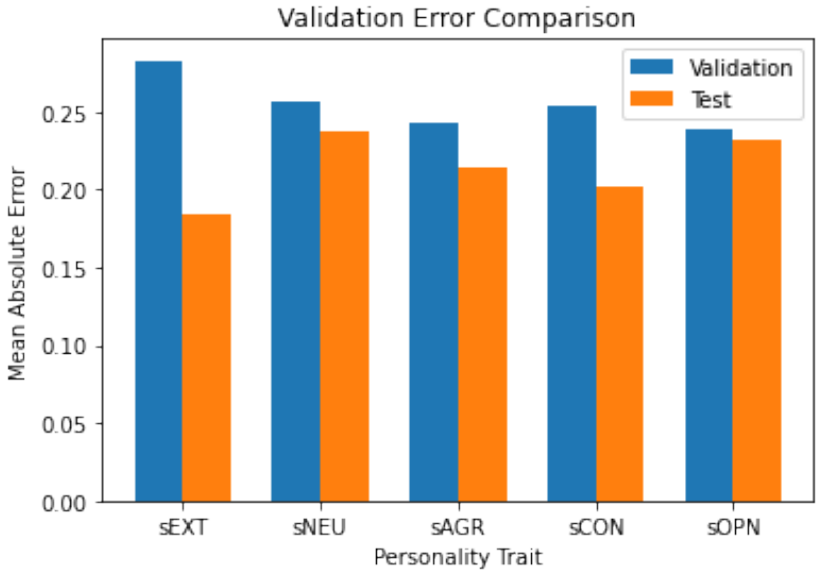


Figura 5.17: Diferencias en error medio absoluto entre conjunto de elaboración propio y conjunto de pruebas de *PAN15* en idioma español con un tamaño de lote de 128. Fuente: Elaboración propia.

.

5.3.2 Tamaño del lote de 512

La tabla de diferencias entre los resultados obtenidos con el conjunto de prueba correspondiente a PAN15 en comparación a los resultados obtenidos con el conjunto de datos de elaboración propia se encuentran expresados mediante el error absoluto medio en la tabla 5.13, mientras que la diferencia de éstos de manera visual se puede ver en la figura 5.18.

	Elab. Propia	PAN15	Diferencia
EXT	0.28433	0.18719	0.09714
NEU	0.26297	0.23638	0.02660
AGR	0.24613	0.21080	0.03532
CON	0.25548	0.20037	0.05511
OPN	0.24052	0.22770	0.01282

Tabla 5.13: Diferencias en error medio absoluto entre conjunto de elaboración propio y conjunto de pruebas de *PAN15* en idioma español con un tamaño de lote de 512. Fuente: Elaboración propia.

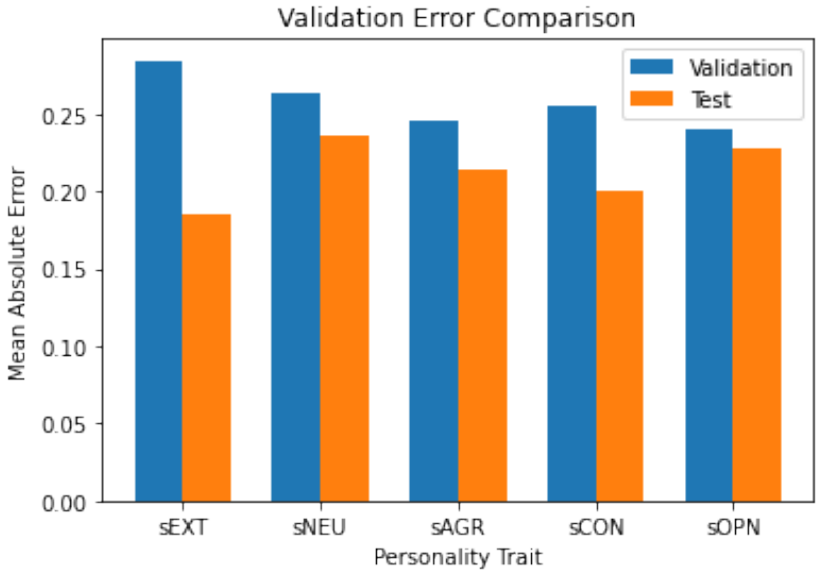


Figura 5.18: Diferencias en error medio absoluto entre conjunto de elaboración propio y conjunto de pruebas de *PAN15* en idioma español con un tamaño de lote de 512. Fuente: Elaboración propia.

.

5.3.3 Tamaño del lote de 4096

La tabla de diferencias entre los resultados obtenidos con el conjunto de prueba correspondiente a PAN15 en comparación a los resultados obtenidos con el conjunto de datos de elaboración propia se encuentran expresados mediante el error absoluto medio en la tabla 5.14, mientras que la diferencia de éstos de manera visual se puede ver en la figura 5.19.

	Elab. Propia	PAN15	Diferencia
EXT	0.28464	0.18505	0.09960
NEU	0.25925	0.24002	0.01923
AGR	0.24601	0.21764	0.02836
CON	0.25520	0.19579	0.05940
OPN	0.23461	0.23515	-0.00054

Tabla 5.14: Diferencias en error medio absoluto entre conjunto de elaboración propio y conjunto de pruebas de *PAN15* en idioma español con un tamaño de lote de 4096. Fuente: Elaboración propia.

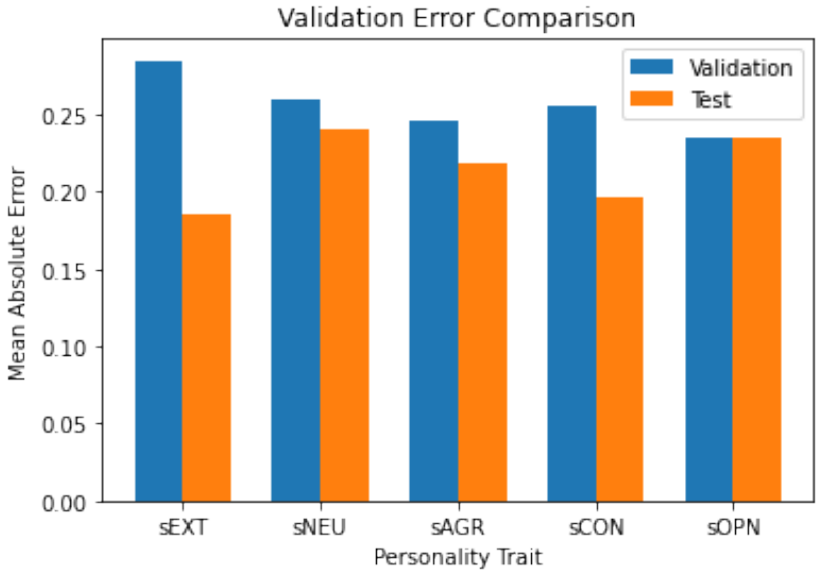


Figura 5.19: Diferencias en error medio absoluto entre conjunto de elaboración propio y conjunto de pruebas de *PAN15* en idioma español con un tamaño de lote de 4096. Fuente: Elaboración propia.

.

5.3.4 Tamaño del lote de 9879

La tabla de diferencias entre los resultados obtenidos con el conjunto de prueba correspondiente a PAN15 en comparación a los resultados obtenidos con el conjunto de datos de elaboración propia se encuentran expresados mediante el error absoluto medio en la tabla 5.15, mientras que la diferencia de éstos de manera visual se puede ver en la figura 5.20.

	Elab. Propia	PAN15	Diferencia
EXT	0.27817	0.19083	0.08734
NEU	0.25895	0.23773	0.02122
AGR	0.22572	0.21299	0.01273
CON	0.25475	0.19623	0.05852
OPN	0.23612	0.23794	-0.00183

Tabla 5.15: Diferencias en error medio absoluto entre conjunto de elaboración propio y conjunto de pruebas de *PAN15* en idioma español con un tamaño de lote de 9879. Fuente: Elaboración propia.

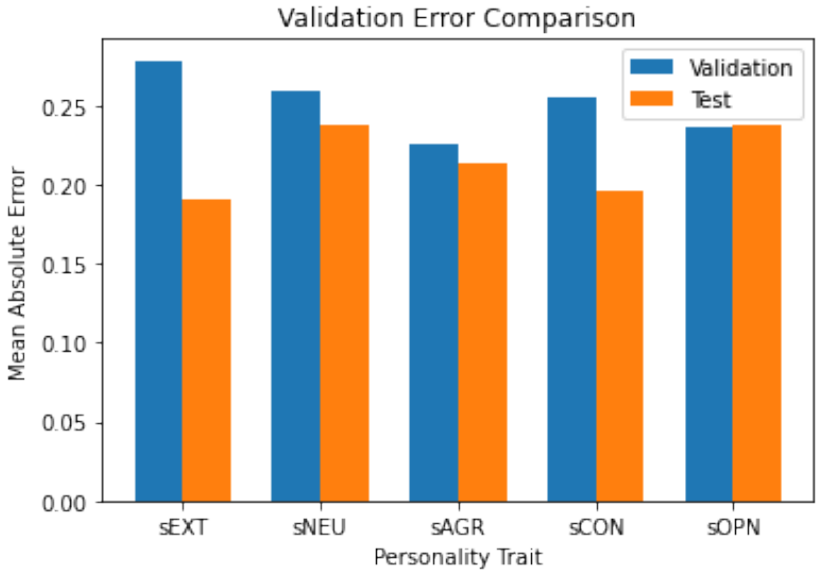


Figura 5.20: Diferencias en error medio absoluto entre conjunto de elaboración propio y conjunto de pruebas de *PAN15* en idioma español con un tamaño de lote de 9879. Fuente: Elaboración propia.

.

5.4 DISCUSIÓN

De acuerdo a los experimentos expuestos anteriormente se muestran las principales diferencias respecto a las diferencias que existen frente a la elección respecto a los tamaños de lote y el impacto que esta decisión posee sobre el entrenamiento y posterior validación con los conjuntos de datos dados.

Para el caso de los experimentos realizados con tamaños de lotes pequeños tanto para el conjunto de datos en inglés con los tamaños de 128 (ver figura 5.1) y 512 (ver figura 5.3) como para el caso del conjunto de datos en español con los tamaños de 128 (ver figura

5.9) y 512 (ver figura 5.11) se obtienen errores bajos para el conjunto de entrenamiento pero el error del conjunto de validación no decrece en las épocas manteniéndose estable y sobre el error del conjunto de entrenamiento, es decir, es un caso de *overfitting* o sobre-ajuste ya que en estos escenarios la curva de entrenamiento decae rápidamente mientras que la curva para el conjunto de validación no lo hace de igual manera incluso se estabiliza sin obtener mejoras consistentes a medida que avanzan las épocas, además, si se analiza la tasa de aprendizaje para estos experimentos de tamaño 128 (ver figuras 5.2 y 5.10) y 512 (ver figuras 5.4 y 5.12) el aprendizaje decae antes de la época 20 para el caso del tamaño de lote 128 y antes de la época 25 para el caso del tamaño del lote de 512.

El *overfitting* se vuelve indeseable ya que el problema es que cuanto más especializado se vuelve el modelo para los datos de entrenamiento, es menos capaz de generalizar y predecir el comportamiento frente a nuevos datos, lo que resulta en un aumento en el error de generalización, esto quiere decir que las predicciones de personalidades futuras a obtener frente a nuevos voluntarios no serán de buena calidad debido a que no forman parte del conjunto de datos de entrenamiento original, como solución a esto se puede considerarse un punto de parada anticipada antes que ocurra este fenómeno, previniendo que la red memorice los datos, sin embargo en los experimentos con tamaños de lote de 128 y 512 el sobre-ajuste ocurre en épocas muy tempranas así que es por esto que se concluye que no es recomendable la utilización de tamaños de lotes muy pequeños, ya que como se explica en la sección 4.8.1 los pesos de la red se actualizan dependiendo del tamaño del conjunto de entrenamiento, ejemplificando el número este fenómeno con el conjunto en inglés y con un tamaño de lote de 128 considerando el conjunto de entrenamiento de 14166 entradas, los parámetros de la red se ajustan 111 veces por cada época y para el caso del conjunto en español con un tamaño de 9879 entradas, los parámetros de la red se ajustan 78 veces por cada época lo cual le brinda inestabilidad a la red lo que produce que se encuentre con mínimos locales sin la posibilidad de salir de ellos.

En el conjunto de datos en español con un tamaño de lote de 9879 (ver figura 5.15) se observa en épocas tempranas que al actualizar solo una vez los pesos de la red el error presente al finalizar la primera actualización de parámetros es cercana a 48% y a medida que con cada época se van actualizando los parámetros de la red se da con ello una disminución sostenida respecto al error presente tanto en el conjunto de entrenamiento como en el conjunto de validación obteniendo un buen ajuste ya que la pérdida de entrenamiento y validación disminuye hasta un punto de estabilidad con una brecha mínima entre los dos valores finales de pérdida, con un error de conjunto de entrenamiento de 0.19874 y 0.20172 para el conjunto de validación en el conjunto de datos del idioma español. En este caso la brecha de generalización (la diferencia entre el entrenamiento y la curva de aprendizaje de pérdida de validación es estrecho y se cumple

que el error del conjunto de pruebas es menor que el de validación, por otro lado, los resultados obtenidos en el conjunto de datos en inglés presentes en la tabla 5.6 y los resultados obtenidos en el conjunto de datos en español presentes en la tabla 5.11 denotan resultados similares en rangos similares, siendo en ambos casos el rasgo de la personalidad de la extraversión (EXT) el que obtiene un menor error con 0.16416 y 0.19083 respectivamente, mientras que la apertura al conocimiento (OPN) el que presenta un mayor error tanto para el idioma inglés como para el idioma español con 0.24056 y 0.23794 respectivamente.

5.5 CONCLUSIONES FINALES

De acuerdo a los datos de la tabla 2.6, los resultados obtenidos de la investigación se encuentran dentro del estado del arte sin embargo no mejoran las técnicas expuestas en uno de las investigaciones con mejores resultados de la competencia PAN15 (Rangel et al., 2015) donde se utilizan técnicas de estilos estilométricos y agrupaciones de funciones junto a *Support Vector Machine Regression* (SVMr) (Grivas et al., 2015), sin embargo, se demuestra que la técnica sí permite clasificar voluntarios de acuerdo a su personalidad a partir de texto libre superando la precisión en un 70% para los 5 rasgos de la personalidad tanto para el idioma español como para el inglés.

Por otro lado, a pesar que los resultados obtenidos son más precisos para el idioma inglés en comparación al idioma español, la diferencia de precisión para el rasgo de la personalidad extraversión posee una diferencia 2.667% de precisión a favor del idioma inglés, mientras que para el caso del rasgo de la personalidad apertura al conocimiento esta diferencia de precisión es de tan solo 0.262% favorable al idioma español, lo cual permite concluir que no existen diferencias significativas de más de un 10% en los resultados para ambos idiomas estudiados con los experimentos expuestos y ocupando los *wordembedding* correspondientes para cada idioma.

Dicho lo anterior, se concluye que sí es posible detectar rasgos de la personalidad a través del análisis de respuestas de texto libre con una precisión promedio para el idioma inglés de 80.739% (ver tabla 5.6) mientras que para el idioma español es una precisión promedio de 79.828% (ver tabla 5.11) y por lo tanto la presente investigación presenta un aporte para las entidades que trabajen con voluntarios ya que este modelo permite obtener mayor información de los postulantes utilizando únicamente las respuestas de textos libres permitiendo clasificar su personalidad con un porcentaje de precisión que se encuentra dentro del estado del arte, sin embargo, el umbral de aceptación de validez dependerán de cada organización de acuerdo a la tolerancia del margen de error. Por otro lado, la interpretación de estos grados de personalidad

son parte del campo de la psicología y en caso de ser utilizado como variable de clases (introvertido o extrovertido) necesitan un puntaje de corte el cual depende del uso y de la organización que desee utilizar el instrumento, de esta manera la toma de decisiones se verá favorecida de acorde a las necesidades de cada institución.

Otro aporte a destacar es el tiempo que la aplicación de este modelo ahorra en comparación a la aplicación del instrumento original del *BFI-40* el cual son 40 preguntas y someter a los voluntarios a un instrumento específico, mientras que el modelo presentado solo requiere de tiempo para ser entrenado, una vez se supera esta fase el tiempo de respuesta es de segundos permitiendo así tener mayor información de los participantes al instante y sin someterlos a un examen complementario ya que se puede utilizar como entrada los textos libres correspondientes a las postulaciones de ellos. Es posible entender este modelo como una parte de un sistema mayor que favorezca la toma de decisiones para voluntariados, un sistema donde en tiempos de paz se registren los voluntarios y llenen su información personal junto al instrumento propio de la organización de modo que el modelo procese los formularios con texto libre de los voluntarios y permita obtener información acerca de la personalidad de ellos, de modo que en tiempos de emergencias la asignación de tareas alimente de datos sobre las tareas que ejecutó el voluntario junto a una retroalimentación de qué tan bien lo hizo generando así datos que permitan estudiar una correlación entre la ejecución correcta de tareas y ciertos rasgos de la personalidad que vean favorezcan esas tareas específicas.

A medida que el tiempo ha avanzado el interés por el estudio del comportamiento humano no ha dejado de crecer, es por esto que diversos estudios han tomado como base los rasgos de la personalidad de los 5 grandes dando origen a nuevos y sofisticadas clasificaciones más precisas, es así como en 2007 cada uno de los 5 rasgos de la personalidad se dividen en 2 facetas originando 10 nuevas facetas las cuales fueron propuestas por los psicólogos Colin DeYoung, Lena Quilty y Jordan Peterson en su artículo, "Entre facetas y dominios: 10 aspectos de los cinco grandes" (DeYoung et al., 2007). Además, el equipo de investigadores de Judge (Judge et al., 2013) se basó en el modelo de 10 facetas, dividiendo cada faceta aún más en 30 sub-facetas (ver figura B.4). Estas sub-facetas se basan en un modelo conocido como el *NEO Personality Inventory*, que fue desarrollado por primera vez en 1978 y revisado en 1990 por los psicólogos Paul Costa y Robert McRae. Estos estudios permiten vislumbrar la posibilidad de trabajos futuros de análisis de texto libre ya no sobre los 5 grandes, sino sobre 10 facetas o 30 sub-facetas permitiendo así una mayor cantidad de variables a considerar lo cual con el mejoramiento continuo de las redes neuronales pueden llevar a estudiar el problema con una mayor profundidad acompañado de una mayor cantidad de datos y variables disponibles, incluso añadir variables de carácter socio-demográfico como la edad o el sexo del postulante.

GLOSARIO

Big Five: Es un modelo que ha ido que busca explicar la personalidad del individuo a través de cinco categorías. Este modelo ha evolucionando con el tiempo a través de autores como (Fiske, 1949), (Norman, 1963) y (Goldberg, 1993). Las personalidades descritas en el modelo son:

- *Openness*: En una alta probabilidad describe personas creativas, con imaginación, abierta a probar nuevas cosas. En una baja probabilidad, describe personas que no le gustan los cambios, no intentan cosas nuevas, resistentes a nuevas ideas.
- *Conscientiousness*: En una alta probabilidad describe personas meticulosas, que se preparan para nuevas tareas y ponen atención a los detalles. En una baja probalidad describe personas que no le gustan la planificación y estructura, fallan en completar tareas.
- *Extraversion*: Caracteriza a personas sociales, emocionales y asertivas. En una baja probabilidad describe a personas solitarias, que piensan antes de hablar.
- *Agreeableness*: Incluye a personas que confiables, altruistas, amables y empáticas. Los individuos que presentan una baja probabilidad describen personas poco preocupadas por el resto, que no tienen interés en los problemas de los demás.
- *Neuroticism*: El neuroticismo es un rasgo caracterizado por la tristeza, el mal humor y la inestabilidad emocional.

Deep Learning: El aprendizaje profundo es un conjunto de algoritmos de *machine learning* que modela abstracciones a través de arquitecturas computacionales de muchas capas que admiten transformaciones no lineales múltiples e iterativas.

Convolutional Neural Network (CNN): Red neuronal convolucional es un tipo de red neuronal artificial es una variación de un perceptron multicapa, sin embargo, debido a que su aplicación es realizada en matrices bidimensionales, son muy efectivas para tareas como en la clasificación y segmentación de imágenes, entre otras aplicaciones.

REFERENCIAS BIBLIOGRÁFICAS

Barrick, M. R. (1991). THE BIG FIVE PERSONALITY DIMENSIONS AND JOB PERFORMANCE: A META-ANALYSIS. *Personnel Psychology*, *44*(1), 1–26.
URL https://doi.org/10.1111%2Fj.1744-6570.1991.tb00688.x

Bawa, G. S., Pakira, K., & Sharma, S. (2018). Extraction and use of personality traits from written commentary. *2018 International Conference on Computational Science and Computational Intelligence (CSCI)*.

Bengio, Y., Ducharme, R., Vincent, P., & Janvin, C. (2003). A neural probabilistic language model. *J. Mach. Learn. Res.*, *3*(null), 1137–1155.

Blázquez, E. (2019). Los 5 Grandes Factores de la Personalidad.
URL https://epsibapsicologia.es/los-5-grandes-factores-de-la-personalidad/

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information.

Celli, F., Pianesi, F., Stillwell, D., & Kosinski, M. (2013). Workshop on computational personality recognition: Shared task.

Collobert, R., & Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, (p. 160–167). New York, NY, USA: Association for Computing Machinery.
URL https://doi.org/10.1145/1390156.1390177

DeYoung, C. G., Quilty, L. C., & Peterson, J. B. (2007). Between facets and domains: 10 aspects of the big five. *Journal of Personality and Social Psychology*, *93*(5), 880–896.
URL https://doi.org/10.1037/0022-3514.93.5.880

Fiske, D. W. (1949). Consistency of the factorial structures of personality ratings from different sources. *The Journal of Abnormal and Social Psychology, 44(3), 329–344*.

Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American Psychologist, 48(1), 26–34*.

Goldberg, Y. (2017). Neural network methods for natural language processing. *Synthesis Lectures on Human Language Technologies*, *10*(1), 1–309.
URL https://doi.org/10.2200/s00762ed1v01y201703hlt037

González, L. (2004). Factores psicológicos asociados a la permanencia y compromiso del voluntariado. *Revista de Psicología*, *13*(2), pág. 21–41.
URL https://revistahistoriaindigena.uchile.cl/index.php/RDP/article/view/17652

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press. http://www.deeplearningbook.org.

Graves, A., & Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, *18*(5-6), 602–610.
URL https://doi.org/10.1016/j.neunet.2005.06.042

Grivas, A., Krithara, A., & Giannakopoulos, G. (2015). Author profiling using stylometric and structural feature groupings. In G. Jones, L. Cappellato, N. Ferro, & E. San Juan (Eds.) *Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum, Toulouse, France, September 8-11, 2015.*, CEUR Workshop Proceedings. CEUR-WS. Null ; Conference date: 08-09-2015 Through 11-09-2015.
URL http://clef2015.clef-initiative.eu/cfl.php

INJUV (2018). Sondeo n°3: Sondeo voluntariado de jóvenes 2018. Último acceso: 18-05-2020.
URL http://www.injuv.gob.cl/storage/docs/Resultados_Sondeo_03__Voluntariado_de_jovenes.pdf

John, O. P., Donahue, E. M., & Kentle, R. L. (1991). Big five inventory.
URL https://doi.org/10.1037/t07550-000

John, S., O. P. Srivastava (1999). The big five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of personality: Theory and research*, (pp. 102–138).
URL https://doi.org/10.1073/pnas.1218772110

Judge, T. A., Rodell, J. B., Klinger, R. L., Simon, L. S., & Crawford, E. R. (2013). Hierarchical representations of the five-factor model of personality in predicting job performance: Integrating three organizing frameworks with two theoretical perspectives. *Journal of Applied Psychology*, *98*(6), 875–925.
URL https://doi.org/10.1037/a0033901

Kelly, D. (2007). Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends® in Information Retrieval*, *3*(1—2), 1–224.
URL https://doi.org/10.1561/1500000012

Kingma, D. P., & Ba, J. (2017). Adam: A method for stochastic optimization.

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, *110*(15), 5802–5805.
URL https://doi.org/10.1073/pnas.1218772110

Lavell, A. (2009). Apuntes para una reflexión institucional en países de la subregión andina sobre el enfoque de la gestión del riesgo predecan. Último acceso: 03-01-2020.
URL http://www.comunidadandina.org/predecan/doc/r1/docAllan2.pdf

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

LeCun, Y., Haffner, P., Bottou, L., & Bengio, Y. (1999). *Object Recognition with Gradient-Based Learning*, (pp. 319–345). Berlin, Heidelberg: Springer Berlin Heidelberg.
URL https://doi.org/10.1007/3-540-46805-6_19

Lillo, M. P. (2016). Prácticas de las organizaciones de la sociedad civil en contexto de desastres en chile: el caso de la fundación para la superación de la pobreza en atacama. *Revista Sul Americana de Psicologia*.

Majumder, N., Poria, S., & Alexander Gelbukh, E. C. (2017). Deep learning-based document modeling for personality detection from text. *IEEE Intelligent Systems (Volume: 32, Issue: 2, Mar.-Apr. 2017)*.

Mehta, Y., Majumder, N., Gelbukh, A., & Cambria, E. (2019). Recent trends in deep learning based personality detection. *Artificial Intelligence Review*, *53*(4), 2313–2339.
URL https://doi.org/10.1007/s10462-019-09770-z

Mercurio, E. (2014). *15 mil voluntarios trabajan en los cerros*. 16 de abril de 2014. p. C6.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013a). Efficient estimation of word representations in vector space.

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., & Joulin, A. (2018). Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Mikolov, T., Yih, W.-t., & Zweig, G. (2013b). Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, (pp. 746–751). Atlanta, Georgia: Association for Computational Linguistics.
URL https://www.aclweb.org/anthology/N13-1090

Millette, G., Valérie (2008). Designing volunteers' tasks to maximize motivation, satisfaction and performance: The impact of job characteristics on volunteer engagement. *Motivation and Emotion*, *32*(1), 11–22.

Munro, K. (2015). Global network of civil society organizations for risk disaster reduction (gndr).

Norman, W. T. (1963). Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personalityratings. *The Journal of Abnormal and Social Psychology, 66(6), 574–583*.

OMS (n.d.). Desarrollo en la adolescencia. Último acceso: 03-01-2020.
URL https://www.who.int/maternal_child_adolescent/topics/adolescence/dev/es

Oracle, E. (2018). Diferencias entre la inteligencia artificial y el machine learning. Último acceso: 03-07-2020.
URL https://medium.com/@experiencIA18/\diferencias-entre-la-inteligencia-artificial-y-el-mach

Pennebaker, J. W., & King, L. A. (1999). Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology*, *77*(6), 1296–1312.
URL https://doi.org/10.1037/0022-3514.77.6.1296

Poria, S., Gelbukh, A., Agarwal, B., Cambria, E., & Howard, N. (2013). Erratum: Common sense knowledge based personality recognition from text. In *Lecture Notes in Computer Science*, (pp. E1–E1). Springer Berlin Heidelberg.
URL https://doi.org/10.1007/978-3-642-45111-9_46

Power, R. A., & Pluess, M. (2015). Heritability estimates of the big five personality traits based on common genetic variants. *Translational Psychiatry*, *5*(7), e604–e604.
URL https://doi.org/10.1038/tp.2015.96

PsycNetDirect, A. (2020). Big five inventory overview.
URL https://psycnet.apa.org/doiLanding?doi=10.1037/t07550-000

Rammstedt, B., & John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the big five inventory in english and german. *Journal of Research in Personality*, *41*(1), 203–212.
URL https://doi.org/10.1016/j.jrp.2006.02.001

Rangel, F., Celli, F., Rosso, P., Potthast, M., Stein, B., & Daelemans, W. (2015). Pan15 author profiling.
URL https://doi.org/10.5281/zenodo.3745945

Ranjith R, C. A., Jothi S (2019). Personality trait analysis by graphology technique using machine learning. *International Journal of Innovative Technology and Exploring Engineering*, *9*(1), 4734–4737.
URL https://doi.org/10.35940/ijitee.a3973.119119

Salminen, J., Rao, R. G., gyo Jung, S., Chowdhury, S. A., & Jansen, B. J. (2020). Enriching social media personas with personality traits: A deep learning approach using the big five classes. In *Artificial Intelligence in HCI*, (pp. 101–120). Springer International Publishing.
URL https://doi.org/10.1007/978-3-030-50334-5_7

Stamatatos, E., Potthast, M., Rangel, F., Rosso, P., & Stein, B. (2015). Overview of the PAN/CLEF 2015 evaluation lab. In *Lecture Notes in Computer Science*, (pp. 518–538). Springer International Publishing.
URL https://doi.org/10.1007/978-3-319-24027-5_49

Tandera, T., Hendro, Suhartono, D., Wongso, R., & Prasetio, Y. L. (2017). Personality prediction system from facebook users. *Procedia Computer Science*, *116*, 604–611.
URL https://doi.org/10.1016/j.procs.2017.10.016

Tausczik, Y. R., & Pennebaker, J. W. (2009). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, *29*(1), 24–54.
URL https://doi.org/10.1177/0261927x09351676

Thurstone, L. L. (1951). The dimensions of temperament. *Psychometrika*, *16*(1), 11–20.
URL https://doi.org/10.1007/bf02313423

Thurstone, L. L. (1953). Thurstone temperament schedule. *Chicago: Science Research Associates*.

UACh (2010). *Un Techo para Chile Busca Sumar 1000 Voluntarios para Este Fin de Semana*. Recuperado el Martes 29 de Julio de 2016, de https://diario.uach.cl/un-techo-para-chile-busca-sumar-1000-voluntarios-para-este-fin-de-semana/.

Yuan, C., Wu, J., Li, H., & Wang, L. (2018). Personality recognition based on user generated content. In *2018 15th International Conference on Service Systems and Service Management (ICSSSM)*. IEEE.
URL https://doi.org/10.1109%2Ficsssm.2018.8465006

ANEXO A. ANEXO

Para mejorar la visualización de elementos presentes en el trabajo son incluidos en este anexo.

A.1 CARTA GANTT COMPLETA

En la figura A.1 se expresa de forma completa y horizontal para una mejor lectura.

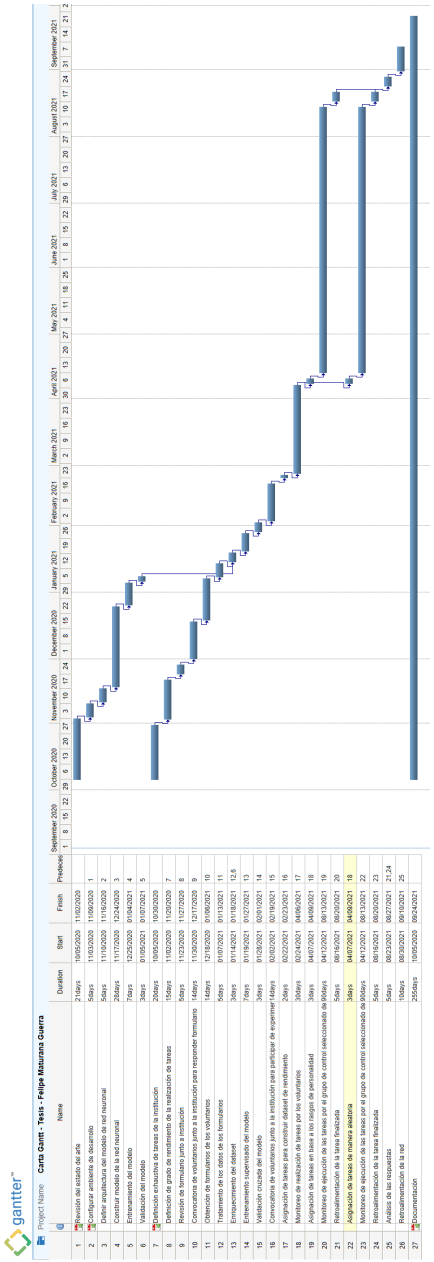


Figura A.1: Carta Gantt completa. Fuente: Elaboración propia.

ANEXO B. COMPLEMENTO MARCO TEÓRICO

En esta sección se abordan de manera conceptual los principales conceptos y teorías que sustentan parte de la formulación de las preguntas de investigación así como el desarrollo propio de este trabajo y favorece a un mejor entendimiento del tema de investigación.

B.1 APRENDIZAJE PROFUNDO

Deep Learning o Aprendizaje Profundo es un subconjunto de la disciplina de inteligencia artificial y específicamente un subconjunto de la disciplina de *Machine Learning* ver Figura B.2. El aprendizaje profundo permite desarrollar modelos computacionales mediante una representación de múltiples capas de procesamiento que poseen funciones de activación de acuerdo a umbrales definidos, imitando así el comportamiento de sinapsis de las neuronas de nuestro cerebro, pero en múltiples capas y niveles de abstracción. En la Figura B.1 se aprecia de mejor forma el paso directo en una red neuronal con dos capas ocultas y una capa de salida, cada una de las cuales constituye un módulo a través de cuál puede propagar hacia atrás (retro-propagar) los gradientes para así ajustar los valores correspondientes.

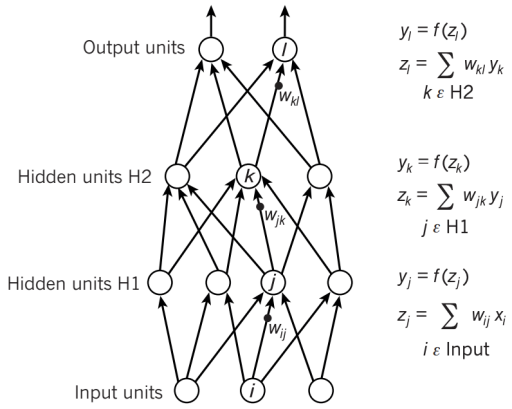


Figura B.1: Red neuronal multi-capa y retro-propagación. Imagen obtenida de la publicación ´Deep Learning´ en *Nature* (LeCun et al., 2015).

Este tipo de técnicas permite descubrir y/o predecir comportamientos y estructuras complejas en conjuntos con alta cantidad de datos mediante un el algoritmo de retro-propagación del error que indica, regula y modifica los parámetros internos de cada uno de los nodos que corresponda en la red neuronal donde se aplique esta técnica para así poder calcular de acuerdo a la representación de la capa o nodo anterior. (LeCun et al., 2015).

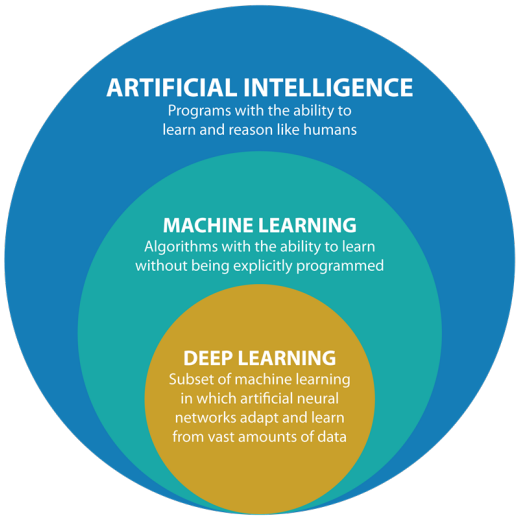


Figura B.2: Disciplinas dentro de la Inteligencia Artificial: *Machine Learning* y *Deep Learning*. Imagen obtenida de ´Diferencias entre la inteligencia artificial y el *machine learning*´(Oracle, 2018)

.

B.2 RED NEURONAL CONVOLUCIONAL

Dentro de la rama de aprendizaje profundo se encuentran las redes neuronales convolucionales o *CNN* por sus siglas en inglés, es comúnmente aplicada al análisis de imágenes. Tienen grandes aplicaciones en reconocimiento de imagen y video, sistemas de recomendación clasificación de imágenes, análisis de imágenes médicas, procesamiento de lenguaje natural (Collobert & Weston, 2008).

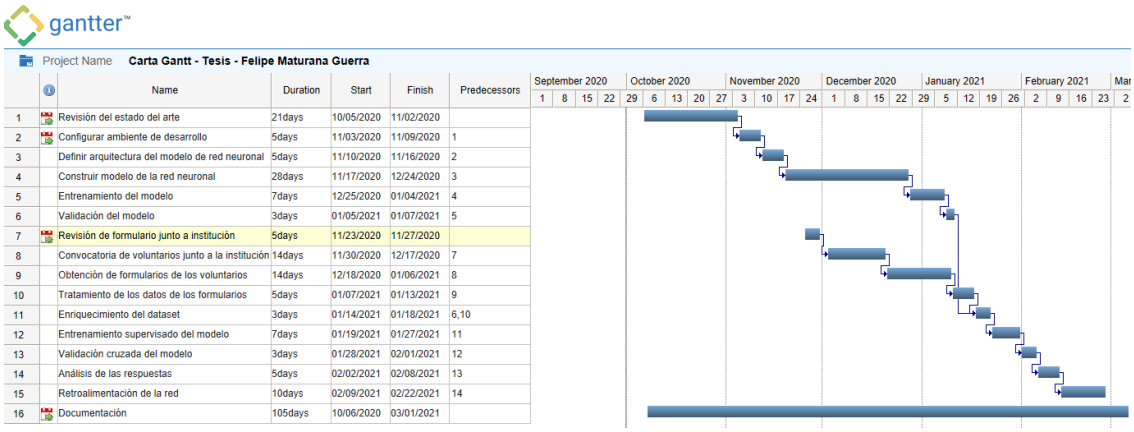
Las redes neuronales convolucionales muestran grandes propiedades al momento de reconocer formas con alta variabilidad como escritos a mano, imágenes e incluso textos. Resulta ventajoso al momento de extraer un conjunto correcto de características. Las CNN Se ha demostrado que son adecuadas para el tipo de tareas descritas (LeCun et al., 1999).

Como su nombre lo indica, la red neuronal aplica una función matemática llamada convolución, es un tipo especial de operación lineal, es decir, son simplemente redes neuronales que utilizan la convolución en lugar de la multiplicación matricial general en al menos una de sus capas. (Goodfellow et al., 2016)

Respecto a la arquitectura una *CNN* consiste en una capa de entrada y una de salida junto a capas ocultas, estas consisten en una serie de capas convolucionales que se relaciona con funciones matemáticas, la función de activación de una capa es generalmente tipo *RELU*, seguida de una función de agrupación seguida de una capa de normalización.

B.2.1 Carta Gantt

Para el desarrollo del proyecto se contemplan 2 semestres, considerando el inicio desde el 5 de octubre de 2020 y el término para el día 01 de marzo de 2021. El detalle de las actividades con su duración y la carta gantt se encuentran en la Figura B.3, además, la versión extendida para mejorar su visualización se encuentra en el anexo en la Figura A.1.



	Name	Duration	Start	Finish	Predecessors
1	Revisión del estado del arte	21days	10/05/2020	11/02/2020	
2	Configurar ambiente de desarrollo	5days	11/03/2020	11/09/2020	1
3	Definir arquitectura del modelo de red neuronal	5days	11/10/2020	11/16/2020	2
4	Construir modelo de la red neuronal	28days	11/17/2020	12/24/2020	3
5	Entrenamiento del modelo	7days	12/25/2020	01/04/2021	4
6	Validación del modelo	3days	01/05/2021	01/07/2021	5
7	Revisión de formulario junto a institución	5days	11/23/2020	11/27/2020	
8	Convocatoria de voluntarios junto a la institución	14days	11/30/2020	12/17/2020	7
9	Obtención de formularios de los voluntarios	14days	12/18/2020	01/06/2021	8
10	Tratamiento de los datos de los formularios	5days	01/07/2021	01/13/2021	9
11	Enriquecimiento del dataset	3days	01/14/2021	01/18/2021	6,10
12	Entrenamiento supervisado del modelo	7days	01/19/2021	01/27/2021	11
13	Validación cruzada del modelo	3days	01/28/2021	02/01/2021	12
14	Análisis de las respuestas	5days	02/02/2021	02/08/2021	13
15	Retroalimentación de la red	10days	02/09/2021	02/22/2021	14
16	Documentación	105days	10/06/2020	03/01/2021	

Figura B.3: Carta Gantt de plan de trabajo (el formato de fechas en la imagen es M/D/A). Fuente: Elaboración propia.

B.2.2 Muestra de respuestas de voluntarios a instrumento aplicado

Se detallan algunas respuestas de los voluntarios frente a las 7 preguntas realizadas de texto libre a través del instrumento utilizado.

Pregunta 1: Describe desde tu perspectiva la importancia de la instancia de los voluntariados, además, de la participación de los voluntarios.

1. Una gran importancia es que son espacios donde la ciudadanía puede participar, crear y buscar apoyar a grupos que se encuentran en situaciones de vulnerabilidad, por lo que abre espacios de acción e involucramiento. Estos, a su vez, son profundamente formadores, desde una mayor capacidad de empatizar con otros, como de conocer realidades que pueden ser ajenas. Finalmente, y en relación a lo mencionado anteriormente, es un acto profundamente político de fácil acceso (a diferencia de ser candidato a alguna elección), donde se declara y actúa de acuerdo a principios, superando el individualismo y actuando de manera colectiva.

2. La importancia radica en que el ser humano es un ser que vive en sociedad y por lo tanto necesitamos trabajar los unos con los otros aunque no nos conozcamos. El Voluntariado es justamente una forma de apoyo abstracto y que nos hace salir de nuestros grupos de confort tales como la familia, amigos, compañeros de trabajo etc. Aporto a la fundación Arturo López Pérez, en mi vida de estudiante participe activamente en voluntariado de mi universidad.

3. El voluntariado es clave, tanto es las tareas especificas a desarrollar en contextos vulnerables, emergencias, etc... dando respuesta a necesidades concretas del territorio; pero así también en la formación personal de esxs voluntarixs, ese complemento de su formación académica no se entrega en niguna aula, mas que en el territorio mismo.

4. Creo que participar en un voluntariado como TECHO, te da la posibilidad de vivir situaciones poco comunes e incluso incómodas, dado que las comunidades se encuentran en una situación precaria y de escasez. Como voluntarios, debemos tomarnos en serio las tareas porque tienen un impacto inmediato en las familias (por ejemplo, un contenedor de agua facilita el día a día de las familias).

Pregunta 2: ¿Cuál es o fue tu principal motivación para participar como voluntario?

1. Fue conocer otras realidades para poder aprender de ellas. Quería tener una mayor experiencia a nivel emocional , por lo que compartir mi tiempo con otros provocaría un crecimiento en mi y una visión mas global de diferente hechos.

2. Pertenezco a bomberos de chile desde que tenía 11 años, llegué motivado por un compañero de curso que participaba como brigadier en un cuartel, y desde allí que sigo participando. Creo que lo primero que me motivó fue lo lúdico y entretenido que era el mundo de los bomberos, pero lo que me mantiene hoy es el servicio a la comunidad y lo bello de ayudar a otros cuando más lo necesitan.

3. Es dar una vuelta de mano, soy una de las primeras profesionales de mi familia y me siento muy afortunada por lo mismo para mi es clave trabajar por una sociedad mejor la cual ha ido transitando desde que fui estudiante y ahora como profesional.

4. Aportar mis conocimientos, perspectivas, experiencias y creatividad para contribuir en la solución de algún problema que afecta a las comunidades humanas o al medioambiente.

Pregunta 3: ¿Qué cualidades te gustaría que tuviesen las personas con las que trabajas en los voluntariados? Indique al menos tres y cuéntanos el porqué.

1. Tolerancia a la frustración, ya que te enfrentas a realidades diversas, desconocidas por tanto es necesario tener la capacidad de adaptación al cambio y a lo que no resulta; innovación ya que a veces con pocos recursos económicos hay que buscar otro tipo de recursos para lograr los objetivos y disciplina, ya que es necesaria la constancia para poder generar cambios.

2. Que tengan mente abierta, generalmente en los trabajos uno se enfrenta a realidades muy distintas a las que uno encuentra en el día a día lo cual puede ser muy fuerte. Luego está el tema de ser empatico, lo cual se desarrolla en los mismo voluntariados pero encuentro que es esencial dado a que da facilidad en el trabajo mismo y a la hora de conocer a los involucrados. Por último que sea motivado, que existan ganas de ayudar dado a que derrepente uno se encuentra con problemas en la obra y hay que encontrar soluciones alternativas.

3. Amabilidad, Compromiso, Disposición y Humildad. Amabilidad porque siempre hay estrés, y la amabilidad y la respuesta amable aplaca el enojo. Compromiso, porque así todos remamos para el mismo lado, con la misma fuerza podemos llegar más lejos. Disposición: Es fácil inscribirse en voluntariados, pero se necesita más que eso para permanecer con una buena actitud, esa es la disposición. Humildad: un voluntario con el ego alto, sólo va a dañar a su entorno. Debe ser fácil de corregir y con corazón humilde.

4. Primero, siempre debe estar la vocacion, esto es el ADN del voluntariado. Segundo, la buena voluntad, no todos sabemos desarrollarnos o tener conocimientos técnicos en algunos temas, pero para apoyar o ayudar lo destacado es la buena voluntad. Tercero es ser empático, tener esa visión de que hay personas que están sufriendo y uno no está ajeno a esa situación, hoy les piede tocar a ellos y otras veces me puede tocar a mí.

Pregunta 4: ¿Cuál fue su rol en la última tarea que desempeñó en un voluntariado? Describa lo que más pueda

1. Acompañamiento a un grupo de jóvenes en busca de su identidad vocacional. Generaba sesiones que les permitiese abordar diferentes temáticas, un espacio de conversación. Preparación del material para cada una de las sesiones.

2. La última vez fui jefe de cuadrilla en una obra de construcción de techo en el área de lo espejo. Nos tocó hacer una vereda para un pasaje que se inundaba en el invierno. Esto requiere mucha fuerza física por lo que había que echar ojo de que los chicos no se fatiguen ni lesionen y de mantenerlos animados. También me tocó reemplazar al camioneta que se enfermó, pero esto fue solamente ir a comprar materiales y descargarlas en la obra.

3. Estuve como jefe de cuadrilla a cargo con la construcción de una capilla en el sur de Chile luego del terremoto y maremoto del 2010. Mi rol fundamental era poder llevar a cabo lo que teníamos como tarea principal, asegurarme que a nivel de estructura quedara todo firme, que el equipo trabajara en conjunto y motivarlos para que a pesar de las condiciones adversas se cumplieran las metas y objetivos diarios.

4. Coordinar construcciones de viviendas de emergencia en el Incendio de la región del Maule 2017 o 2018 (no recuerdo muy bien el año). Tarea de liderar el equipo que castatraba los casos; coordinar la logística de recepción, distribución y construcción de viviendas; coordinar y apoyar el trabajo voluntario de construcción de las diferentes redes que quisieron apoyar, universidades, centros de formación tecnica, etc.

Pregunta 5: De acuerdo a las tareas que has desarrollado como voluntario ¿Cómo evalúas tus capacidades y cuáles son?

1. Las evaluó de manera positiva, soy muy responsable y comprometida. Pero si debo ser realmente honesta, creo que en esos tiempo, me falto madurez para enfrentar el reto de compartir con otros voluntarios de mejor manera, a pesar de que nunca tuve un inconveniente ni enfrentamiento importante, creo que las habilidades sociales con las que contaba en ese tiempo eran mucho menor a las que poseo en estos momentos.

2. En primer lugar he desarrollado una gran proactividad de buscar hacerse cargo de los problemas e intentar sumar más gente a distintas iniciativas. En segundo lugar, la capacidad de organización, dado que tener responsabilidad importantes (aunque sea en un voluntariado) junto con estudiar, requiere poder organizarse muy bien. En tercer lugar, la capacidad de liderar, mediante el ejemplo y motivando a los demás, en proyectos que son voluntarios pero altamente emocionales, donde uno busca que todos se involucren lo más posible para lograr un gran resultado.

3. Poder hablar con gente, tanto como para pasar el rato, calmar, organizar, pedir cosas,etc. poder improvisar cuando hay problemas. poder explicar que se esta haciendo y por que. Mantener la calma. organizar comida. etc. creo que lo que mas me ayudo fue a crecer mis habilidades blandas y sociales ya que era una area donde tenia problemas y que queria trabajar.

4. Desarrollé capacidades que creía no tener y mejoré o potencié otras. La capacidad de organizar, coordinar y liderar un equipo de logística de emergencia en una región diferente a la cual me desempeñaba.... capacidad de ubicación, sin conocer las localidades al segundo o tercer día ya debía reconocer rutas, dar referencias a los camiones de despachos, etc.

Pregunta 6: ¿Cómo describirías tu experiencia y actuar frente a situaciones que se presentan imprevistos o emergencias?

1. Muy buena, en general uno aprende a moverse con pocos recurso y con urgencia, de que uno no puede esperar meses. Además, al estar siempre en situaciones no tan cómodas en cantidad de recursos y otros aspectos, suelen surgir imprevistos o emergencias que uno se mueve de la manera necesaria para resolverlos.

2. Soy resolutiva. Creo respuestas rápidas. Me gusta todo lo que es acampar y trekking y siempre ocurren imprevistos o emergencias, y hay que saber arreglárselas, me encanta crear respuestas raras, que nos saquen del paso.

3. Creo que trabajar en emergencias siempre es un desafío, siempre hay algo nuevo que aprender o algo a lo que prestar atención. En mi experiencia, creo que siempre hay que tener cuidado de los imprevistos, y ser cauto a la hora de actuar.

4. Desde que inicie hasta el día que me retire del voluntariado,afrontaba de la mejor manera los imprevistos que se presentaban siempre de manera racional y de rápida ejecución. Falle varias veces al inicio y más de una vez me lleve un reto de la gente de oficina pero con los años lo mejoré.

Pregunta 7: ¿Consideras a futuro realizar una nueva actividad de voluntariado? ¿Qué factores influyen en tu respuesta?

1. Principalmente influye en lo que este mi vida en el momento en que esto pudiera ocurrir, hijos, familia, trabajo etc. Y por supuesto que lo consideraría, dedicar parte de tu tiempo a otros es un gesto hermoso de amor al prójimo y por uno mismo.
2. Si, suelo hacerlo de manera política, participando como voluntario en campañas. Probablemente no hago en temas físicos de construcción, por la demanda de tiempo (varios días seguidos) que suelen requerir. En mi respuesta influye todo lo que he aprendido y las personas que he conocido, lo que me ha hecho crecer de gran manera.
3. Sí lo considero, siempre hay problemas que resolver en las comunidades. Actualmente, el principal factor es la situación sanitaria que no permite o restringe la realización de actividades presenciales.
4. Si, es algo que tengo adquirido en mi ADN por lo que en lo que se necesite y pueda aportar un granito, ahí estoy.

B.2.3 Representación jerárquica de rasgos de personalidad, incluidas facetas y sub-facetas

La evolución constante del estudio del comportamiento humano ha dado origen a facetas y sub-facetas tomando como base los cinco grandes de la personalidad ahondando más en ellas lo que permite un estudio más preciso de acuerdo a la personalidad y cómo impacta en la vida de las personas. La figura B.4 permite vislumbrar el avance de estos estudios a lo largo de los años, en primer lugar transformando estos 5 grandes estudiados en 10 facetas y posteriormente en 30 sub-facetas.

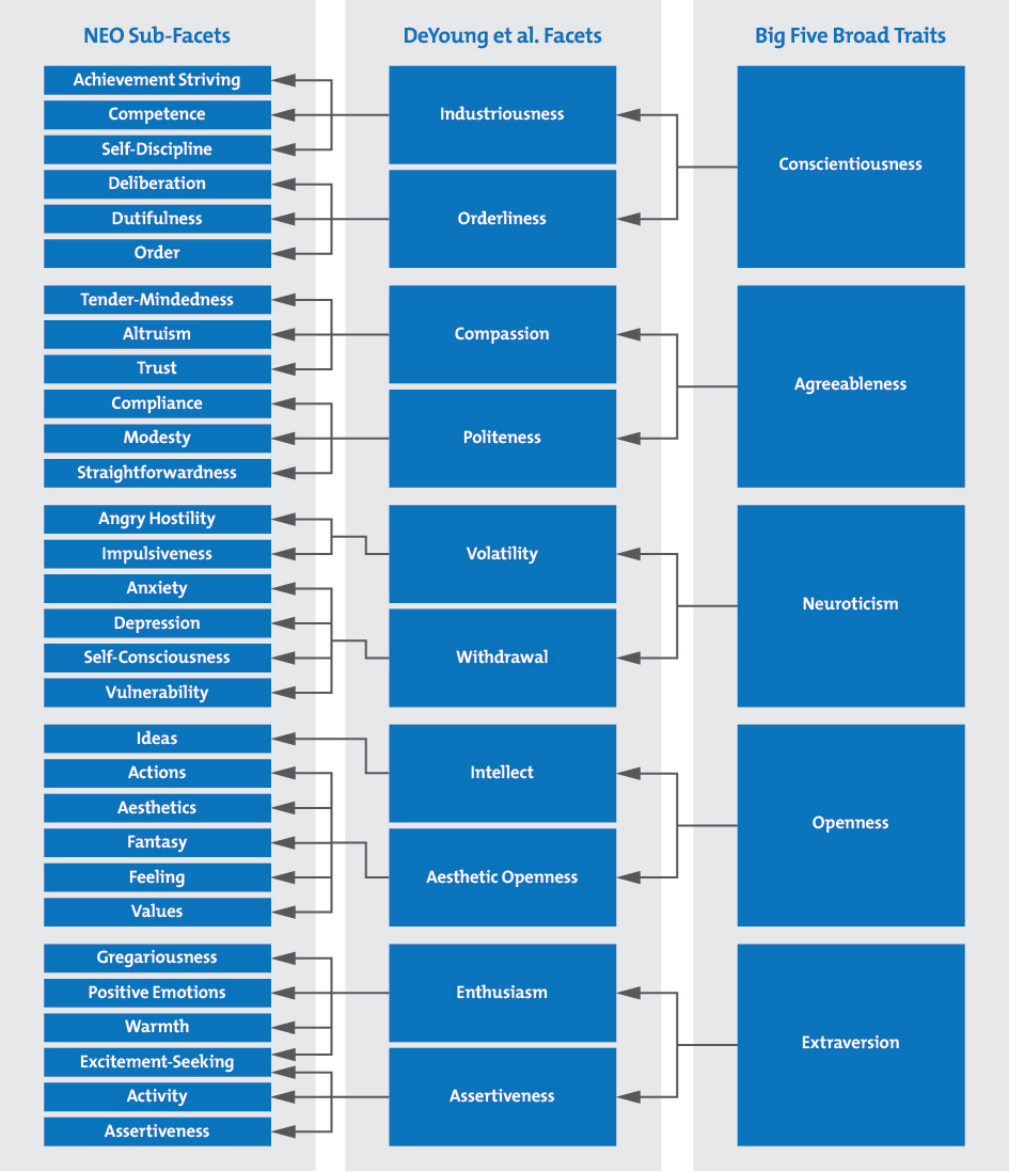


Figura B.4: Representación jerárquica de rasgos de personalidad, incluidas facetas y sub-facetas (Judge et al., 2013).