UNIVERSIDAD DE SANTIAGO DE CHILE

INGENIERÍA ESTADÍSTICA
UNIVERSIDAD DE SANTIAGO DE CHILE

Deep Learning

Codificación CIUO 08 bajo redes neuronales

Andrés Rojas Elgueta
Bayron Espinoza Venegas
Bastián Díaz Vergara

Agosto, 2024

Contenido de la presentación

Recordemos: ¿Qué es el CIUO 08?

UNIVERSIDAD DE SANTIAGO DE CHILE

• Según INE (2018), la Clasificación Internacional Uniforme de Ocupaciones permite ordenar jerárquicamente y agrupar diversas ocupaciones de acuerdo al tipo de tareas realizadas en un puesto de trabajo y las competencias requeridas para ello, dependiendo del nivel y de la especialización de estas competencias.

¿Por qué es importante?

- Proporciona un marco de datos comparables internacional.
- Permite la producción de datos estructurados, útiles para su análisis e investigación.
- Facilita la toma de decisiones específicas y las actividades orientadas a la acción.

Objetivos

Objetivo General: Desarrollar distintos modelos de aprendizaje profundo para la codificación de al menos dos niveles de la Clasificación Internacional Uniforme de Ocupaciones 2008 (CIUO-08).

Objetivos Específicos

- Representar de manera vectorial las respuestas textuales, capturando la relación semántica de las respuestas mediante la técnica de Words Embeddings.
- Implementar una red neuronal recurrente tipo LSTM bidireccional con MLP para codificar cada nivel de la clasificación a partir de la representación vectorial.

Descripción de Datos

Nombre	Descripción	Tipo	Valor
Ocupación	Ocupación del Encuestado	Texto	Hasta 249 Caracteres
Tarea	Tareas que cumple en su Ocupación	Texto	Hasta 492 Caracteres
CIUO_N	Clasificación Internacional Uniforme de Ocupación , desde Nivel 1 a Nivel 4	Factor	Nivel 1: 10 Grupos Nivel 2: 44 Grupos Nivel 3: 162 Grupos Nivel 4: 649 Ocupaciones

Tabla 1: Descripción de Variables

Proporción por categorías nivel 1

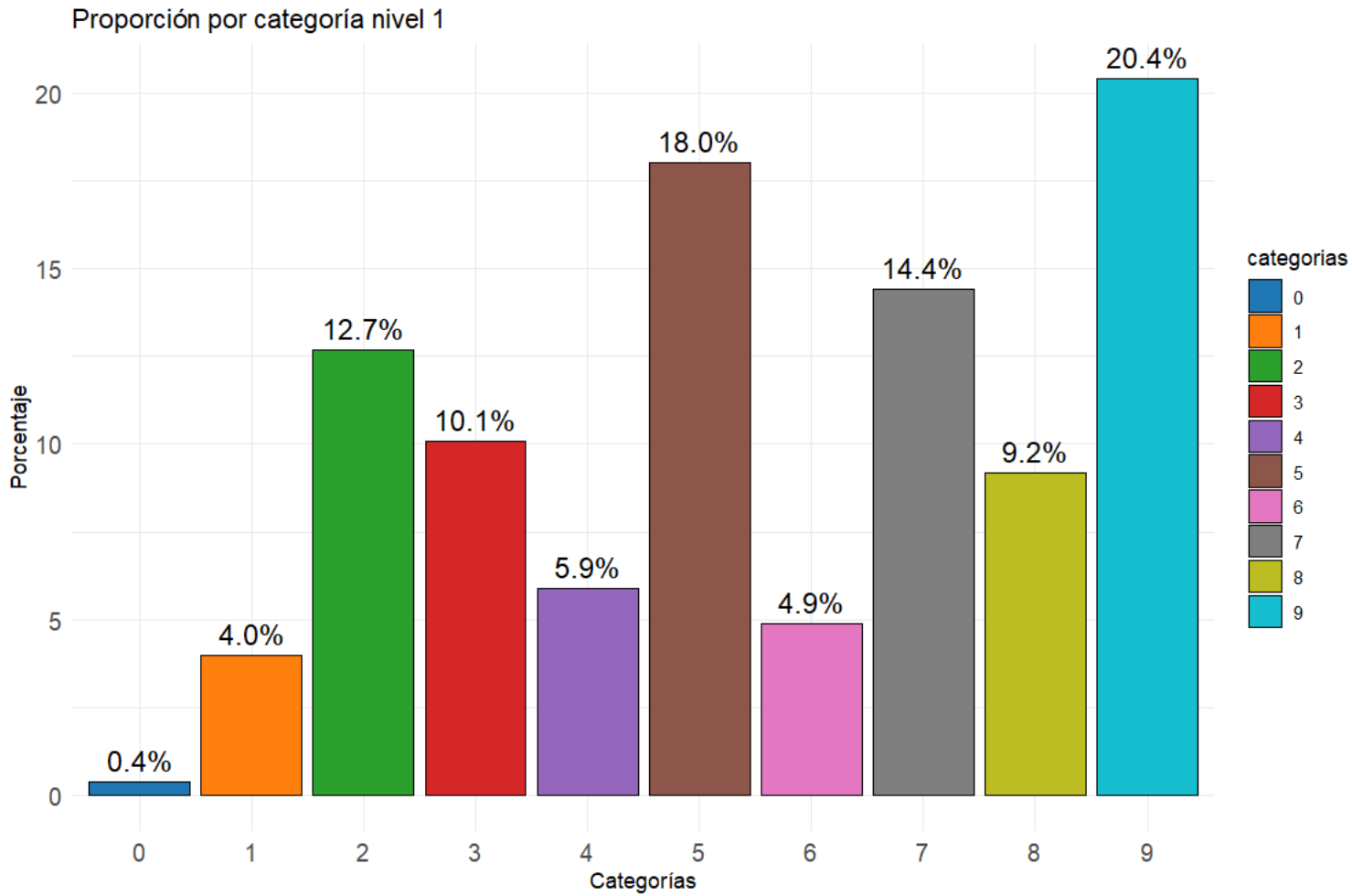


Figura 1: Gráfico de barras porcentual

Frecuencia categorías nivel 2 dado el nivel 1

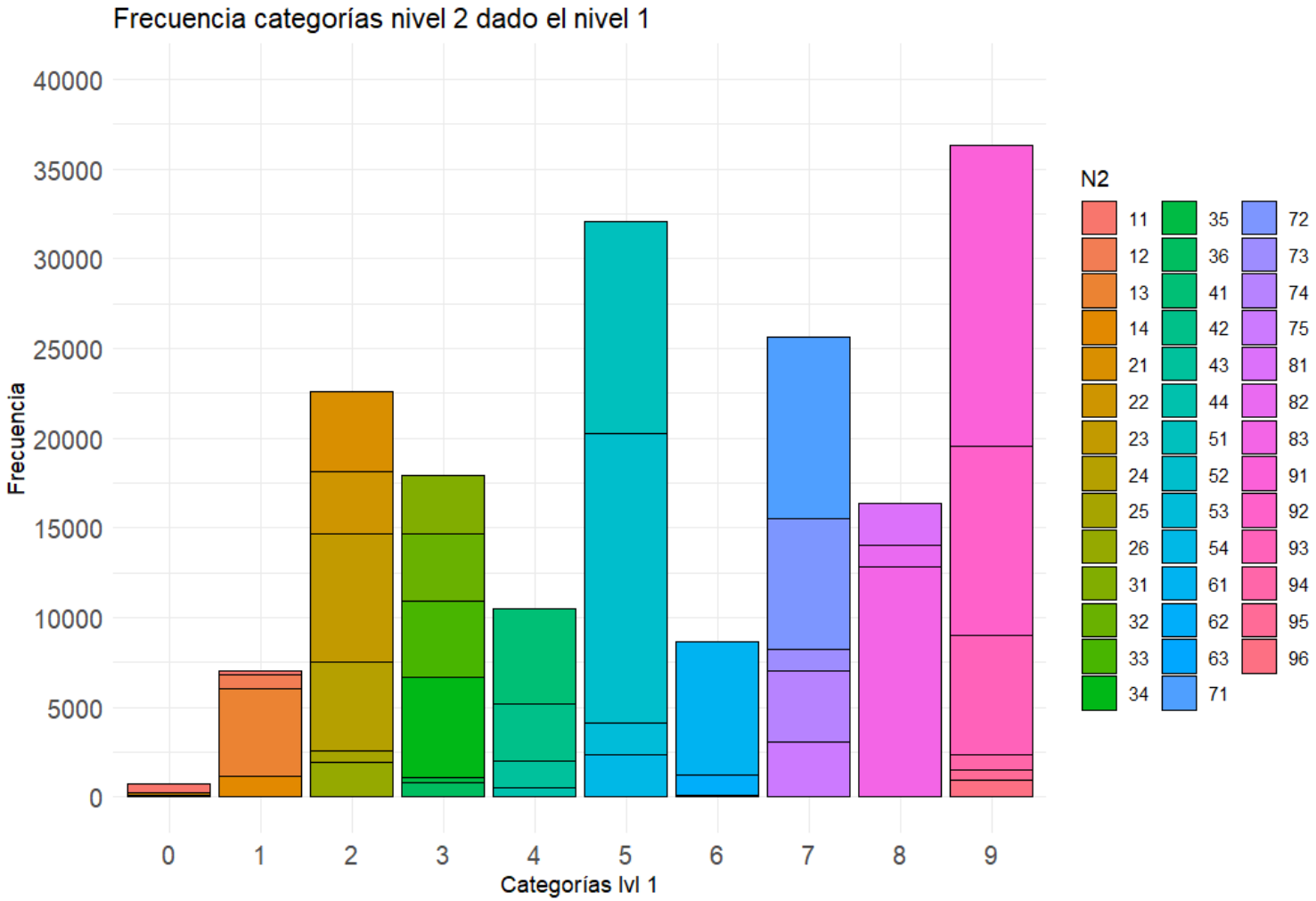
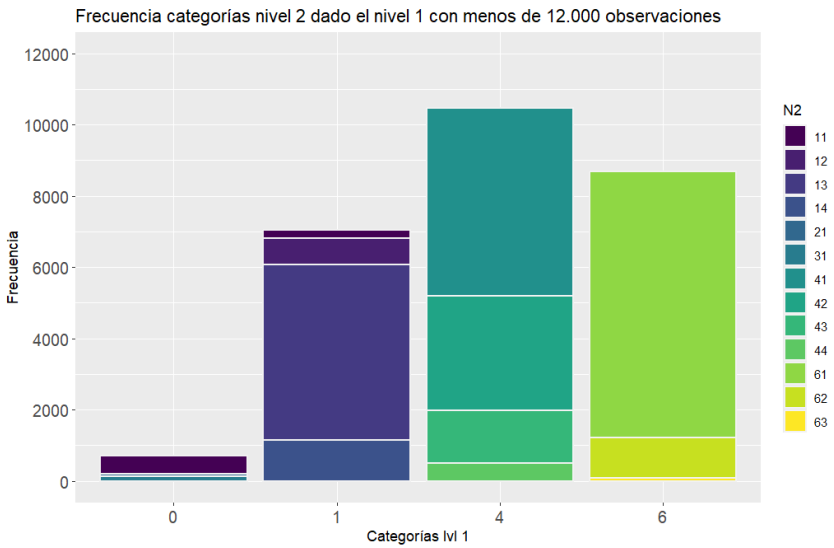
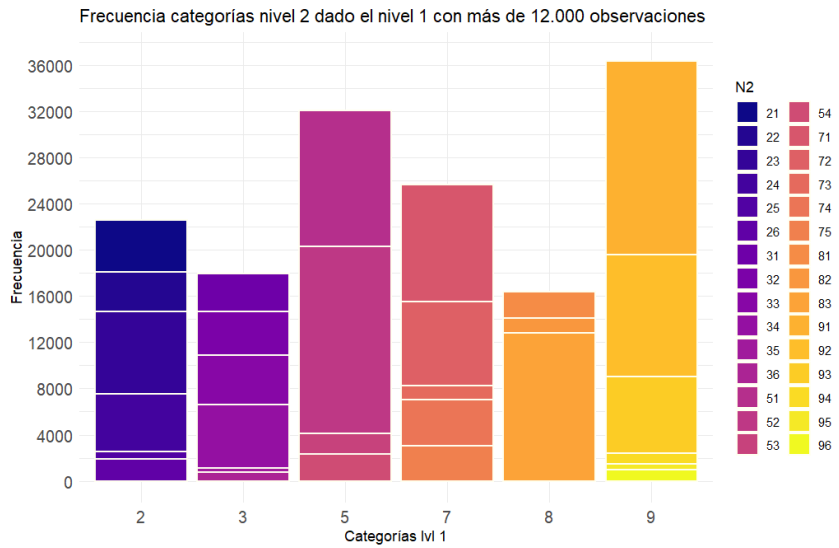


Figura 2: Gráfico de barras de frecuencia

Frecuencia categorías nivel 2 dado el nivel 1



(a) Frecuencias menores a 12.000 observaciones.



(b) Frecuencias mayores a 12.000 observaciones.

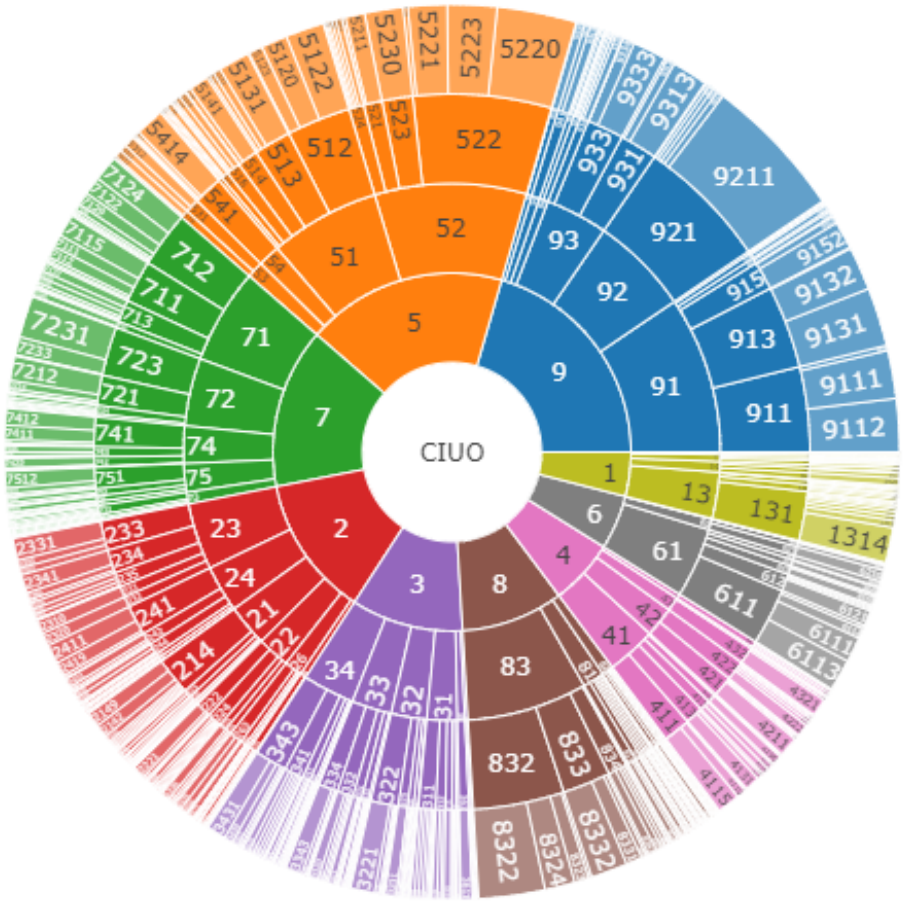
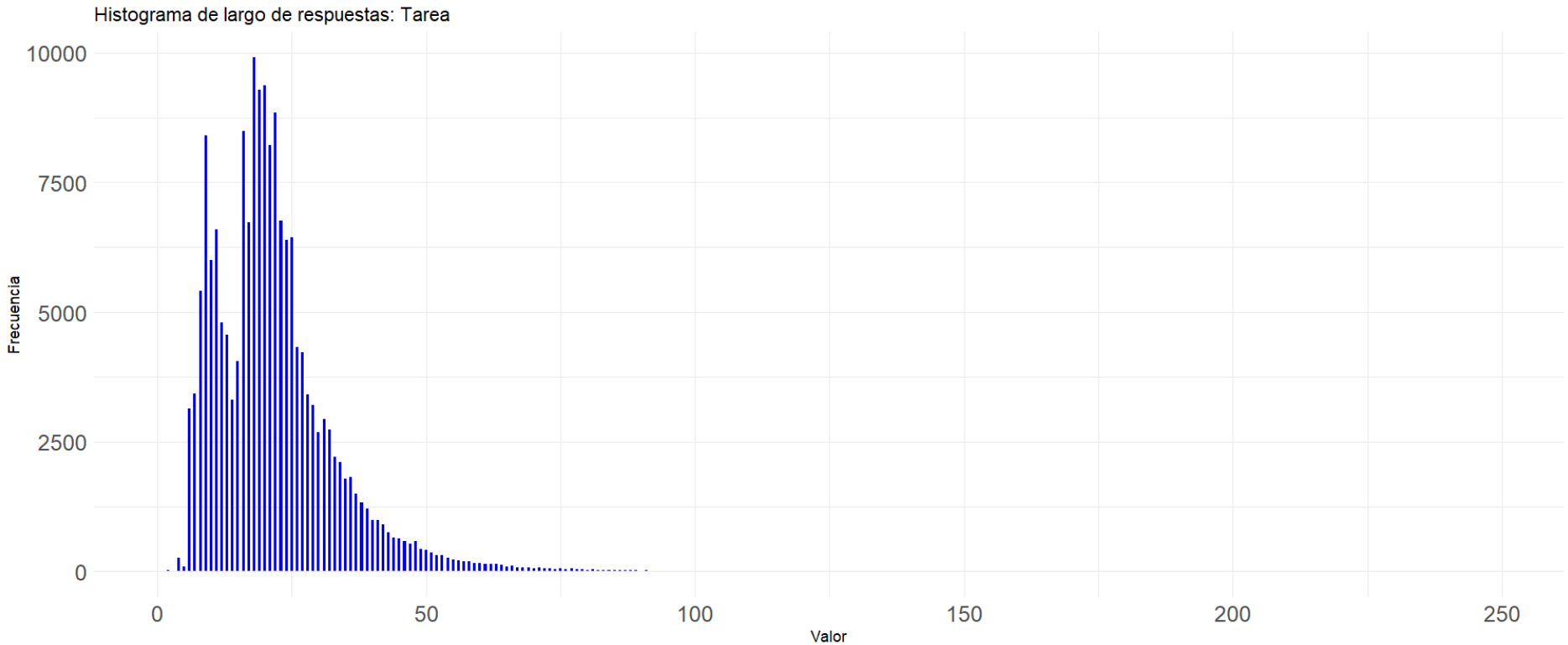
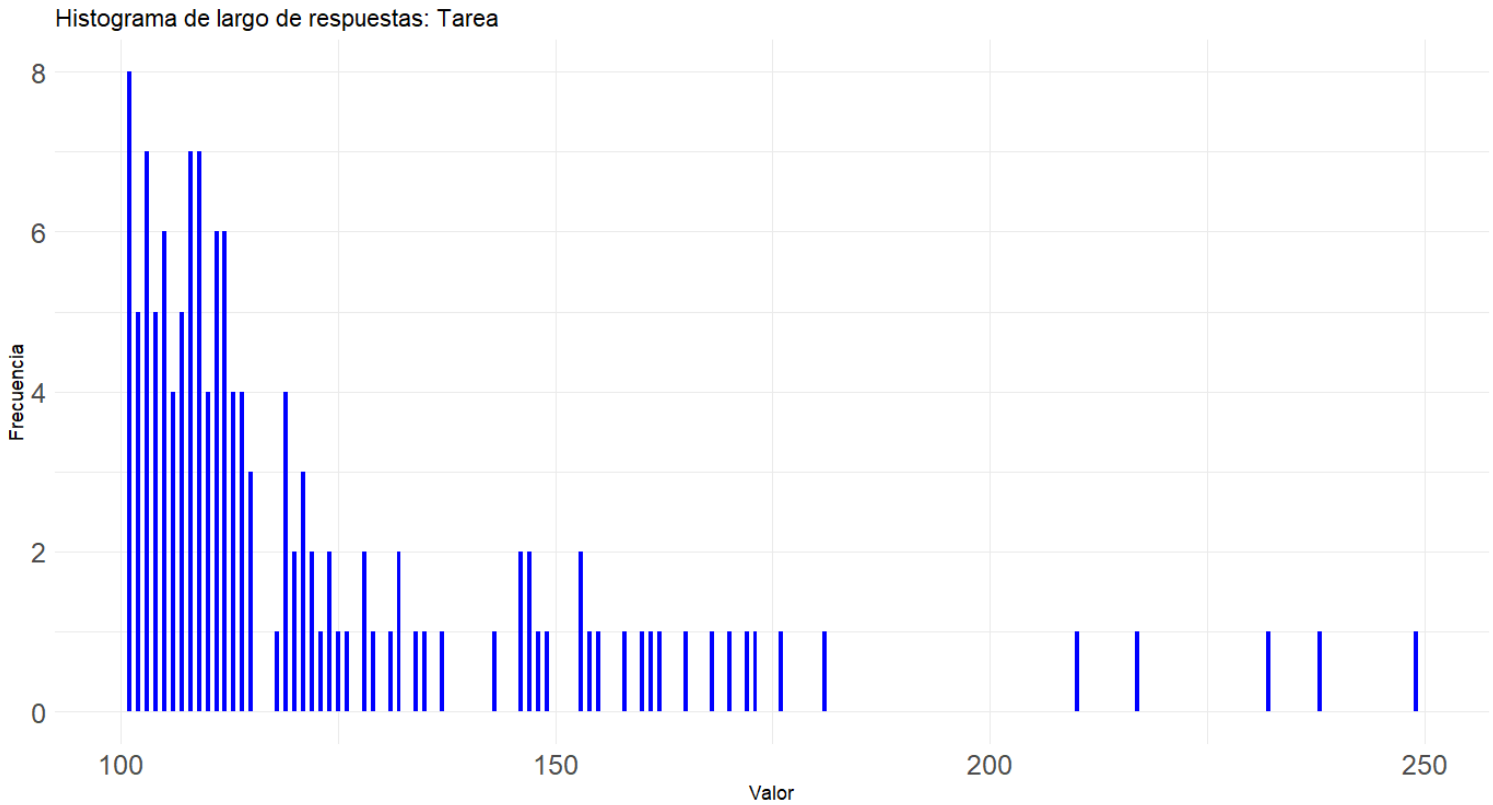


Figura 4: Clasificación jerárquica según gráfica rayos de sol

Largo de respuestas variable Tarea

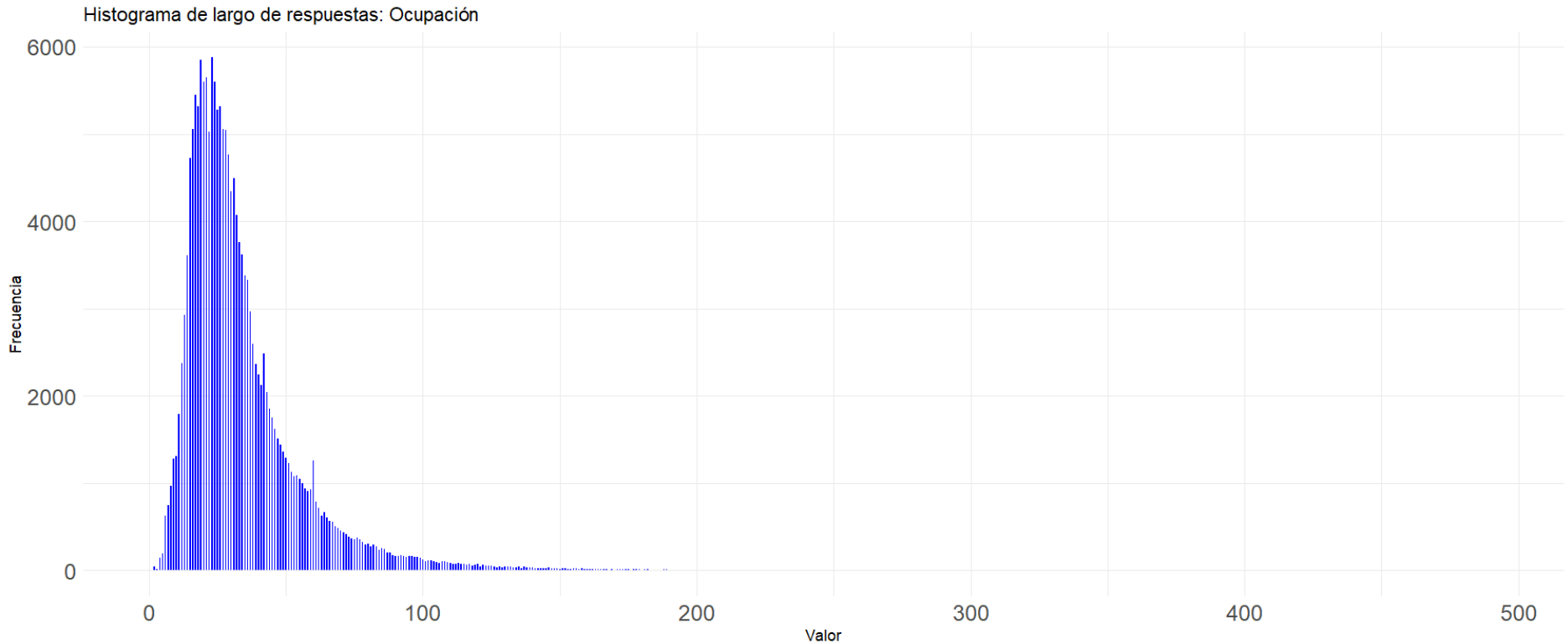


Zoom largo de respuestas variable Tarea

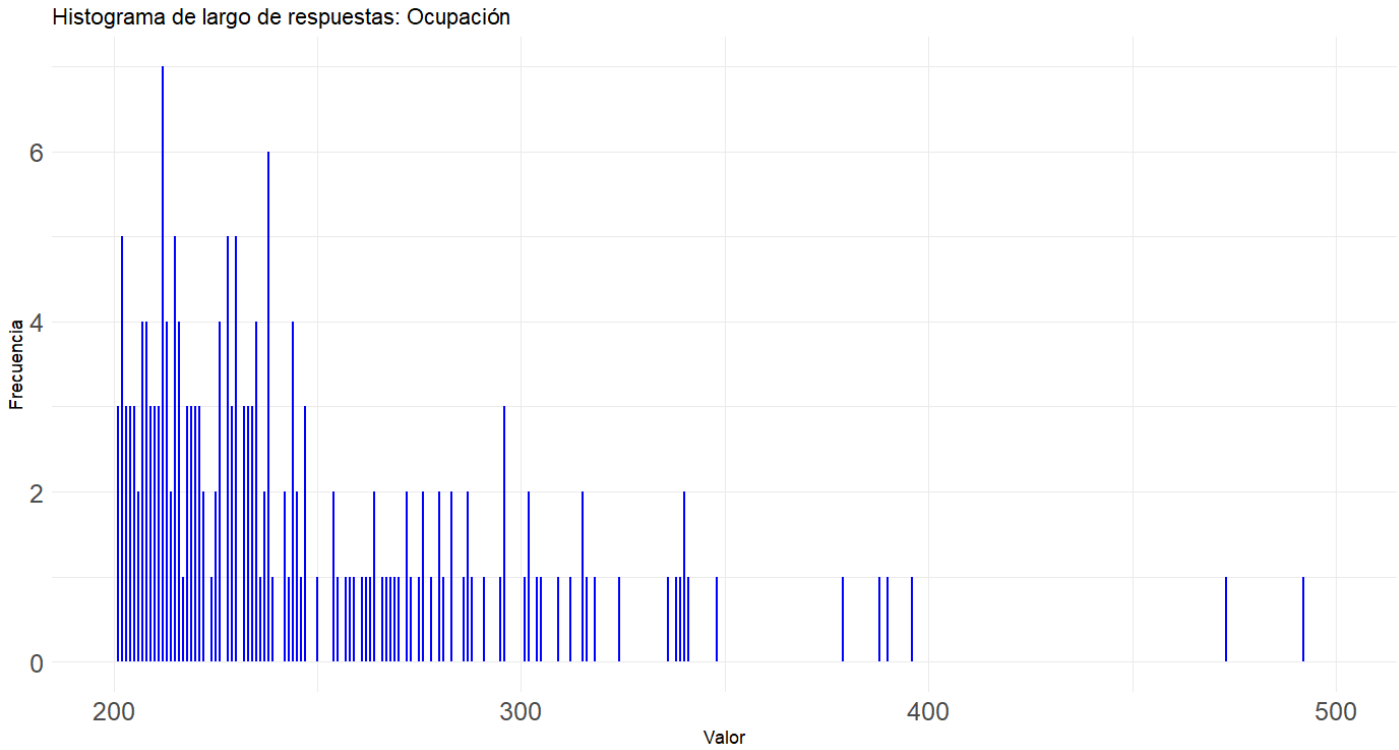


Ejemplo: "Realiza contabilidad compras y ventas de productos atención de clientes y trámites del exportaciones de productos."

Largo de respuestas variable Ocupación

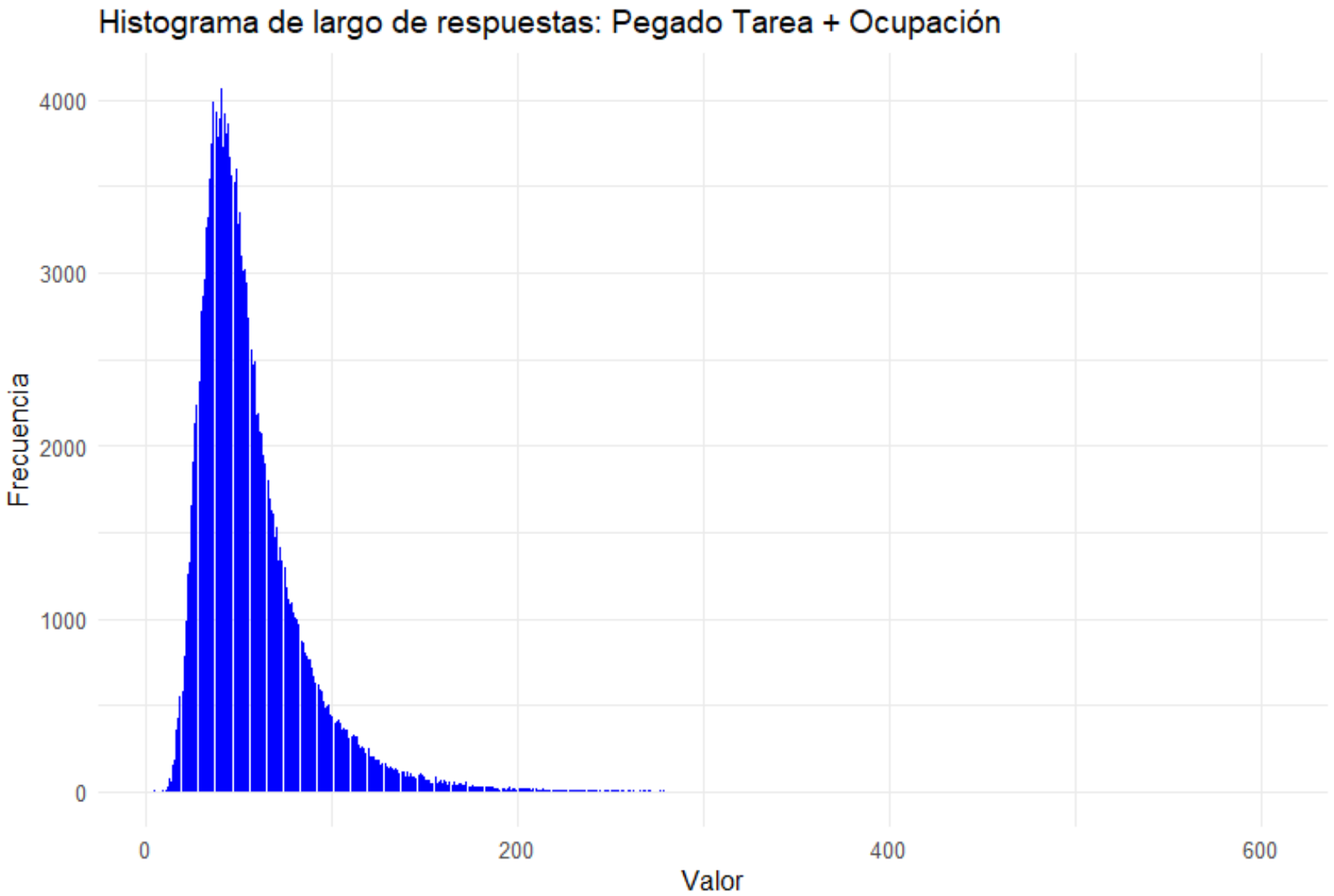


Zoom largo de respuestas variable Ocupación



Ejemplo: “Tiene un emprendimiento familiar de decoración de repostería, elabora productos para decorar pasteles, utiliza fondant, glaseado, merengues, entre otros insumos para personalizar tortas y otros productos de repostería.”

Largo de respuestas del pegado de Tarea + Ocupación



Recuerdo: Word Embedding

¿Qué es Word Embedding?

Corresponde a una técnica de NLP que representa las palabras en forma de vectores en un espacio vectorial, en donde se captura la relación semántica de las palabras, reduciendo la dimensionalidad.

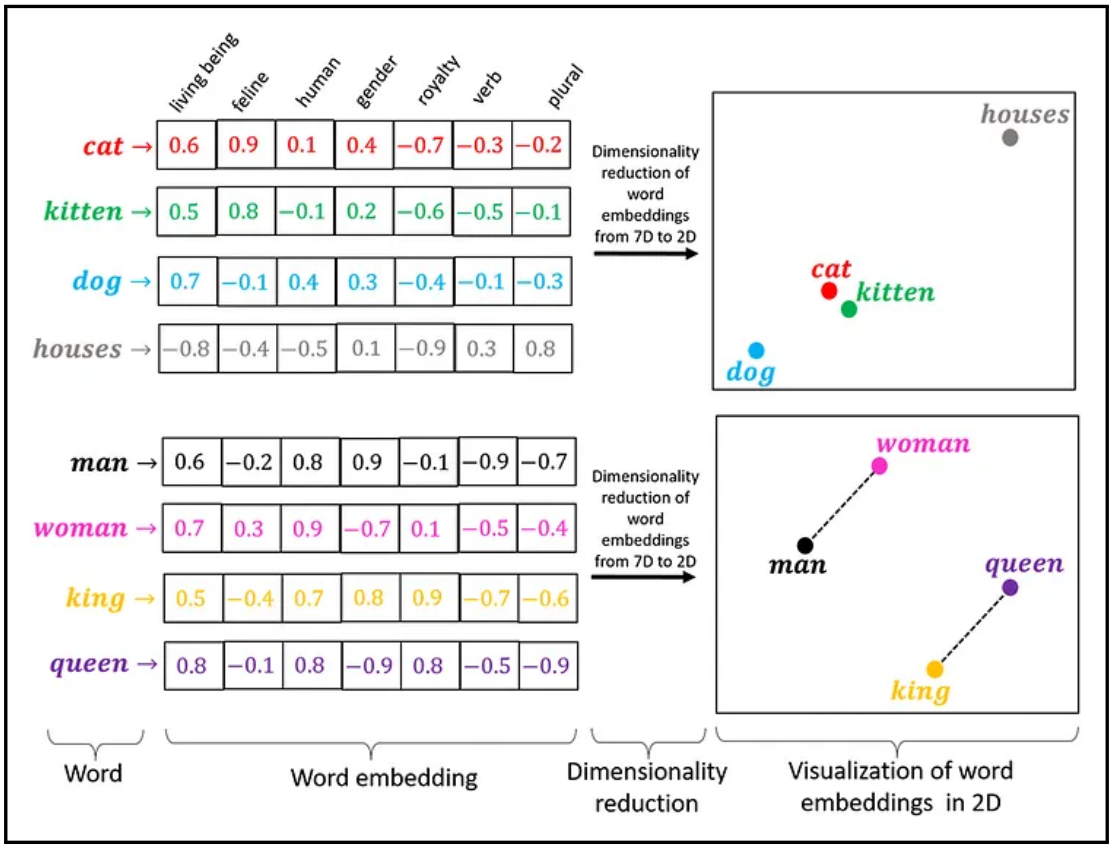


Figura 5: Word Embedding

Esquema Metodológico

Para la realización de este proyecto se seguirá la siguiente metodología de trabajo:

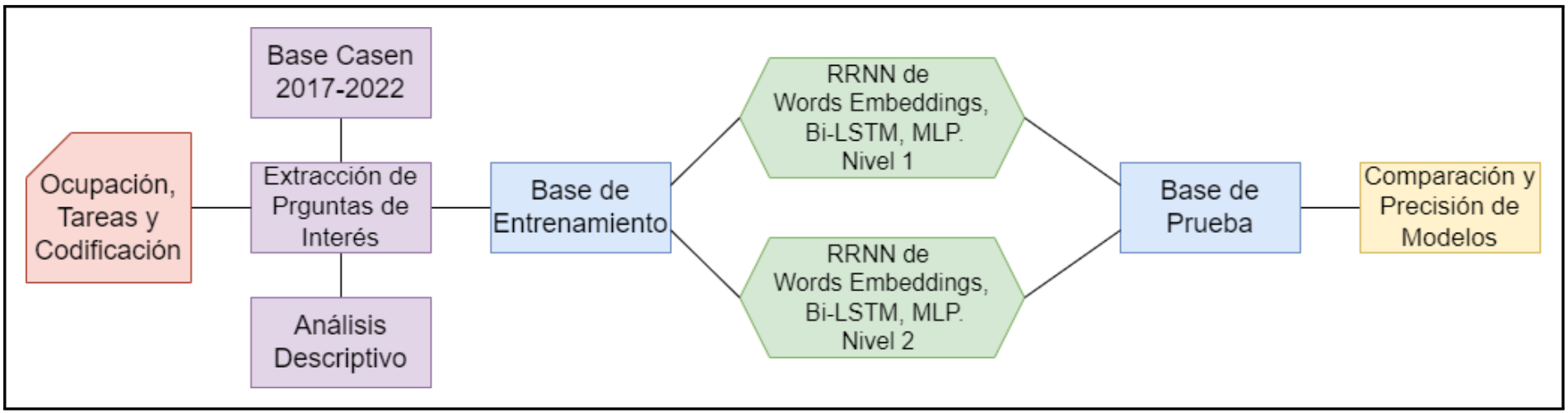


Figura 6: Esquema Metodológico

Esquema metodológico: Base de entrenamiento

¿Cómo se entrenará cada red? Para cada nivel de clasificación se implementará la misma arquitectura del modelo, calibrando diferentes hiperparámetros debido a la fase de entrenamiento se realiza con diferentes variables del conjunto de datos.

Ocupación + Tarea	CIUO_N1
⋮	⋮

Tabla 2: "Set de datos" para entrenar red de nivel 1

Ocupación + Tarea	CIUO_N1	CIUO_N2
⋮	⋮	⋮

Tabla 3: "Set de datos" para entrenar red de nivel 2

Además, al estar probando solo una arquitectura de modelo se optará por un enfoque clásico, donde el conjunto de datos para cada red neuronal, se particiona en un 85 % para entrenamiento y un 15 % de validación.

Descripción del Modelo

- Al poseer 4 niveles de codificación, se entrenarán 4 redes bajo el mismo esquema, utilizando el correspondiente set de datos para entrenar la red en cada nivel

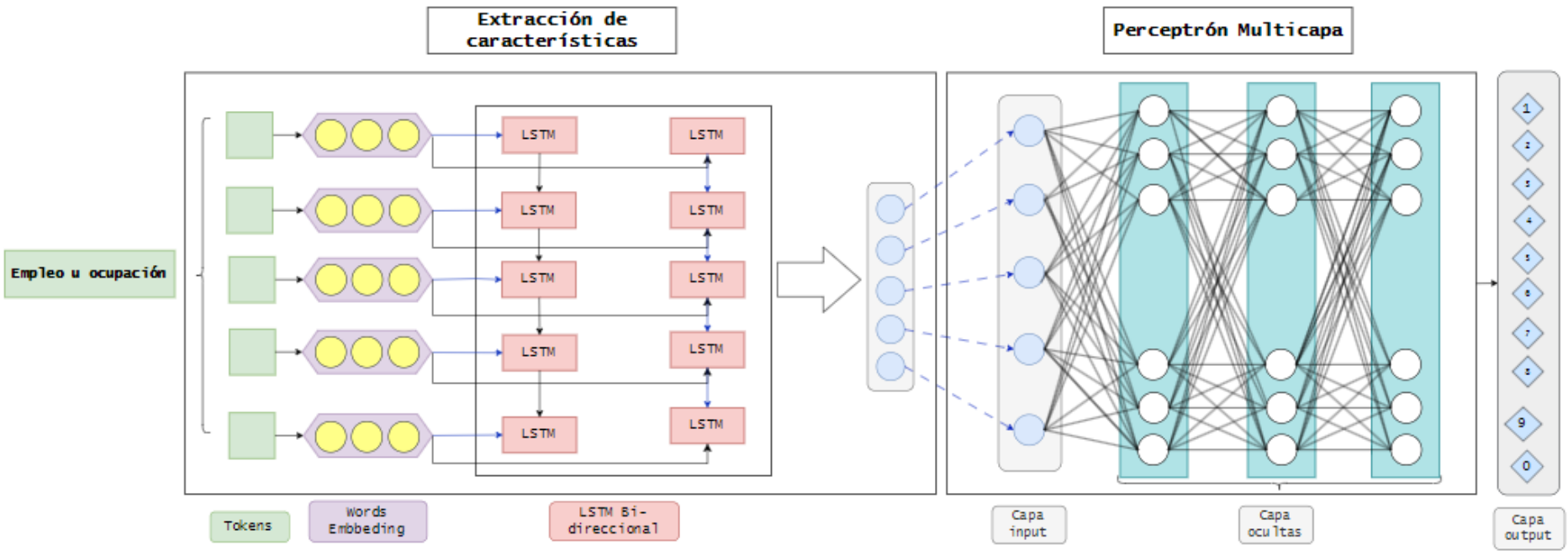


Figura 7: Metodología para el nivel uno de codificación

Descripción del Modelo

Estamos ante un problema de NLP, el cual consiste en recopilar respuestas de interés de la Encuesta CASEN y obtener su clasificación CIUO-08. Para ello se escogió la arquitectura,

Componente	Descripción	Detalle
Words-Embeddings	Representaciones vectoriales densas	• Contexto semántico • Operaciones algebraicas
bi-LSTM	Extracción de características	• *sparse_categorical_crossentropy*$=-\sum_{i=1}^{to} y_i \cdot \ln(\hat{y}_i)$ • 2 direcciones de procesamiento • Puerta de olvido: $f_t = \sigma(W_f x_t + U_f h_{t-1} + V_f c_{t-1} + b_f)$ • puerta de entrada $\begin{cases} i_t = \sigma(W_i x_t + U_i h_{t-1} + V_i c_{t-1} + b_i) \\ \tilde{c}_t = \tanh(W_c x_t + U_c h_{t-1} + V_c c_{t-1} + b_c) \\ c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \end{cases}$ • La puerta de Salida o determina el valor de salida de la unidad LSTM $\begin{cases} o_t = \sigma(W_o x_t + U_o h_{t-1} + V_o c_{t-1} + b_o) \\ h_t = o_t \odot \tanh(c_t) \end{cases}$
MLP	Clasificación de características	• Función de activación Relu • Dropout del 40 % • Función softmax para la capa de salida • Tasa de aprendizaje adaptativa.

Tabla 4: Arquitectura elegida para la clasificación

Referencias

1. Géron, Aurélien. 2022. Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow. O'Reilly Media, Inc
2. Instituto Nacional de Estadísticas. (2018). Clasificador Chileno de Ocupaciones CIUO 08.CL. Instituto Nacional de Estadísticas, Subdirección Técnica, Departamento de Infraestructura Económica, Sección de Nomenclaturas. Disponible en `https://www.ine.gob.cl/docs/default-source/buenas-practicas/clasificaciones/ciuo/clasificador/ciuo-08-cl.pdf`
3. Gautam, H. (2020, marzo 1). Word Embedding. Basics. `https://medium.com/@hari4om/word-embedding-d816f643140`
4. DataScientest. (s.f.). Memoria a largo plazo a corto plazo (LSTM): ¿Qué es?. DataScientest. (2024, mayo 20) `https://datascientest.com/es/memoria-a-largo-plazo-a-corto-plazo-lstm`