



UNIVERSIDAD BERNARDO O'HIGGINS
DIRECCIÓN GENERAL DE POSTGRADO
MAGISTER EN INGENIERÍA EN INFORMÁTICA

Priorización de Interconsultas Hospitalarias basadas en modelos de Inteligencia Artificial.

Oswaldo Mellado Beltrán

Profesor Guía: Broderick Crawford

Abstract

Las interconsultas hospitalarias son documentos de atención primaria para pacientes que poseen alguna enfermedad y son atendidos por la red de salud pública. El problema que existe en la actualidad es que dichas interconsultas deben ser priorizadas como urgentes o no. Para ello existe un equipo de especialistas que se encargan de su priorización. Este estudio pretende entregar el cómo un modelo de IA puede ayudar en dicha tarea.

Keywords: Interconsultas Hospitalarias, Bert, modelo IA

Santiago, Chile

2024

- estructuras neuronales
x capitulos
- [+] ; considerar
más referencias

Índice



Capítulo I

- 1.1 **Introducción**
- 1.2 **Objetivo General y Específicos.**
- 1.3 • **Desarrollo de los ítems temáticos de cada proyecto**
- 1.4 **Conclusiones.**
- 1.5 • **Referencias Bibliográficas**

Introducción

El servicio de salud metropolitano sur oriente (SSMSO) tiene dentro de sus funciones el registrar cada consulta que pueda generar un paciente en particular , junto con ello y debido a que la red de salud es muy extensa, produce que entre distintos centros hospitalarios se tenga que diagnosticar y examinar a una misma persona , para ello se creo el documento de interconsultas electrónicas para cada paciente (PULSO APP), dicha interconsulta es el resumen del tratamiento y las instrucciones generadas por cada medico de cada especialidad para un correcto diagnostico y tratamiento del paciente.

Como regla general dichas interconsultas ingresan al sistema después de una primera intervención médica, donde se genera y completa por el medico inicial, cuando la patología del paciente indica ciertos exámenes o tratamientos más específicos estas interconsultas son derivadas a las respectivas áreas donde el paciente deberá tratarse.

Debido a la cantidad de ordenes de interconsulta estas son priorizadas como urgentes o no, debido a este paso fundamental, un equipo especializado de médicos y contralores trabajan en determinar la urgencia de cada interconsulta (paciente) en relación a la gravedad del asunto.

Objetivo General

Si bien es cierto que las interconsultas se deben derivar en el menor plazo posible, la cantidad de interconsultas es muy alta, complejizando la correcta distribución de las urgencias de los pacientes, provocando retrasos en los diagnósticos, toma de muestras y consultas a los médicos especialistas traduciendo en que la propia interconsulta demore la efectividad de un tratamiento de salud.

Para solucionar dicho problema hemos propuesto que un modelo de inteligencia artificial pueda tomar la decisión de la urgencia asociada a una interconsulta, para ello la propia interconsulta cuenta con ciertos campos que son útiles para analizar la información escrita y emitir de manera más rápida una asignación de prioridad, en esta etapa del proyecto queremos abordar que nuestro modelo sea capaz de colocar si una interconsulta es urgente o no.

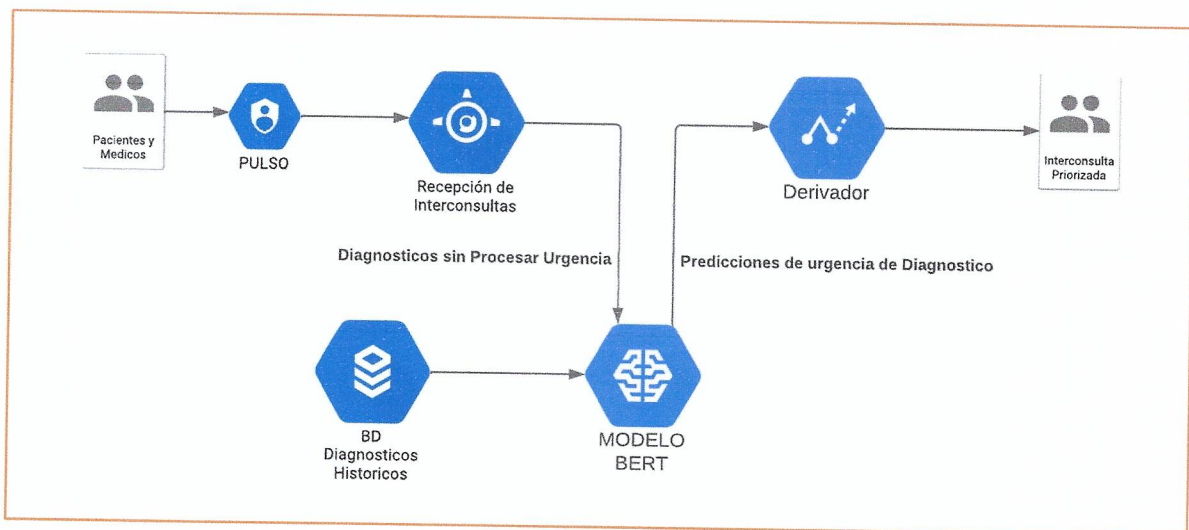


Fig # _____

Objetivos específicos:

Estudiar el flujo de la interconsulta actual y datos Útiles a utilizar en él modelo.

Una de las principales tareas para la realización de este trabajo es explicar ¿Por qué sería aplicable una solución de IA?, pues esto es básicamente el ayudar en la metodología de priorización de las interconsultas incluyendo un análisis real de los campos de información relacionados a la urgencia misma de la interconsulta y que los controladores médicos utilizan para basar su opinión de salida.

En el flujo de creación de la interconsulta el médico y paciente generan la primera entrevista donde en su ficha clínica se consignan los valores encontrados en su diagnóstico, junto con eso se crea una interconsulta seriada y asociada al paciente el cual es derivada a la especialidad que requiere el paciente a través del software institucional de la red llamado "PULSO", dicha derivación es analizada cuando llega al casilla de entrada de las interconsultas a priorizar para asignar una hora al médico especialista que corresponda. Dicha acción es realizada en la actualidad por un grupo de médicos controladores que indican la categoría de urgencia y son entregadas a un derivador para asignar la hora en base al estudio realizado. Finalmente, dicha hora es comunicada al paciente y agendada en base a la categorización antes entregada.



Fig. # _____

La interconsulta presenta ciertos campos que se emplean en la actualidad son usados para definir el nivel de urgencia de ellas. Como se logra observar en la lámina siguiente tenemos los siguientes campos:

CAMPOS ACTIVOS INTERCONSULTA	
Servicio de Salud	Prais
Establecimiento	Clasificación Previsional
Unidad	Se deriva para ser atendido en
Especialidad origen	Se deriva para
Nombre	Hipótesis Diagnostica
N° Archivo	Sospecha problema AUGE
Run	Fundamento del Diagnóstico y exámenes realizados
Run Recién Nacido	Datos del Profesional que deriva
Sexo	Fecha Interconsulta
Fecha y hora de nacimiento	Fecha de Atención
Edad	Diagnostico o sospecha clínica
Domicilio	Indicaciones
Comuna	Alta medica
Fono	Continúan en control
Celular	Datos del Profesional RUN
E-mail	

TABLA #

Como se aprecia en la tabla anterior la cantidad de campos que utiliza la interconsulta son varios , ello se explica que cada campo representa un estado de la Interconsulta en el transcurso del tiempo , se deduce que existe el campo de hipótesis diagnostica que es lo que cataloga la primera revisión del paciente y en el punto donde se genera la interconsulta dentro del flujo de la misma , luego tenemos el diagnostico o sospecha clínica , la cual da los fundamentos para realizar exámenes y otras acciones investigativas de la enfermedad o patología encontrada.

Dentro de la misma interconsulta también tenemos varios campos que en realidad no se utilizaran ya que para determinar la urgencia de la interconsulta no son necesarios los demás campos.

CAMPOS ACTIVOS INTERCONSULTA		Campo de la Base de datos
Servicio de Salud	Prais	
Establecimiento	Clasificación Previsional	
Unidad	Se deriva para ser atendido en	
Especialidad origen	Se deriva para	
Nombre	Hipótesis Diagnostica	IC_DIAG_SOSPECHA
Nº Archivo	Sospecha problema AUQE	
Run	Fundamento del Diagnóstico y exámenes realizados	
Run Recién Nacido	Datos del Profesional que deriva	
Sexo	Fecha Interconsulta	
Fecha y hora de nacimiento	Fecha de Atención	
Edad	Diagnostico o sospecha clínica	IC_DIAG_CONFIRMA
Domicilio	Indicaciones	
Comuna	Alta medica	
Fono	Continúan en control	
Celular	Datos del Profesional RUN	
E-mail		

Para seleccionar el set de datos nos concentraremos en los campos IC_DIAG_SOSPECHA y IC_DIAG_CONFIRMA, dichas columnas se deberían juntar (concatenar), para obtener un texto mas exacto en lo que respecta a la descripción del tratamiento del paciente a seguir y su posterior derivación.

Ejemplo de datos a transferir al modelo.

IC_DIAG_SOSPECHA + IC_DIAG_CONFIRMA
R33 RETENCION DE ORINA PCTE 85 AÑOS CON ANTEC HTA DM2 ARTROSIS OBS FA HOSPITALIZADA POR INFARTO CEREBRAL EVOLUCIONA CON ITU A REPETICIÓN, REQUERIMIENTOS DE SONDEO INTERMIENTE ACTUALMENTE CON Sonda FOLEY Y TAMSOLUSINA OBSERVACIONES: R33 RETENCION DE ORINA Pcte 85 años con antec HTA DM2 Artrosis obs FA Hospitalizada por infarto cerebral Evolucion a repetición, requerimientos de sondeo intermite Actualmente con sonda foley y tamsolusina
A51 SIFILIS PRECOZ PL CON CITOQUIMICO Y VDRL EN LCR TIENE PLAQUETAS 348 000 E INR 1 16 NO NECESITA TAC vdrl 1:128 pwi OBSERVACIONES: A51 SIFILIS PRECOZ PL CON CITOQUIMICO Y VDRL EN LCR Tiene Plaquetas 348 000 e INR 1 16 No necesita TAC
R32X INCONTINENCIA URINARIA, NO ESPECIFICADA DICE ORINAR MUCHO, SU FUNCION TENAL ES NORMAL NO VISTO POR POR UROLOGO ID OBSERVACIONES: R32X INCONTINENCIA URINARIA, NO ESPECIFICADA DICE ORINAR MUCHO, SU FUNCION TENAL ES NORMAL NO VISTO POR POR UROLOGO
COLICO RENAL am. noaqx. apendicectomía clásica alergias. metamizol evaluado hace 2 días por cuadro de dolor en hcd y flanco derecho. se le solicitó eco abdominal realizada hoy que informa: · moderada hidroureteronefrosis sin identificar un factor obstructivo mediante esta modalidad de estudio. (15.45 hrs) so leuc 11-21, hemat 0-2 pielotc: urolitiasis proximal derecha de 5 mm, asociado a moderada hun derecha plan: alta con tratamiento antibiótico / control en cdt urología / urgencia sos ***- por dr(a) fnt -**
Apnea (adulto) Apnea (adulto) sahos paciente de 53 años, sin ant conocidos, alergia no, acude debido a que le preocupan periodos de apnea nocturna que presenta, con periodos de apnea largos en donde se despierta agitado, roncador nocturno, que esto a ido aumentando con el tiempo, acude con exámenes generales los cuales estan normales pendiente ecg ex fisico bcg hde afebril hidratado clote rr2tss, mp + sra, abdomen bdl rha + normales masas- sin signos de irritacion peritoneal indicaciones derivó a neurología pbr obs sahos OBSERVACIONES: Apnea (adulto)
Paciente con EQZ de larga data, previamente usuario de AP de deposito + Clozapina Estable psicopatológicamente Desde hace años con postura de tronco en anteroflexión lateralizado a derecha, con desviación cefálica a derecha, sospechoso de distonia tardía de larga data Además, paciente es usuario de larga data de Tonaril (desde los tiempos de uso de AP de deposito) Dado que ya no lo usa, queda la duda del beneficio de mantener tonaril en cuadro actual Favor solicito evaluación diagnóstica y sugerencia de conducta a seguir De antemano muchas gracias OBSERVACIONES: G249 DISTONIA, NO ESPECIFICADA Derivo a Neurología Trastornos del Movimiento (Dr Juri o Dra Lazcano)
Epilepsia - EPILEPSIA; Usuaría hace 1 mes vive en Pirque y es fonasa (antes era ISAPRE) Antecedentes de Epilepsia desde el 2009 Trae documento donde se diagnostica Epilepsia Temporal Mesial Derecha por Dr Reinaldo Uribe San Martín Refiere que ultima crisis fue ayer, 4 episodios seguidos Además con Tr depresivo desde hace 2 años En tto con levetiracetam 1000mg 2 veces al día, acido valproico 500mg 2 veces al día, lamotrigina 50mg 2 veces al día y litio 600mg 2 veces al día Trae constancia GES Epilepsia no refractaria 15 años y más del 08 09 2017 del extrasistema Por lo que solicito evaluación por especialista para continuar tratamiento y seguimiento OBSERVACIONES: Realizar Tratamiento

De los campos anteriores hemos elegido solo los que requerimos para ser utilizados en nuestro modelo de IA, ello debido a que si bien es cierto que existen parámetros que hoy en día los médicos consideran, tomaremos solo los de diagnostico para nuestro modelo, dejando para un nuevo estudio la inclusión de dichos parámetros como son la edad, domicilio, sexo, etc.

Con lo anterior ya podríamos estimar que tenemos los datos necesarios para llevarlos al modelo de aprendizaje de nuestra IA.

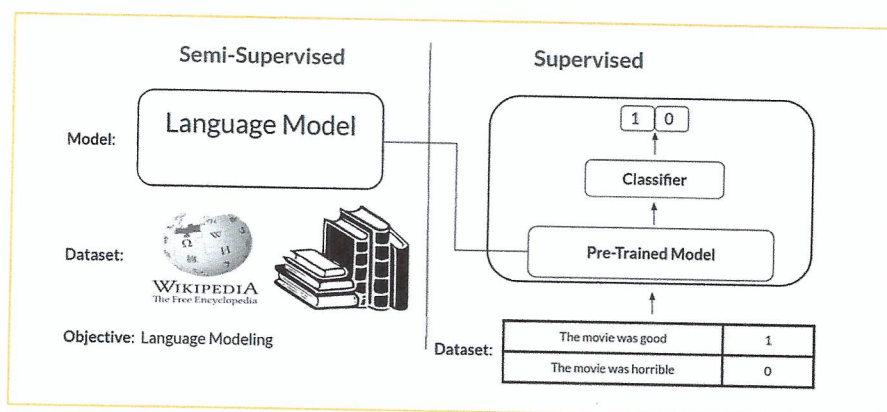
Pruebas de Laboratorio y preentrenamiento 1.-

Una vez construido el modelo y aplicativos en Google colab, procedimos a ejecutar el preentrenamiento del modelo con los datos mencionados en el apartado anterior, este laboratorio tiene la intención de verificar la cantidad de los datos, la calidad de los datos y la exactitud de los datos, para ello cargamos un Excel con 20.000 registros con sus respectivas salidas, para ello tomamos que en el modelo tendríamos dos tablas una llamada diagnóstico y el otro sentimiento. La columna sentimiento tienen la salida en base a la experiencia alcanzada en los datos de muestreo y esta escala como normal o urgente.

Hicimos el programa y lo ejecutamos dando por resultado un ciclo infinito y sin resultados positivos ni término del análisis de la data.

Conclusiones al preentrenamiento.

En primer lugar, al analizar el modelo Bert vemos que este se encuentra entrenado con mucha data como ve en la siguiente lamina.



El detalle que encontramos es que el dataset utilizado para alimentar el modelo proviene de Wikipedia, pero en el idioma Ingles, entonces nos preguntamos que el problema puede ser el Idioma, para ello tomamos nuestro dataset y los traducimos al inglés.

Este paso lo hicimos en Google traductor, y si bien, no es lo ideal hemos pensado que debemos colocar otra capa a nuestro modelo, y esta sería la de traducción automática al inglés para el análisis de los datos.

Una vez hecho ese proceso podemos observar la data final que incluiremos en una segunda revisión de laboratorio.

review	
R33 Urine Retention PCTE 85 years with HTA DM2 Antec AF Hospitalized due to cerebral infarction Evolves with repeated UTI, requires intermittent catheterization Currently with foley cath	feeling URGENT
A51 EARLY SYPHILIS PL WITH CYTOCHEMICAL AND VDRL IN CSF HAS PLATELETS 348,000 AND INR 1.16 DOES NOT NEED CT vdr1 1:128 pvti OBSERVATIONS: A51 EARLY SYPHILIS PL WITH CYTOCHEMICAL	NORMAL
R32X URINARY INCONTINENCE, UNSPECIFIED HE SAYS HE URINATES A LOT, HIS TENSILE FUNCTION IS NORMAL NOT SEEN BY UROLOGY ID OBSERVATIONS: R32X URINARY INCONTINENCE, UNSPECIFIED	URGENT
RENAL COLICO am. not herex. classic appendectomy allergies. metamizole evaluated 2 days ago due to pain in the hcd and right flank. An abdominal echo was requested today, which rep	NORMAL
Apnea (adult) Apnea (adult) oesus 53-year-old patient, no known ant, no allergy, comes because he is concerned about periods of nocturnal apnea that he presents, with long periods of	NORMAL
G249 DYSTONIA, UNSPECIFIED REFERRED TO NEUROLOGY MOVEMENT DISORDERS (OR JURI OR DRA LAZCANO) Patient with long-standing EQ2, previously user of depot AP + Clozapine Psychopa	URGENT
Epilepsy - EPILEPSY; User 1 month ago lives in Pirque and is a phonase (previously ISAPRE) History of Epilepsy since 2009 Brings document where Right Mesial Temporal Epilepsy was diagn	NORMAL
A530 LATENT SYPHILIS, NOT SPECIFIED AS EARLY OR LATE SYPHILIS, MEMORY FAILURES exam OBSERVATIONS: A530 LATENT SYPHILIS, NOT SPECIFIED AS EARLY OR LATE sypilis, memory failures	NORMAL
Tremor The patient reports a condition of 3-4 months of evolution characterized by tremor in the right hand, at rest and when carrying out activities, apparently there would be no trigger. I	NORMAL
HEMATURI OBS. MECHANICAL COMPONENT (Diagnostic Hypothesis) nsp patient consulted for long-term hematuria. Today's episode of 24 hours of evolution. without abdominal or back pai	NORMAL
Meningitis, Unspecified Tension-Featured Headache - Clear Liquid Meningitis Probably Viral -	NORMAL
G40 EPILEPSY PATIENT IN CONTROL WITH AC V 200 = 5 COMP PER DAY, CARBAMAZEPINE 200 = 3 5, CNZ 1MG AT NIGHT + VENLAFAXINE AND RISPERIDONE FOR MENTAL HEALTH MAINTAINS CRISES M	NORMAL
Testicular malignant neoplasia Testicular malignant neoplasia - 31-year-old patient with a history of severe esophageal hypomotility 1 week ago felt a local increase in volume in the	NORMAL
G62 OTHER POLYNEUROPATHIES patient under study for sensory polyneuropathy, presents with bladder urgency and urinary retention, evaluation is requested OBSERVATIONS: G62 OTHER PO	URGENT
RENAL COLICO VS AP UTI: denies medical and surgical aptologies. Denies allergy to medicines, patient reported right renal lithotripsy two years ago. He reported intense right lumbar pain	URGENT
OBS TRIGEMINAL NEURALGIA am psoriasis aqx cca meds (-) allergies (-) tbq (-) since yesterday with tremor in the right eyelid associated with intense stabbing pain of short duration 3-5 se	NORMAL
OBS FACIAL PARALYSIS OBS LUMBAGO healthy, no allergies mc: 1 week ago with lower back pain. without symptoms, neurological compromise or red flags. -lumbar spine with pain palpati	NORMAL
Peyronie's disease nc Peyronie's disease npatient referred to mental hospital due to suspicion of depression, history of hta and dlp reports alteration of penile curvature of more or les	NORMAL
URINARY RETENTION. ant doctors: dm ii - hta ant qc: inguinal hernia - appendectomy allergies (-) pcte with urinary retention. 09h50 place Foley probe. ***- by dr(a) gtf -***	URGENT
D410 TUMOR OF UNCERTAIN OR UNKNOWN BEHAVIOR OF THE KIDNEY COMPLEX RENAL CYST Patient in post-hernioplasty controls in abdominal ultrasound a complex renal cyst was found, ew	NORMAL
Inguinal hernia nc - right and umbilical herni Inguinal hernia nc - right and umbilical herni patient who for approximately 1 year has been experiencing pain in the periumbilical and right	NORMAL
Bladder stone - BLADDER STONE; 61 years old Morbid history: sequela stroke, GTT user, cystostomy, HBP, epilepsy, DM2 IR Med allergies: no Patient belonging to a bedridden program, CU	URGENT
Brain Tumor Left Frontal Tumor Obs. Oligodendroglioma Left Frontal Tumor - Obs. Oligodendroglioma Opportunistic infection - Klebsiella UTI ESBL Epilepsies - Secondary epilepsy oral ca	URGENT
CHRONIC HEADACHE Patient with long-standing chronic migraine-type headache, poor response to analgesics increase in frequency and intensity OBSERVATIONS: CHRONIC HEADACHE	NORMAL
MALIGNANT TUMORS OF THE PROSTATE Patient referred from PSA due to hematuria Patient operated on for prostate cancer in January 2018 presents with hematuria of 3 months of evolutio	NORMAL
N19X UNSPECIFIED RENAL FAILURE CRF IN HD LEFT PQ TESTICLE evaluation OBSERVATIONS: N19X UNSPECIFIED RENAL FAILURE CRF in HD left PQ testicle	URGENT

Como se aprecia en la lamina anterior los datos resultantes están en optimas condiciones para ingresar al preentrenamiento.

Elección del modelo a emplear

BERT es el elegido para ayudarnos en desarrollar nuestro proyecto esto es porque BERT, que significa Bidirectional Encoder Representations from Transformers, es un modelo de lenguaje basado en transformadores desarrollado por Google. Fue presentado en 2018 y es una de las arquitecturas de modelos de lenguaje más avanzadas hasta la fecha.

Bidirectional Encoder Representations from Transformers:

- **Encoder Representations:** sistema de modelado del lenguaje, pre entrenado con datos sin etiquetar. Luego se ajustan sus parámetros.
- **from Transformer:** basado en un poderoso algoritmo de NLP. Define la arquitectura de BERT.
- **Bidirectional:** usa un contexto de izquierdas y derechas al trabajar con las palabras. Define el proceso de entrenamiento.

Lo que hace a BERT particularmente poderoso es su capacidad para comprender el contexto y la relación entre las palabras en una oración. A diferencia de los modelos anteriores que procesaban el texto de izquierda a derecha o de derecha a izquierda, BERT utiliza un enfoque bidireccional, lo que significa que considera el contexto tanto antes como después de cada palabra al hacer predicciones sobre la siguiente palabra en una oración. Esto le permite capturar mejor las sutilezas del significado y la semántica en el texto.

BERT ha demostrado ser muy efectivo en una variedad de tareas de procesamiento del lenguaje natural (NLP), como el reconocimiento de entidades, la clasificación de texto, la traducción automática y la respuesta a preguntas, entre otras. Además, se puede pre-entrenar en grandes cantidades de texto sin supervisión, lo que le permite capturar un conocimiento lingüístico profundo que luego puede ser transferido a tareas específicas con solo un ajuste fino en conjuntos de datos más pequeños y específicos.

Arquitectura BERT

La arquitectura BERT es un modelo de lenguaje basado en la arquitectura de Transformers, que consiste en codificadores y decodificadores. A diferencia de los modelos de lenguaje anteriores que procesaban el texto de izquierda a derecha o de derecha a izquierda, BERT utiliza un enfoque bidireccional, lo que le permite capturar mejor el contexto y la relación entre las palabras en una oración.

Descripción general de la arquitectura de BERT:

La arquitectura Bert posee diferentes capaz de funcionalidad siendo las principales:

Codificador Transformer: BERT utiliza únicamente los codificadores de la arquitectura Transformer. Los codificadores se componen de múltiples capas, cada una de las cuales realiza dos tareas principales: la atención multi-cabeza y la transformación de las características. La atención multi-cabeza permite al modelo asignar diferentes pesos a diferentes partes de la entrada, mientras que la transformación de las características ayuda a modelar y procesar la información en cada capa.

Capas de codificación: BERT emplea una pila de capas de codificación (Encoder layers), donde cada capa aplica la atención multi-cabeza y la transformación de características a la entrada. Cada capa tiene conexiones residuales alrededor de las operaciones de atención y transformación, lo que ayuda a mitigar el problema del desvanecimiento del gradiente durante el entrenamiento.

Modelo pre-entrenado: BERT se entrena prestando atención a la tarea de predicción de palabras enmascaradas (Masked Language Model, MLM) y a la tarea de predicción de la siguiente oración. En la tarea de MLM, se ocultan aleatoriamente ciertas palabras en las oraciones de entrada y el modelo debe predecir cuáles son estas palabras en base al contexto proporcionado por las palabras circundantes. En la tarea de predicción de la siguiente oración, el modelo debe determinar si dos oraciones dadas son secuenciales o no.

Capa de salida: La salida del modelo BERT puede ser utilizada en una variedad de tareas de procesamiento del lenguaje natural (NLP), como la clasificación de texto, el etiquetado de entidades, la traducción automática, entre otras. Dependiendo de la tarea específica, se puede agregar una capa de salida adicional para ajustar finamente el modelo pre-entrenado en un conjunto de datos particular.

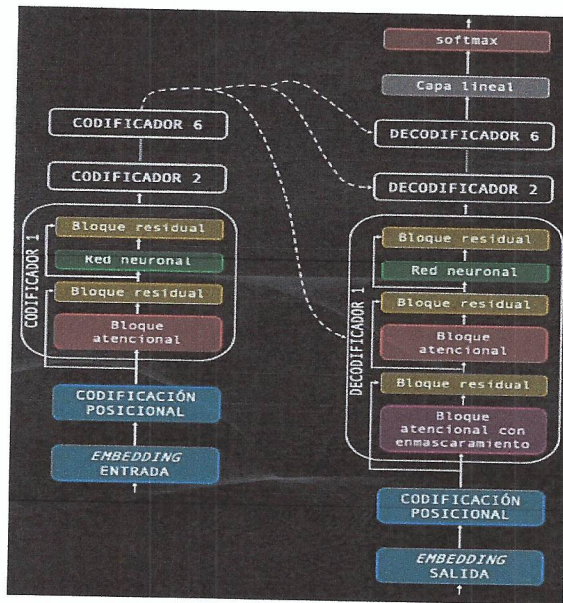


Fig # _____

En resumen, la arquitectura de BERT se basa en los codificadores de la arquitectura Transformer y utiliza un enfoque bidireccional para capturar el contexto y la relación entre las palabras en un texto. Esto le permite capturar un conocimiento lingüístico profundo que puede ser transferido a una variedad de tareas de procesamiento del lenguaje natural mediante el ajuste fino en conjuntos de datos específicos.

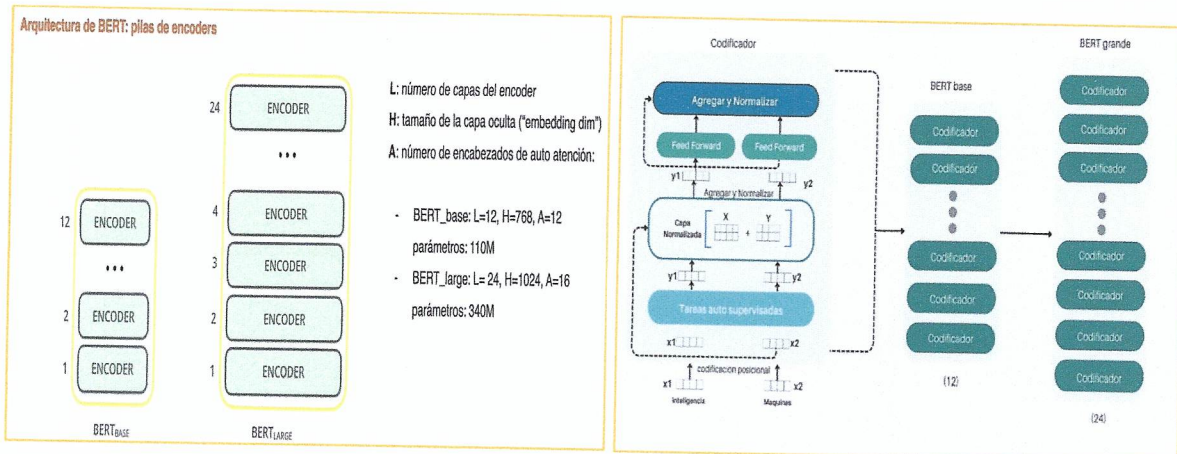


Fig # _____

Para el reconocimiento de la acción de la frase Bert utiliza token, estos funcionan de manera que el modelo entiende los verbos, con ello puede empezar a parametrizar las frases en cercanas y lejanas.

Tokenizador de BERT:

Tokenizador WordPiece:

- 30,522 "palabras"
- Trata una nueva palabra combinando palabras conocidas
- Compensación entre el tamaño del vocabulario y la palabra fuera del vocabulario
- Cada uno corresponde a un ID, por lo que de strings obtenemos números, utilizables por ordenadores

```
1 tokenizer.tokenize("I love Transformer")  
['i', 'love', 'transform', '##er']
```

```
1 tokenizer.tokenize("I love covfefe")  
['i', 'love', 'co', '##v', '##fe', '##fe']
```

Next sentence prediction:

A partir de dos frases A y B (separadas por [SEP]): IsNext/NotNext.

A: Ecology is one of the biggest challenge of this century.

B: Reducing our energy consumption could help saving living beings



IsNext

A: Ecology is one of the biggest challenge of this century.

B: Camels have three eyelids to help protect themselves from sand blowing in their eyes.



NotNext

Desarrollo del aplicativo de preentrenamiento

Para nuestro ambiente de desarrollo utilizaremos visual studio code y Python , este es un ambiente de desarrollo el que después será replicado en el servidor de priorización para utilizarlo en nivel de producción.

Nuestro Programa

En la creación del código para el modelo iremos por cada parte del programa para explicar la función de cada trozo de código.

```
Priorizacion_Interconsultas.ipynb x
C: > Users > Osvaldo > Desktop > Proyectos > Magister > Proyecto de titulo > Priorizacion_Interconsultas.ipynb > # Entrenamiento!!!
+ Código + Markdown | □ Interrumpir ↻ Reiniciar ☰ Borrar todas las salidas ↻ Ir a | 📄 Variables ☰ Esquema ...
```

```
[2] ✓ 11.2s
from transformers import BertModel, BertTokenizer, AdamW, get_linear_schedule_with_warmup
import torch
import numpy as np
from sklearn.model_selection import train_test_split
from torch import nn, optim
from torch.utils.data import Dataset, DataLoader
import pandas as pd
from textwrap import wrap
import matplotlib.pyplot as plt
```

```
[3] ✓ 0.0s
# Inicialización
RANDOM_SEED = 42
MAX_LEN = 200
BATCH_SIZE = 16
DATASET_PATH = r'C:\Users\Osvaldo\Desktop\Proyectos\Magister\Proyecto de titulo\2024_2_Ingles.xlsx'
NCLASSES = 2
np.random.seed(RANDOM_SEED)
torch.manual_seed(RANDOM_SEED)
device = torch.device("cuda:0" if torch.cuda.is_available() else "cpu")
print(device)
```

```
[4] ✓ 1.0s
df = pd.read_excel(DATASET_PATH)
df = df[0:20000]
```

```
print(df.head())
print(df.shape)
#print("\n".join(wrap(df['review'][:200])))
```

[5] ✓ 0.0s

```
...
                                review feeling
0  R33 Urine Retention PCTE 85 years with HTA DM... URGENT
1  A51 EARLY SYPHILIS PL WITH CYTOCHEMICAL AND VD... NORMAL
2  R32X URINARY INCONTINENCE, UNSPECIFIED HE SAY... URGENT
3  RENAL COLICO am. not herex. classic appendecto... NORMAL
4  Apnea (adult) Apnea (adult) oesus 53-year-old ... NORMAL
(14124, 2)
```

```
# Reajustar dataset
df['label'] = (df['feeling']=='URGENT').astype(int)
df.drop('feeling', axis=1, inplace=True)
df.head()
```

[6] ✓ 0.0s

	review	label
0	R33 Urine Retention PCTE 85 years with HTA DM...	1
1	A51 EARLY SYPHILIS PL WITH CYTOCHEMICAL AND VD...	0
2	R32X URINARY INCONTINENCE, UNSPECIFIED HE SAY...	1
3	RENAL COLICO am. not herex. classic appendecto...	0
4	Apnea (adult) Apnea (adult) oesus 53-year-old ...	0

```
# TOKENIZACIÓN
```

```
PRE_TRAINED_MODEL_NAME = 'bert-base-cased'
tokenizer = BertTokenizer.from_pretrained(PRE_TRAINED_MODEL_NAME)
```

[7] ✓ 1.2s

```
# Ejemplo tokenización
sample_txt = ' UNSPECIFIED REFERRED TO NEUROLOGY MOVEMENT DISORDERS (DR JURI OR DRA LAZCANO) Patient with long-standing EQZ, previously user of depo'
tokens = tokenizer.tokenize(sample_txt)
token_ids = tokenizer.convert_tokens_to_ids(tokens)
print('Frase: ', sample_txt)
print('Tokens: ', tokens)
print('Tokens numéricos: ', token_ids)
```

[8] ✓ 0.0s

```
...
Frase: UNSPECIFIED REFERRED TO NEUROLOGY MOVEMENT DISORDERS (DR JURI OR DRA LAZCANO) Patient with long-standing EQZ, previously user of depo
Tokens: ['UN', '##SP', '##EC', '##IF', '##IE', '##D', 'R', '##EF', '##ER', '##RE', '##D', 'TO', 'NE', '##UR', '##DL', '##O', '##G', '##Y', 'M', '##O', '##VE', '##']
Tokens numéricos: [7414, 15735, 8231, 15499, 17444, 2137, 155, 14663, 9637, 16941, 2137, 16972, 26546, 19556, 13901, 2346, 2349, 3663, 150, 2346, 17145, 14424, 15]
```



```
# Codificación para introducir a BERT
encoding = tokenizer.encode_plus(
    sample_txt,
    max_length = 10,
    truncation = True,
    add_special_tokens = True,
    return_token_type_ids = False,
    pad_to_max_length = True,
    return_attention_mask = True,
    return_tensors = 'pt'
)
```

[9] ✓ 0.0s

... C:\Users\Oswaldo\AppData\Roaming\Python\Python312\site-packages\torch\tensor.py:2645: FutureWarning: warnings.warn()

```
encoding.keys()
```

[0] ✓ 0.0s

```
dict_keys(['input_ids', 'attention_mask'])
```

```
print(tokenizer.convert_ids_to_tokens(encoding['input_ids'][0]))
print(encoding['input_ids'][0])
print(encoding['attention_mask'][0])
```

[11] ✓ 0.0s

```
... [['CLS'], 'UN', '##SP', '##EC', '##IF', '##IE', '##D', 'R', '##EF', '[SEP]']
tensor([ 101, 7414, 15735, 8231, 15499, 17444, 2137, 155, 14663, 102])
tensor([1, 1, 1, 1, 1, 1, 1, 1, 1, 1])
```



```

# CREACIÓN DATASET

class IMDBDataset(Dataset):

    def __init__(self, reviews, labels, tokenizer, max_len):
        self.reviews = reviews
        self.labels = labels
        self.tokenizer = tokenizer
        self.max_len = max_len

    def __len__(self):
        return len(self.reviews)

    def __getitem__(self, item):
        review = str(self.reviews[item])
        label = self.labels[item]
        encoding = tokenizer.encode_plus(
            review,
            max_length = self.max_len,
            truncation = True,
            add_special_tokens = True,
            return_token_type_ids = False,
            pad_to_max_length = True,
            return_attention_mask = True,
            return_tensors = 'pt'
        )

        return {
            'review': review,
            'input_ids': encoding['input_ids'].flatten(),
            'attention_mask': encoding['attention_mask'].flatten(),
            'label': torch.tensor(label, dtype=torch.long)
        }

```

[12] ✓ 0.0s

```

# Data Loader:

def data_loader(df, tokenizer, max_len, batch_size):
    dataset = IMDBDataset(
        reviews = df.review.to_numpy(),
        labels = df.label.to_numpy(),
        tokenizer = tokenizer,
        max_len = MAX_LEN
    )

    return DataLoader(dataset, batch_size = BATCH_SIZE, num_workers = 4)

```

[13] ✓ 0.0s

```

df_train, df_test = train_test_split(df, test_size = 0.2, random_state=RANDOM_SEED)

train_data_loader = data_loader(df_train, tokenizer, MAX_LEN, BATCH_SIZE)
test_data_loader = data_loader(df_test, tokenizer, MAX_LEN, BATCH_SIZE)

```

[14] ✓ 0.0s

```

# EL MODELO!

class BERTSentimentClassifier(nn.Module):

    def __init__(self, n_classes):
        super(BERTSentimentClassifier, self).__init__()
        self.bert = BertModel.from_pretrained(PRE_TRAINED_MODEL_NAME)
        self.drop = nn.Dropout(p=0.3)
        self.linear = nn.Linear(self.bert.config.hidden_size, n_classes)

    def forward(self, input_ids, attention_mask):
        _, cls_output = self.bert(
            input_ids = input_ids,
            attention_mask = attention_mask
        )
        drop_output = self.drop(cls_output)
        output = self.linear(drop_output)
        return output

```

[15] ✓ 0.0s

```

model = BERTSentimentClassifier(NCLASSES)
model = model.to(device)

```

[16] ✓ 1.4s

```

print(model)

```

[17] ✓ 0.0s

```

... BERTSentimentClassifier(
  (bert): BertModel(
    (embeddings): BertEmbeddings(
      (word_embeddings): Embedding(28996, 768, padding_idx=0)
      (position_embeddings): Embedding(512, 768)
      (token_type_embeddings): Embedding(2, 768)
      (LayerNorm): LayerNorm((768,), eps=1e-12, elementwise_affine=True)
      (dropout): Dropout(p=0.1, inplace=False)
    )
    (encoder): BertEncoder(
      (layer): ModuleList(
        (0-11): 12 x BertLayer(
          (attention): BertAttention(
            (self): BertSelfAttention(
              (query): Linear(in_features=768, out_features=768, bias=True)
              (key): Linear(in_features=768, out_features=768, bias=True)
              (value): Linear(in_features=768, out_features=768, bias=True)
              (dropout): Dropout(p=0.1, inplace=False)
            )
            (output): BertSelfOutput(
              (dense): Linear(in_features=768, out_features=768, bias=True)
              (LayerNorm): LayerNorm((768,), eps=1e-12, elementwise_affine=True)
              (dropout): Dropout(p=0.1, inplace=False)
            )
          )
        )
      )
    )
  )
  (drop): Dropout(p=0.3, inplace=False)
  (linear): Linear(in_features=768, out_features=2, bias=True)
)

```