Identificación de reacciones adversas a medicamentos en redes sociales basada en recuperación de información

Por:

José Alberto Fuentes Carbajal

Tesis sometida como requisito parcial
para obtener el grado de:

MAESTRÍA EN CIENCIAS EN EL ÁREA DE CIENCIAS COMPUTACIONALES

en el

Instituto Nacional de Astrofísica,
Óptica y Electrónica

Marzo, 2023

Tonantzintla, Puebla

Dirigida por:

Dr. Manuel Montes-y-Gómez

Dr. Luis Villaseñor-Pineda

Tabla de contenido

Lista de figuras

Lista de tablas

Agradecimientos

Esta investigación fue realizada gracias al apoyo otorgado por el Consejo Nacional de Ciencia y Tecnología (CONACYT), a través de la Beca No. 1080699. De igual forma, agradezco inmensamente a mis asesores el Dr. Manuel Montes y Gómez y al Dr. Luis Villaseñor Pineda por todo el conocimiento aportado, su constancia, disposición y compromiso con la investigación, permitiendo que el presente trabajo se realizara con éxito. Gracias por todos sus comentarios y críticas constructivas.

También agradezco a mis sinodales: Dr. Jesús Ariel Carrasco Ochoa, Dr. Hugo Jair Escalante Balderas y Dr. José Francisco Martínez Trinidad por brindarme su orientación y experiencia, permitiendo la mejora del trabajo realizado. Agradezco al INAOE, a sus trabajadores y a todos los profesores que nos trasmitieron sus conocimientos para poder alcanzar la meta final.

Dedicatoria

A mi familia

María, Benjamín y Miguel Angel

por todo el esfuerzo, sacrificios y apoyo incondicional a lo largo de mi vida.

A mi pareja, Analuz

por el apoyo incondicional, por tu comprensión, consejos y palabras de aliento a lo largo de este camino.

Resumen

El rápido crecimiento de las redes sociales se ha convertido en una fuente de información útil para diferentes tareas debido a que los usuarios publican información valiosa sobre varios aspectos de su vida, incluida la atención médica. Esta forma de intercambiar opiniones y experiencias ha proporcionado una rica fuente de información sobre los medicamentos y su eficacia y, lo que es más importante, sus posibles reacciones adversas. Las reacciones adversas a medicamentos (RA) son una causa importante de morbilidad en pacientes hospitalizados y una carga financiera para los sistemas de salud. Para abordar esta problemática, diferentes métodos se han desarrollado para la identificación de RA que van desde una simple clasificación de textos hasta la normalización de los tramos de texto que mencionen dicha RA a un diccionario médico por medio de técnicas de Aprendizaje Computacional y Aprendizaje Profundo. Estos últimos suelen tener buenos resultados, sin embargo, los conjuntos de datos existentes (específicamente de Twitter) cuentan con unos pocos miles de tweets etiquetados como positivos en comparación con los negativos. Además de esto, los problemas asociados a las redes sociales como la polisemia de las palabras, el lenguaje coloquial y los errores ortográficos dificultan el correcto entrenamiento de los sistemas. Con base en lo anterior, en el presente trabajo proponemos un método alternativo de dos etapas para la normalización de RA. Para la primera etapa utilizamos un modelo de Transformer pre-entrenado con el que abordaremos

la clasificación como una clasificación de pares de oraciones al considerar oraciones auxiliares generadas a partir de los tweets de entrada como información contextual adicional y como segunda etapa utilizamos un sistema de recuperación de información con el que normalizaremos las menciones de RA a términos médicos con ayuda de un diccionario. Los diferentes experimentos realizados muestran un enfoque diferente al estado del arte mediante un sistema de recuperación, brindando un panorama distinto para construir herramientas que ayuden a expertos en la fármaco-vigilancia.

Abstract

The rapid growth of social networks has made them to become a useful data source for different tasks as users post valuable information about various aspects of their lives, including medical care details. This way of exchanging opinions and experiences has provided a rich source of information about drugs and their efficacy and, more importantly, their possible adverse drug reactions. Adverse drug reactions (ADR) are a major cause of patient morbidity and a source of financial burden for healthcare systems. To address this problem, different methods have been developed for the identification of ADRs ranging from simple text classification to the normalization of the text sections mentioning such ADR to a medical dictionary by means of Machine Learning and Deep Learning techniques. The latter usually have good results, however, existing datasets specifically from Twitter have a few thousand tweets labeled as positive compared to negative ones, in addition to the problems associated with social networks such as polysemy of words, colloquial language and misspellings that hinder the correct training of the systems. Based on the above, in the present work, we propose an alternative two-stage method for the normalization of ADRs. For the first stage, we use a pre-trained Transformer model with which we will approach the classification as a sentence pair classification by considering auxiliary sentences generated from the input tweets as additional contextual information and, as a second stage, we use an information retrieval system with which we will

normalize mentions of ADRs to medical terms, helped by a dictionary. The different experiments performed show a different approach to the state of the art through a retrieval system, providing a different approach to build tools to help experts in pharmacovigilance.

Capítulo 1

Introducción

La identificación de reacciones adversas a medicamentos (RA) es un tema de gran importancia en el ámbito de la salud pública, ya que las RA son una de las principales causas de morbimortalidad en todo el mundo. Estas reacciones son una carga importante para los recursos sanitarios, en países occidentales se estima que las reacciones graves ocurren en el 6.7% de pacientes hospitalizados y son responsables de entre 5% a 9% de los costos de hospitalización (Kohn et al., 2000; Lazarou et al., 1998). Una reacción adversa es definida por la Organización Mundial de la Salud (OMS) como "*Una respuesta a un fármaco que es nociva y no intencionada, y que se produce a dosis normalmente utilizadas en el hombre para la profilaxis, el diagnóstico o el tratamiento de una enfermedad, o para la modificación de funciones fisiológicas*" Lee (2008). Sin embargo, esta definición suele ser un poco ambigua, por lo que una definición más precisa es dada por el Centro de Monitoreo Uppsala que la define como "*Una reacción apreciablemente dañina o desagradable, resultante de una intervención relacionada con el uso de un medicamento, que predice el peligro de la administración futura y justifica la prevención, el tratamiento específico, la alteración del régimen de dosificación o retiro del producto*" Edwards and Aronson (2000).

Identificar los efectos negativos de los medicamentos es un proceso arduo y complicado que puede disminuir significativamente el impacto que estos tienen en la salud humana. Los ensayos clínicos son uno de los principales métodos utilizados antes de la comercialización de medicamentos para detectar y medir los riesgos asociados a ellos. Sin embargo, estos ensayos presentan limitaciones importantes, como una muestra limitada, un período de prueba corto y la falta de diversidad entre los pacientes evaluados. Por lo que, se han buscado otras opciones para obtener información valiosa que permita identificar los problemas asociados a los medicamentos, incluyendo el uso de registros médicos no estructurados. Sin embargo, el acceso a estos registros a menudo resulta complicado debido a la privacidad con la que se manejan.

Hoy en día, existen recursos poco explorados pero valiosos en internet donde los internautas comparten experiencias utilizando medicamentos, por lo que analizar esta información es de suma importancia, ya que pueden contener información relevante relacionada con el consumo, uso incorrecto, combinación de diferentes medicamentos y la disminución o aumento de la dosificación recomendada (Sultana et al., 2013). De manera que mantener fuentes alternas de monitoreo puede ser de gran utilidad, ayudando a reducir los riesgos para los grupos más vulnerables (Pirmohamed et al., 2004; Sultana et al., 2013). Debido a esto, el monitoreo activo de medicamentos comúnmente llamado *fármaco-vigilancia* juega un papel importante, su objetivo es supervisar, evaluar y prevenir los riesgos asociados a medicamentos, así como la extracción de autoinformes y el desarrollo de estudios observacionales prospectivos de cohortes o retrospectivos de base de datos con el fin de contribuir al uso seguro de medicamentos. Esto permite dar un seguimiento prolongado a los pacientes con una gama más amplia de características, proporcionando medios valiosos para la detección temprana, identificación de factores de riesgo, estimación riesgo/beneficio o falta de eficacia; así como una reducción en los costos de atención médica (Coloma et al., 2013; Berlin et al., 2008).

La manera tradicional de identificar las RA posterior al lanzamiento es mediante historias clínicas de pacientes. Estas historias son registros médicos sistemáticos que incluyen todas las condiciones médicas, enfermedades y tratamientos pasados, con énfasis en los eventos específicos que afectan al paciente durante el episodio actual de atención. La información contenida en estos documentos es creada por todos los profesionales de la salud que brindan atención al paciente y es utilizada para dar continuidad a un tratamiento. Por lo que, siguiendo este registro se puede analizar los factores que están afectando al paciente y así concluir si se trata de una reacción adversa que debe ser informada. Sin embargo, uno de los principales problemas que se tiene con estos registros es el acceso debido a la privacidad con la que se manejan y la pequeña cantidad existente debido al poco o nulo seguimiento de los pacientes, por lo que utilizar fuentes alternativas de información pública es importante para un monitoreo activo.

El uso de fuentes de información como Internet es de suma importancia debido al gran número de personas que comparten información día a día. De acuerdo con las estadísticas reportadas por la Asociación de Internet[1] tan solo en el año 2020 en México se registró un total de 81.1 millones de personas activas en internet. Además, éste estudio reportó que las redes sociales son las páginas más utilizadas con un total del 86% del total de los internautas, seguidas por las aplicaciones de mensajería con un 83.9%. De la misma manera, de acuerdo con las estadísticas de la Unión Internacional de Telecomunicaciones[2] (ITU, por sus siglas en inglés) en el año 2021 el 64.05% equivalente a 4.9 millones de personas a nivel global se encontraban utilizando internet, reportando de la misma manera que las redes sociales fueron las páginas más visitadas. Con base en estas estadísticas se pudo observar que los usuarios suelen

[1] https://www.asociaciondeinternet.mx

[2] www.itu.int

invertir su tiempo en páginas como Facebook[3], WhatsApp[4], YouTube[5], Instagram[6], Twitter[7], Telegram[8], etc.

El uso de redes sociales como las antes mencionadas nos ha permitido conectarnos con una mayor cantidad de personas con las que podemos interactuar de manera inmediata y así compartir información en un corto periodo de tiempo, por lo que esto las hace adecuadas para compartir noticias, ideas, documentos y experiencias como el uso de medicamentos. Debido a esto, páginas como Twitter han cobrado gran relevancia, ya que los usuarios suelen compartir todo tipo de experiencias y opiniones. Lo que las ha convertido en fuentes de información de interés para una gran cantidad de investigadores enfocados en el área de la salud, gracias a su atmósfera comunicativa y colaborativa donde pacientes, médicos, instituciones e investigadores intercambian información de forma libre; como se ha demostrado en diferentes estudios (Raghupathi and Raghupathi, 2014; Edo-Osagie et al., 2020; Ginn et al., 2014; O'Connor et al., 2014). Algunos de los temas de investigación que han cobrado relevancia son la identificación de reacciones adversas a medicamentos (Nikfarjam et al., 2015), identificación de trastornos mentales (Coppersmith et al., 2014), trastorno bipolar (Thompson et al., 2015), abuso de drogas (Hu et al., 2019), síntomas COVID-19 (Lian et al., 2022), etc. Sin embargo, existen riesgos asociados al uso de redes sociales como fuente de información, sobre todo en temas de cohorte médico, ya que pueden incluir: altas tasas de información errónea, dificultades para verificar las fuentes, grandes volúmenes de información, etc. Adicional a esto, el uso incorrecto del lenguaje, faltas ortográficas y uso excesivo de abreviaturas, suelen ser los principales problemas al momento de analizar dicha información.

[3]https://www.facebook.com
[4]https://www.whatsapp.com
[5]https://www.youtube.com/
[6]https://www.instagram.com
[7]https://twitter.com/
[8]https://web.telegram.org/

Un ejemplo de esto puede ser observado en los siguientes tweets.

"*@user I've had Cipro before. Luckily for me, the only side fx I tend to get from AB is gastic upset. But I RARELY use AB.*", "*My medication is making me stupid, brain meds are not ok, fuck u paxil.*", "*the pathway of cipro is definitely murdering my mind right now? #donedonedone*"

donde podemos notar que el uso de lenguaje coloquial, errores ortográficos, uso de abreviaturas, errores gramaticales y uso de lenguaje ofensivo suele predominar en este tipo de redes sociales. Por lo que es importante aplicar diferentes enfoques y criterios al momento de utilizar este tipo de información debido al conocimiento clínico necesario. Hoy en día se han realizado grandes esfuerzos por desarrollar un conjunto de datos para la identificación de RA en Twitter. Sin embargo, debido a la necesidad de conocimiento clínico y el tiempo que toma etiquetar un conjunto de datos ha provocado que los datos existentes sean limitados y cuenten con un des-balance entre clases.

Debido a lo antes mencionado, en este trabajo abordaremos la identificación de reacciones adversas a medicamentos como un problema de dos etapas. Con la primera etapa pretendemos solventar los problemas relacionados con el lenguaje coloquial, los errores ortográficos y errores gramaticales, por medio de una clasificación binaria donde el objetivo es identificar si un tweet contiene al menos una reacción adversa a medicamento o no, etiquetando como Ra si contiene al menos una reacción adversa o como NoRa en caso contrario. Para abordar la tarea de clasificación proponemos un enfoque basado en *Transformers* pre-entrenado debido a la cantidad de datos con la que contamos y a las ventajas con las que cuentan los *transformers* para procesar texto coloquial. Adicional a esto, éste enfoque fue elegido basado en trabajos previos que han demostrado que el uso de *transformers* como BERT han logrado obtener resultados sobresalientes en esta tarea (Ramesh et al., 2021; Sakhovskiy et al., 2021; Aji et al., 2021a). Sin embargo, a diferencia

de los trabajos previos, que utilizan BERT[9] de manera tradicional para tareas de clasificación de una sola oración, en nuestro trabajo proponemos abordar la tarea como un problema de clasificación de pares de oraciones, al considerar un generador de oraciones a partir de los tweets de entrada para agregar información adicional contextual. Planteamos dicho enfoque debido al surgimiento de trabajos recientes que utilizan la segunda entrada de BERT como mecanismo para guiar al modelo en la clasificación (Sun et al., 2019; Yu et al., 2019; Ma et al., 2020; Sánchez-Vega and López-Monroy, 2021). Para ello proponemos tres enfoques para generar oraciones auxiliares a partir de los tweets de entrada, el primero consiste en utilizar una pregunta simple relacionada con el tema de interés, el segundo enfoque se basa en utilizar el tweet con mayor similitud textual dentro del conjunto de entrenamiento y el tercero en utilizar las palabras con mayor relevancia dentro de la clase positiva.

Como segunda etapa, dado que no se cuenta con un conjunto de datos para la identificación y la normalización de las reacciones adversas nos enfocaremos en utilizar un método basado en recuperación de información para la normalización de las menciones de reacciones adversas a una terminología médica mediante el uso de un diccionario[10]. La normalización es un proceso que consiste en relacionar las reacciones adversas dentro de un tweet (mencionada de manera coloquial) a una terminología médica utilizada por el personal de la salud. Para ejemplificar dicha normalización, tomemos en cuenta el siguiente tweet: "*between the fucking redbull and vyvanse i popped to energize my triple yesterday...* ***couldn't fall asleep*** *for the life of me.*" donde el tramo del texto que indica dicha reacción adversa se encuentra marcada en negritas y al término médico al que queremos normalizar de acuerdo al diccionario médico es "*Initial insomnia*". Para resolver este problema proponemos un enfoque basado en un Sistema de Recuperación de Información (SR). Para nuestro sistema de recuperación, nos enfocaremos principalmente en utilizar diferentes representaciones de los datos y en la expansión de los documentos (términos médi-

[9]Este modelo se describe a detalle en la subsección 2.3.4

[10]https://www.meddra.org/

cos) con la finalidad de enriquecer nuestro vocabulario para una mejor asociación de los tweets y los términos médicos.

El utilizar un enfoque de recuperación de información para la identificación de RA nos permite abordar la problemática de una manera diferente al estado del arte, reducir el error en cascada del que sufren los mismos y explorar diferentes enfoques utilizando un sistema de recuperación.

1.1 Planteamiento del problema

Las reacciones adversas a medicamentos representan un importante problema de salud pública, ya que son causantes de morbilidad y suponen una carga financiera para los sistemas de salud. Dichas reacciones son causadas por ensayos clínicos limitados que no logran detectar todas las reacciones adversas debido a factores como la cantidad de muestras, la duración de las pruebas y la falta de diversidad entre los pacientes de estudio (Sultana et al., 2013). Bajo este escenario, la vigilancia después de la comercialización, también conocida como fármaco-vigilancia, juega un papel fundamental en la identificación y prevención de riesgos asociados al uso de los medicamentos (Coloma et al., 2013).

Para resolver esta problemática se han propuesto diferentes soluciones que utilizan como fuente de información las redes sociales y foros médicos donde las personas comparten sus experiencias utilizando diferentes medicamentos. Sin embargo, los trabajos que tratan de solucionar dicha problemática se han enfocado mayormente a una clasificación de textos debido a la poca cantidad de datos etiquetados con los que se cuentan actualmente, dejando de lado la normalización de las menciones de reacciones adversas a un vocabulario médico. Por otra parte, los trabajos que se han enfocado en la normalización han obtenido bajos resultados debido a que no se cuenta con un conjunto de entrenamiento y al enfoque de tres etapas que utilizan.

Éstas etapas se basan en una clasificación automática de tweets mediante el uso de algoritmos de Aprendizaje Computacional y Aprendizaje Profundo, seguido de un enfoque basado en Reconocimiento de Entidades Nombradas (NER) y Campos Aleatorios Condicionales (CRF) para la identificación del tramo del texto que menciona dicha reacción adversa y la normalización del tramo del texto a vocabulario médico con ayuda de *embeddings* de palabras y medidas de similitud. Para la etapa de clasificación se ha desarrollado un conjunto de datos de Twitter donde solo el 7% corresponden a una reacción adversa. Por otra parte, para la etapa de identificación y normalización no se cuenta con un conjunto de datos de entrenamiento, lo que ha provocado que los modelos basados en redes neuronales obtengan un bajo rendimiento.

Debido a lo antes mencionado, para resolver esta problemática en nuestro trabajo proponemos un método que consta de dos etapas. Para la primera etapa de clasificación abordaremos el problema mediante un modelo de *Transformer* preentrenado que no necesitan grandes cantidades de datos para ajustarse a una tarea objetivo. Además, utilizaremos un enfoque de pares de oraciones al considerar algunas oraciones auxiliares derivadas de los datos de entrada como información contextual adicional. Como segunda etapa, nos enfocaremos directamente en la normalización de las menciones de reacciones adversas a un vocabulario médico, mediante un sistema de recuperación de información que a diferencia de los métodos propuestos no necesita un conjunto de entrenamiento. Además de esto, al utilizar un enfoque de dos etapas el error en cadena que se produce con el enfoque utilizado en el estado del arte es reducido.

1.2 Objetivos

1.2.1 Objetivo general

Diseñar e implementar un método para la clasificación y normalización de reacciones adversas a medicamentos en Twitter, utilizando técnicas de recuperación de información, que nos permita obtener resultados que igualen o superen el estado del arte.

1.2.2 Objetivos específicos

Los objetivos específicos se listan a continuación:

- Desarrollar e implementar un método que sea capaz de clasificar tweets que mencionen reacciones adversas a medicamentos.
- Desarrollar y evaluar diferentes tipos de sistemas de recuperación que nos ayuden en la normalización de las menciones de reacciones adversas a medicamentos.
- Evaluar y analizar diferentes representaciones del sistema de recuperación para la normalización de reacciones adversas a medicamentos.

1.3 Contribución

En este trabajo se plantean dos contribuciones:

- Propuesta de un método para la normalización de reacciones adversas a medicamentos basado en un sistema de recuperación de información en Twitter.
- Propuesta de un módulo de generación de oraciones auxiliares para la arquitectura de Transformers para una clasificación de pares de oraciones.

1.4 Alcance y limitaciones

El presente trabajo de tesis abarca el diseño, implementación y evaluación del método propuesto para la clasificación y normalización de menciones de reacciones adversas a medicamentos en Twitter en el idioma Inglés. Los conjuntos de datos y las métricas de evaluación utilizada se ajustaron de acuerdo al foro *Social Media Mining for Health* del año 2021.

1.5 Organización de tesis

Los capítulos se estructuran de la siguiente manera:

- Capítulo 2: Marco teórico. En esta sección se presentan las definiciones y conceptos clave para comprender nuestra propuesta.
- Capítulo 3: Trabajo relacionado. En esta sección se hace una revisión del estado del arte. Este capítulo tiene como objetivo mostrar los enfoques y técnicas empleadas para resolver esta problemática.

- Capítulo 4: Método propuesto. En esta sección se describe el enfoque propuesto, comenzando con una descripción general, seguido de la etapa de clasificación y la normalización de las menciones de reacciones adversas a medicamentos. El objetivo es explicar a detalle cada fase de nuestro método, el funcionamiento y los diferentes enfoques propuestos.

- Capítulo 5: Experimentos. En esta sección se detallan los conjuntos de datos utilizados, el pre-procesamiento seguido en cada etapa, las configuraciones de los diferentes modelos y arquitecturas, los resultados obtenidos con cada configuración y el análisis de resultado. El objetivo es brindar la información necesaria que facilite la reproducibilidad de los resultados, realizar una comparación de nuestros resultados con los del estado del arte para analizar y discutir las fortalezas y debilidades de nuestro método.

- Capítulo 6: Conclusiones y Trabajo futuro. En esta sección se describen las contribuciones del enfoque propuesto, así como las conclusiones y el trabajo futuro.

CAPÍTULO 2

MARCO TEÓRICO

En este capítulo se presentan los conceptos y definiciones clave para comprender el enfoque propuesto. En la primera sección se definen los diferentes tipos de representaciones utilizadas para el procesamiento de texto. En la segunda sección se define el concepto de clasificación de texto. En la tercera sección se describe el concepto de Aprendizaje Profundo y se explican las diferentes arquitecturas utilizadas en este trabajo para la clasificación de texto, así como las métricas de evaluación. En la cuarta sección se explica la idea básica del Sistema de Recuperación, así como las métricas utilizadas para evaluar dicho sistema. En la quinta sección se explica el concepto de Normalización de reacciones adversas y se describe el contenido del diccionario utilizado para este procedimiento, así como sus alcances y limitaciones.

2.1 Representación de textos

El modelado del lenguaje natural es la manera en que la máquina logra distinguir e identificar la forma en que nos comunicamos. Por lo que es importante construir una entrada que la máquina logre interpretar, ya sea texto, audio e incluso imágenes. La complejidad del análisis del modelo estará directamente relacionada con el tipo de representación utilizada, agregando costo de tiempo, memoria, etc. Por lo que elegir

una representación adecuada es de suma importancia para cada tarea. Algunas de las representaciones que se pueden utilizar son las siguientes.

2.1.1 Bolsa de palabras

Una de las técnicas más utilizadas en el Procesamiento del Lenguaje Natural (PLN) y en la Recuperación de Información (IR) para representar texto es la Bolsa de palabras (BoW por sus siglas en inglés). Esta es una técnica donde cada documento de un conjunto $D = \{d_1, d_2, \ldots, d_i\}$ es representado por un conjunto de palabras/términos $V = \{t_1, t_2, \ldots, t_j\}$ donde el orden de aparición de cada palabra no importa. De igual manera cada documento d_i es representado por un vector v_i cuya dimensión será igual a $|V|$ y cada elemento del vector contará con un peso $w_{i,j}$. La Tabla 2.1 muestra un ejemplo de una Bolsa de palabras, en donde se tiene un conjunto de documentos $D = \{$"*Humira made me lose my appetite*", "*trazodone made me feel sick*"$\}$ del cual se genera un vocabulario $V = \{$ "*humira*", "*trazadone*", "*made*", "*me*", "*feel*", "*lose*", "*my*", "*sick*", "*apetite*"$\}$.

Tabla 2.1: Ejemplo de una representación de bolsa de palabras.

	humira	*trazadone*	*made*	*me*	*feel*	*lose*	*my*	*sick*	*apetite*
d_1	$w_{1,1}$	$w_{1,2}$	$w_{1,3}$	$w_{1,4}$	$w_{1,5}$	$w_{1,6}$	$w_{1,7}$	$w_{1,8}$	$w_{1,8}$
d_2	$w_{2,1}$	$w_{2,2}$	$w_{2,3}$	$w_{2,4}$	$w_{2,5}$	$w_{2,6}$	$w_{2,7}$	$w_{2,8}$	$w_{2,8}$

Donde el peso $w_{i,j}$ es asignado a cada uno de los términos dentro del documento. Este peso a su vez puede ser calculado de diferentes maneras y puede incluir diferentes tipos de información a cerca de los documentos. Los esquemas de pesado más utilizados son siguientes:

- **Binario:** el peso $w_{i,j}$ toma el valor de 1 si el término t_j aparece en el documento d_i y 0 en el caso de no aparecer.

$$w_{i,j} = \begin{cases} 1 \text{ si el termino } t_j \text{ aparece en el documento } i \\ 0 \text{ en caso contrario} \end{cases} \tag{2.1}$$

- **Frecuencia de Término (TF):** En este enfoque el peso es el número de repeticiones (o frecuencia) de cada término t_j en el documento i. Así un término será más importante entre más aparezca en el documento.

$$w_{i,j} = f_{i,j} \tag{2.2}$$

- **Frecuencia de Términos - Frecuencia Inversa de Documentos (TF-IDF):** En los enfoques anteriores no se considera la frecuencia del término a través de todos los documentos en la colección. Por lo que en este enfoque denotado comúnmente como *tf-idf* se determina el peso de un término j en el documento i en proporción directa al número de veces que el término aparece en el documento, e inversamente proporcional al número de documentos en el conjunto D_t que contienen el término j. En particular, el peso está dado por la ecuación siguiente,

$$w_{i,j} = tf_{i,j} \times idf_i \tag{2.3}$$

donde $tf_{i,j}$ es la frecuencia del término t_j en el documento d_i e idf_i es la frecuencia inversa del documento definida como $\log\left(\frac{N}{dfi}\right)$; $N = |D_t|$ y df_i es el número de documentos que contienen el término t_j. El factor idf_i es utilizado para eliminar el impacto de términos frecuentes que existen en la mayoría de los documentos.

2.1.2 *N*-gramas de palabras

Un n-grama de palabras es un conjunto de n elementos consecutivos en un documento de texto, que puede incluir palabras, números, símbolos y puntuación. El uso de n-gramas de palabras como representación de texto es utilizada en clasificación de textos y sistemas de recuperación de información. Esta representación son subsecuencias consecutivas que dependen de un valor N. Para valores de uno hasta tres los n-gramas recibirán el nombre de unigrama, bigrama y trigrama, mientras que para valores mayores se utilizará el número seguido de grama (4-grama). En este tipo de representaciones el valor de N es de gran importancia, ya que mientras mayor sea el número, la representación será más específica. La manera en que un documento puede ser representado por los n-gramas de palabras es similar al explicado en la representación de Bolsa de Palabras. Tomando como ejemplo el siguiente texto "*Humira made me lose my appetite*", podemos extraer los siguientes N-gramas de palabras:

- unigrama = {"*humira*", "*made*", "*me*", "*lose*", "*my*", "*appetite*"}
- bigrama = {"*humira made*", "*made me*", "*me lose*", "*lose my*", "*my appetite*"}
- trigrama = {"*humira made me*", "*made me lose*", "*me lose my*", "*lose my appetite*"}

2.1.3 Embeddings de palabras

Hoy en día, los *embeddings* de palabras son una de las técnicas más utilizadas como representación de texto, la clasificación de texto y en recuperación de información. Los *embeddings* son vectores de palabras que permiten que las palabras con un significado similar tengan una representación similar dentro de un espacio n-dimensional. Si bien esta técnica se remonta a los años 90, los *embeddings* no se popularizaron

sino hasta 2013, cuando un grupo de científicos de la computación desarrollaron un método simple para la creación de *embeddings* de palabras llamado "word2vec". Los *embeddings* son técnicas de modelado del lenguaje que transforman el vocabulario de un corpus de entrada en una representación vectorial continua y de baja dimensión, utilizando una Red Neuronal Artificial no recurrente para generar los vectores de representación (Wohlgenannt et al., 2016). Los *embeddings* han demostrado un rendimiento de vanguardia para estimaciones de similitud de palabras, pero también para operaciones más sofisticadas como analogías de palabras y sirven como componentes de entrada para varias tareas de PLN (Mikolov et al., 2013; Ghannay et al., 2016). A continuación, se describen los métodos utilizados en este trabajo para generar *embeddings* de palabras.

Word2vec

Word2Vec es una herramienta creada por Mikolov et al. (2013) que aplica una red neuronal para obtener el contexto lingüístico de las palabras o frases. La entrada de la red está determinada por un corpus de gran dimensión donde la salida de la red son los vectores de los embeddings que representan a las palabras del corpus, donde por lo general la dimensionalidad de cada vector varía entre 50 y 1000. Estos vectores generados a su vez pueden ser visualizados dentro de un espacio vectorial y la cercanía entre ellos puede traducirse como la similitud que tienen entre ellos.

Existen dos arquitecturas diferentes para que el algoritmo de Word2vec pueda crear dichos vectores, una de ellas es Bolsa de Palabras Continuas (CBOW) y la otra es Skip-gram continuo. El modelo Skip-gram aprende a predecir una palabra objetivo gracias a una palabra cercana. Por otro lado, el modelo CBOW predice la palabra objetivo según su contexto. El contexto se representa como una bolsa de palabras contenidas en una ventana de tamaño fijo alrededor de la palabra objetivo. La Figura 2.1 muestra estas dos arquitecturas donde podemos ver que la matriz de

peso entre la entrada y la capa de proyección es compartida en todas las posiciones de las palabras de la misma manera que en un Modelo de Lenguaje de Red Neuronal, lo que da como resultado un vector de longitud N de cada palabra (Wohlgenannt et al., 2016).

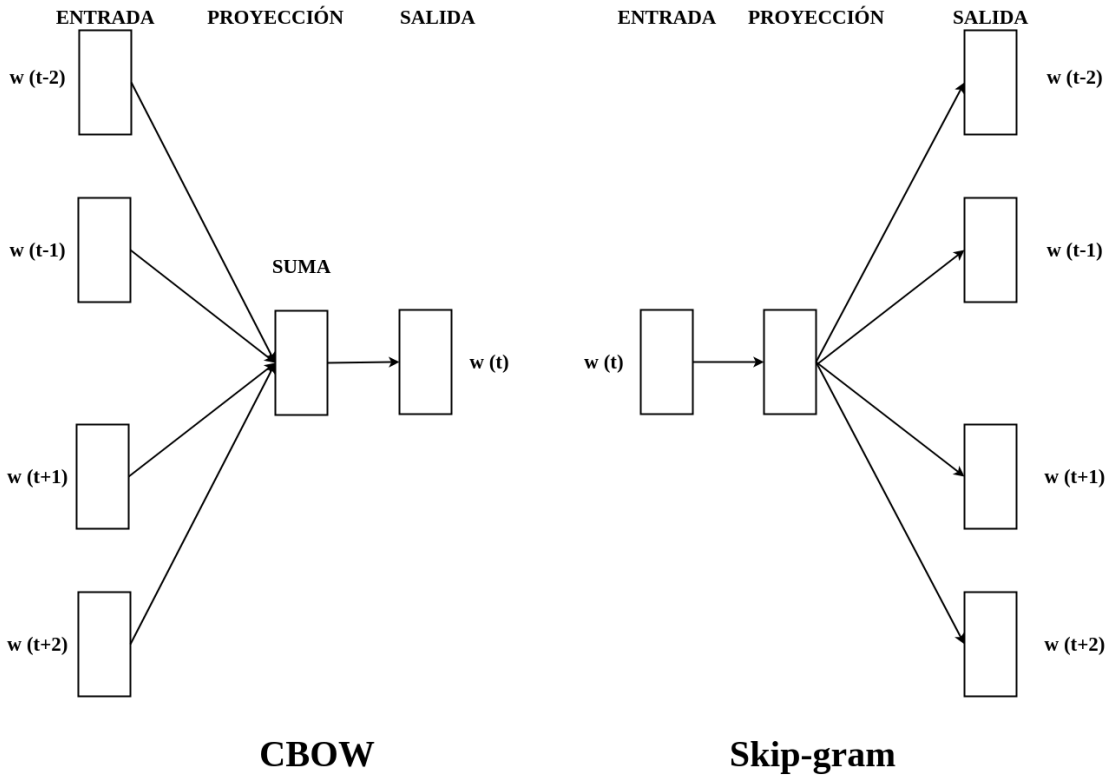

Figura 2.1: Arquitectura CBOW y Skip-gram contiguo. La arquitectura CBOW predice la palabra actual en función del contexto y Skip-gram predice las palabras circundantes dada la palabra actual. Figura basada en el trabajo de Wohlgenannt et al. (2016).

GloVe: Vectores Globales para la representación de las palabras

GloVe es un algoritmo de aprendizaje no supervisado para obtener representaciones vectoriales de palabras. El modelo GloVe fue introducido por Pennington et al. (2014) y está compuesto por dos enfoques: el primero es un método de factorización matricial y el segundo se basa en ventanas poco profundas. El entrenamiento se

realiza utilizando estadísticas globales agregadas de la co-ocurrencia entre palabras dentro del corpus, y las representaciones resultantes muestran sub-estructuras lineales interesantes del espacio vectorial de palabras. La idea principal de GloVe es aprender representaciones de palabras con el objetivo de predecir palabras en un contexto local. Los modelos pre-entrenados más relevantes basados en este modelo han sido construidos con corpus de gran dimensionalidad provenientes de Twitter, Common Crawl y Wikipedia, y a diferencia de los modelos pre-entrenados de Word2Vec, estos modelos presentan una menor dimensionalidad en sus vectores de palabras, rondando entre 50 y 300. El uso de este modelo por lo general suele destacar en tarea de clasificación de textos (Wang et al., 2019).

2.1.4 BERT Emebeddings

Desde la introducción de los embeddings de palabras previamente entrenadas como Word2vec o GloVe. Se han dedicado muchos esfuerzos para desarrollar *embeddings* de uso general. Uno de los modelos más recientes que aprovecha en gran medida el modelo de lenguaje y que ha logrado un rendimiento de vanguardia en tareas de comprensión del lenguaje natural como la clasificación de secuencias, pares de secuencias y respuesta a preguntas es el modelo de *Transformers* BERT (Devlin et al., 2018). La ventaja de estos modelos pre-entrenados es su facilidad para ajustarse a una tarea específica con solo una capa de salida adicional lo que lo convierte en un modelo de última generación que sugiere que las representaciones de BERT pueden ser de uso general. Los *Transformers* tienen muchos beneficios sobre los modelos secuenciales convencionales como las redes LSTM, RNN, GRU, etc. Una de las ventajas de estos modelos es que modela el texto de manera más efectiva, ya que toma en cuenta la relevancia de una palabra con respecto a las palabras anteriores y siguientes. Para ello BERT parte de una arquitectura de codificador-decodificador que utiliza mecanismos de atención para transmitir al decodificador una imagen más completa de toda

la secuencia a la vez en lugar de secuencialmente. Sin embargo, para la capa de entrada BERT utiliza un vector de tokens de secuencia junto con los tokens especiales (Una explicación más detallada de este modelo se encuentra en la sección 2.3.4). Para obtener dicha representación, BERT utiliza *WordPiece* (Wu et al., 2016) que divide las palabras en tokens (sub-palabras) donde las palabras fuera del vocabulario son divididas en sub-palabras como por ejemplo la palabra "playing" que puede ser dividida en "play" y "ing" lo que le permite al modelo abarcar un mayor número de palabras fuera del vocabulario. Para generar esta nueva representación BERT utiliza diferentes tipos de representaciones así como doce capas de codificadores donde cada capa genera una representación diferente que al final puede ser agregada a un solo vector de representación. Esto permite obtener representaciones en diferentes niveles, sin embargo, los autores recomiendan sumar estas representaciones ya que consideran que agrega una mayor información a la representación final.

2.2 Clasificación de textos

Con el crecimiento desmesurado de la información generada a nivel mundial, la búsqueda de información precisa y específica se ha vuelto más rigurosa. La clasificación de textos es una de las aplicaciones del Aprendizaje Automático y consiste en clasificar textos en función de su contenido, es decir, realizar un análisis de las palabras para decidir qué tipo de texto es el que se está identificando. La clasificación de textos surge de la necesidad de separar documentos de un tema o clasificación específica de un conjunto de documentos de diferentes temas. Al lograr clasificar los documentos por temas, la búsqueda de información se puede realizar de manera más sencilla. La clasificación de textos es la tarea de asignar un valor booleano a cada par $d_j, c_i \in D \times C$, donde D es un dominio de documentos y $C = \{c_1, \ldots, c_{|C|}\}$ es un conjunto de categorías predefinidas. Un valor de T asignado a $\langle d_j, c_i \rangle$, indica una decisión de clasificar d_j bajo c_i , mientras que un valor de F indica una decisión

de no presentar d_j bajo c_i (Sebastiani, 2002).

2.3 Aprendizaje profundo

El término aprendizaje profundo fue introducido por primera vez a la comunidad de aprendizaje computacional por Dechter (1986), años más tarde fue introducido a las redes neuronales por Aizenberg et al. (2000). El aprendizaje profundo, también conocido como redes neuronales profundas, es una técnica de aprendizaje automático capaz de extraer atributos de bajo nivel y asociarlos a conceptos de alto nivel utilizando arquitecturas jerárquicas. Los algoritmos de aprendizaje profundo extraen características de los datos con las que el ordenador aprende y reconoce patrones. Las arquitecturas más populares son la Redes Neuronales Convolucionales (CNN), la Redes Neuronales Recurrentes (RNN) y las Redes de Memoria a corto y largo plazo (LSTM). A continuación, en este capítulo se describen las Arquitecturas utilizadas en nuestro trabajo para la clasificación de textos.

2.3.1 Arquitectura basada en CNN

Una de las arquitecturas de redes neuronales más utilizadas para la clasificación de textos son las redes convolucionales. Los fundamentos de estas redes se basan en el Neocognitron introducido por Fukushima and Miyake (1982). Este modelo fue más tarde mejorado por LeCun et al. (1998) al introducir un método de aprendizaje basado en la propagación hacia atrás para poder entrenar el sistema correctamente. Las Redes Neuronales Convolucionales utilizan capas con filtros de convolución que son aplicadas a la extracción de características, estas redes originalmente fueron creadas para tareas de visión por computadora, sin embargo, estas redes han tenido una variedad de aplicaciones en procesamiento de lenguaje natural, alcanzando resultados de vanguardia en tareas como: el Etiquetado de las Partes de la Oración

(PoS), el Reconocimiento de Entidades Nombradas (NER), entre otras tareas tradicionales del PLN (Otter et al., 2020). Uno de los primeros trabajos que se enfocaron en la clasificación de texto con redes convolucionales fue realizado por Kim (2014).

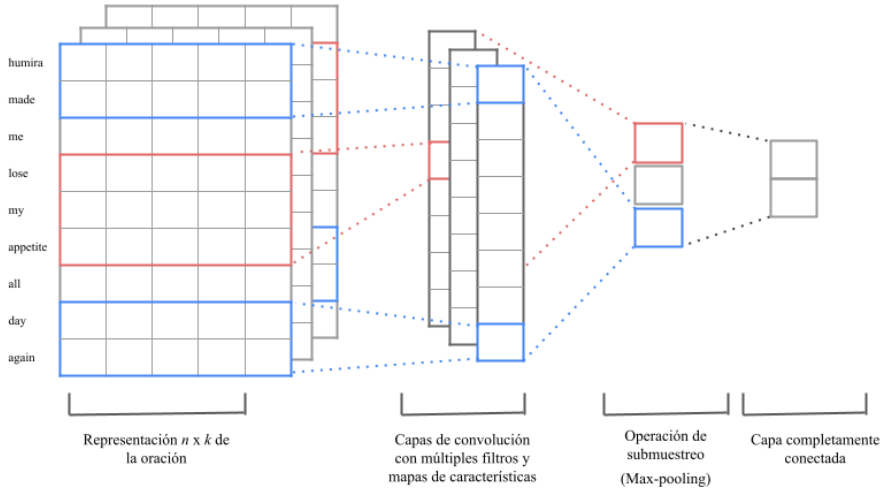

Figura 2.2: Arquitectura de Red Neuronal Convolucional para la clasificación de texto. Figura basada en el trabajo de Kim (2014).

En este trabajo se presenta una arquitectura basada en una representación de n-gramas de palabra para generar una representación en diferentes niveles. La Figura 2.2 muestra la arquitectura CNN utilizada donde podemos ver que el modelo cuenta con tres capas. La primera es la capa de *embedding* que convierte las palabras en sus respectivos vectores. La segunda es la capa de convolución donde tiene lugar el procesamiento principal del modelo donde los filtros aleatorios pasan por encima de la matriz de oraciones y la reducen a una matriz de baja dimensión. La tercera capa es la capa *max-pooling*, esta capa es una operación de agrupación que calcula el valor máximo para los filtros del mapa de características y lo usa para crear un mapa de características con muestreo reducido (agrupado). Finalmente, con una capa completamente conectada y una función *softmax* el modelo es capaz de obtener las probabilidades de pertenencia de cada clase para finalmente asignarle una de ellas.

2.3.2 Redes neuronales recurrentes

El fundamento de las redes neuronales recurrentes, se basa en la inclusión de memoria introducida por retrasos de tiempo en la estructura sináptica de la red (Williams and Zipser, 1989). Una Red Neuronal Recurrente no tiene una estructura de capas definida, sino que permite conexiones arbitrarias entre las neuronas, pudiendo crear ciclos, con esto se consigue crear la temporalidad, permitiendo que la red tenga memoria. Las RNN integran bucles de realimentación, los cuales permiten que la información persista durante algunos pasos o épocas de entrenamiento, a través de conexiones desde las salidas de las capas, que "incrustan" sus resultados en los datos de entrada, la Figura 2.3 ilustra un ejemplo de los bucles de realimentación.

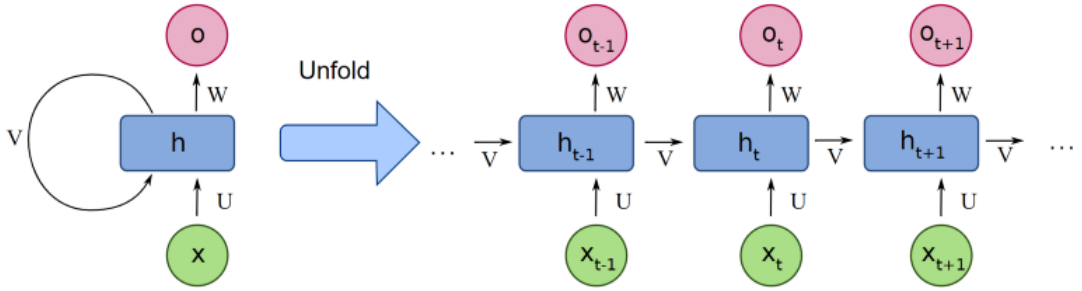

Figura 2.3: Ilustración del despliegue temporal en una Red Neuronal Recurrente. Figura obtenida de Pedro Borges (2018)

.

donde x corresponde a la entrada, O es la salida, h es el bloque principal de la RNN que contiene los pesos y las funciones de activación de la red y V representa la comunicación de un paso de tiempo a otro. Las conexiones entre nodos forman grafos dirigidos a lo largo de una secuencia temporal. Esta red funciona como una red con múltiples copias de sí misma, cada una con un mensaje a su sucesor. Estas redes son útiles para el análisis de secuencias, como puede ser el análisis de textos, sonido o vídeo. Hoy en día existe una gran variedad de tipos de redes neuronales recurrentes, dependiendo del número de capas ocultas y la forma de realizar la retro-propagación.

Red LSTM

La red LSTM (*Long Short Term Memory*) es un tipo especial de red recurrente, diseñada específicamente para evitar el problema del desvanecimiento del gradiente (Hochreiter et al., 2001), originalmente fue propuesta por Hochreiter and Schmidhuber (1997). Las redes LSTM se componen de una secuencia de unidades o celdas encadenadas, la estructura de las unidades se muestra en la Figura 2.4 donde por cada instante de tiempo t, un conjunto de vectores es procesado para generar un nuevo estado oculto:

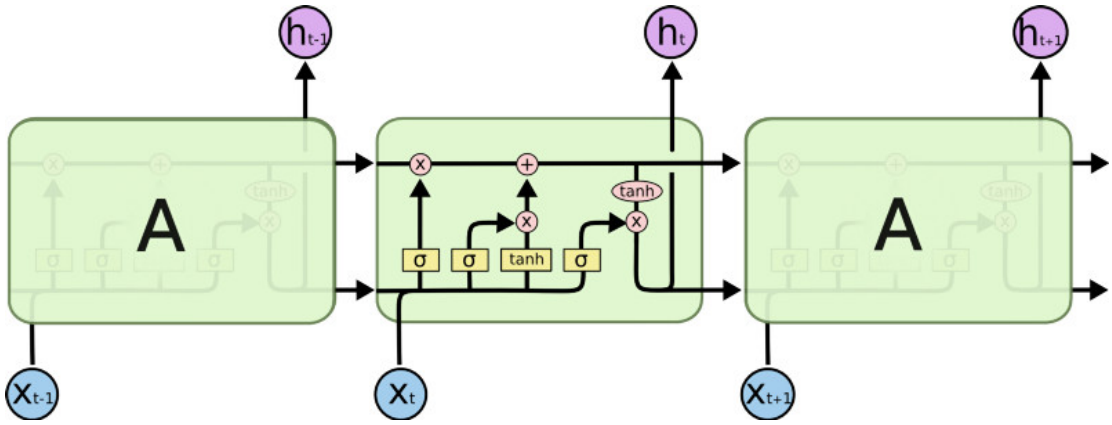

Figura 2.4: Ilustración de la unidad LSTM. Figura obtenida de Christopher Olah (2015)

.

1.- Compuerta para descartar:

$$f_t = \sigma(W_f \times x_t + U_f \times h_{t-1} + b_f) \tag{2.4}$$

2.- Compuerta de entrada:

$$i_t = \sigma(W_i \times x_t + U_i \times h_{t-1} + b_i) \tag{2.5}$$

3.- Vector de nuevos candidatos para el estado de la unidad:

$$\hat{C}_t = \tanh(W_c \times x_t + U_c \times h_{t-1} + b_C) \tag{2.6}$$

4.- Compuerta de salida:

$$o_t = \sigma(W_o \times x_t + U_o \times h_{t-1} + b_o) \tag{2.7}$$

5.- Memoria de la unidad:

$$C_t = i_t \times \hat{C}_t + f_t \times C_{t-1} \tag{2.8}$$

6.- Capa de salida oculta:

$$h_t = o_t \times \tanh(C_t) \tag{2.9}$$

Donde W_f, U_f, W_i, U_i son matrices de pesos y b_f, b_i, b_C, b_o son vectores de sesgos.

El proceso interno se describe de acuerdo a:

- Paso 1: La unidad decide qué información va a ser descartada. Esto se hace computando f_t. Una vez que ocurre esto, un número entre 0 y 1 se crea para cada número en el estado de la celda C_{t-1} , donde un 1 representa mantener completamente y un 0 representa olvidar completamente.

- Paso 2: La celda decide qué información se va a almacenar, usando $\hat{C}_t$ y i_t.
- Paso 3: La unidad actualiza el estado anterior C_{t-1} con el nuevo C_t. Esto se hace computando la fórmula C_t.
- Paso 4: La salida h_t es determinada en función de una versión filtrada del estado de la celda.

Red GRU

La red GRU (Gated Recurrent Unit) es un tipo especial de red recurrente, propuestas por Chung et al. (2014), con la finalidad de abordar el problema del desvanecimiento del gradiente (Kolen and Kremer, 2001), esta red es una versión simplificada de la red LSTM. Las red GRU se compone de una secuencia de unidades o celdas encadenadas, la estructura de las unidades se muestra en la Figura 2.5.

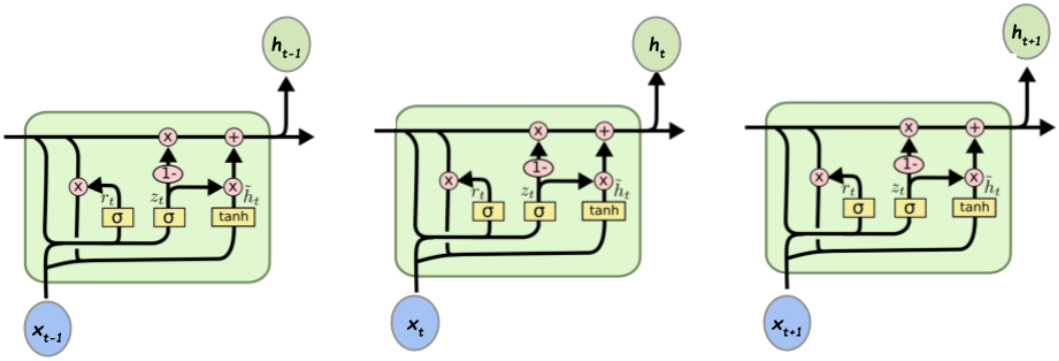

Figura 2.5: Ilustración de la unidad GRU. Figura inspirada en Perambai, Abhishek. (2019)

.

Con la finalidad de resolver el problema de desvanecimiento del gradiente, la red GRU utiliza las compuertas de actualización y de reinicio, las cuales deciden que información pasará a la salida, así como la cantidad de información que pasará. Lo que la hace diferente de la red LSTM, es la mayor cantidad de tiempo que la información puede permanecer en las celdas. El primer paso que realiza la red GRU

es la activación h_t de la GRU al tiempo t, ésta es una interpolación lineal entre la activación previa h_{t-1} y el candidato de activación $\hat{h}_t$.

$$h_t = (1 - z_t)h_{t-1} + z_t\hat{h}_t \tag{2.10}$$

Donde la compuerta de actualización está definida por: z_t, esta compuerta decide cuánto actualiza en la unidad de activación. La actualización de la compuerta es calculada por:

$$z_t = \sigma(W_z x_t + U_z h_{t-1}) \tag{2.11}$$

El procedimiento de tomar una suma lineal entre los estados existentes y el nuevo estado calculado es similar al de la unidad LSTM. A diferencia de la red LSTM, la red GRU no tiene ningún mecanismo para controlar el grado al cual es expuesto el estado, pero expone todo el estado a la vez. El candidato de activación $\hat{h}_t$ es calculado de manera similar a la unidad recurrente tradicional:

$$\hat{h}_t = \tanh(W x_t + U(r_t \odot h_{t-1})) \tag{2.12}$$

Donde r_t es un conjunto de compuertas de reinicio y $\odot$ es una multiplicación de elementos. Cuando está apagado (r_t es cercano a 0) la compuerta de reinicio hace que la unidad actué como si estuviera leyendo el primer símbolo de la secuencia de entrada, permitiéndole que olvide el cómputo del estado anterior. La compuerta de reinicio r_t es calculada de manera similar a la compuerta de actualización:

$$\hat{r}_t = \sigma(W_r x_t + U_r h_{t-1}) \tag{2.13}$$

2.3.3 Atención contextual

La atención es un concepto conocido en la psicología humana, donde los humanos limitados por el manejo de grandes cantidades de información tienen un enfoque selectivo en las partes más relevantes de la misma. Al mapear el mismo concepto de la psicología humana al aprendizaje profundo, se crea el uso de un mecanismo de atención el cual permite tener un enfoque en ciertas partes de secuencias o regiones, desdibujando las partes irrelevantes en secuencias de datos como: secuencias de texto o secuencias de audio.

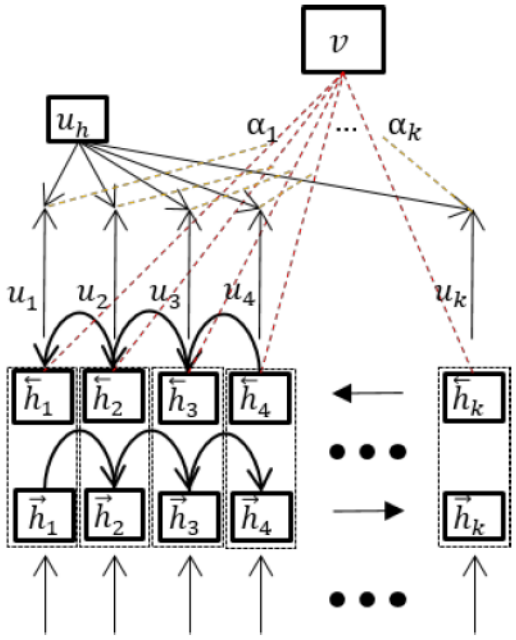

Figura 2.6: Ilustración del mecanismo de atención. Figura inspirada por de Yang et al. (2016)

.

La atención contextual fue introducida por Yang et al. (2016), con la propuesta de una red de atención jerárquica para clasificación de documentos, la red maneja la atención en dos diferentes niveles, el primer nivel extrae la importancia de las palabras con respecto a las oraciones y el segundo nivel extrae la importancia de las oraciones con respecto al documento. Previo a la extracción de la importancia de los

elementos de una secuencia a través de la atención contextual, es necesario aplicar una codificación en la secuencia, con la finalidad de extraer el contexto de cada uno de sus elementos; para ello se utiliza una red Bidireccional-GRU como se puede ver en la Figura 2.6, donde la salida h_i de cada palabra x_i se alimenta a una capa de un perceptrón multicapa, con la finalidad de obtener una representación oculta u_i de la palabra, esto se puede ver en la siguiente ecuación.

$$u_i = \tanh(W_h h_i + b_h) \tag{2.14}$$

Una vez que u_i es calculado, se calcula la importancia de la palabra, con la similitud entre u_i y u_c a través de un producto punto, posteriormente se obtiene un peso de importancia normalizado α_i a través de la función de *softmax*, la siguiente ecuación presenta el cálculo de la importancia de la i-ésima palabra.

$$\alpha_i = \frac{exp(u_i^T u_h)}{\sum exp(u_j^T u_h)} \tag{2.15}$$

El vector u_c es el vector de contexto a nivel de palabra, el cual es inicializado aleatoriamente y sus parámetros se ajustan conforme se realiza el entrenamiento de la red, el vector de contexto se puede ver como una medida de importancia global para las palabras en el texto. Por último se calcula la representación general del mensaje v, como una suma pesada de los elementos codificados por la red Bidireccional-GRU y sus respectivos pesos de importancia normalizados, la siguiente ecuación muestra el cálculo de esta representación.

$$u = \sum_j \alpha_j h_j \tag{2.16}$$

2.3.4 Arquitectura basada en Transformer

Como sabemos uno de los mayores desafíos en el Procesamiento de Lenguaje Natural es la escasez de los datos para entrenar un modelo. Esto se debe a que el PLN es un campo diversificado con una gran cantidad de tareas distintas donde la mayoría de los conjuntos de datos específicos contienen solo unos pocos cientos o miles de ejemplos de entrenamiento etiquetados por humanos. Sin embargo, los modelos de PLN modernos basados en aprendizaje profundo se ven beneficiados de las cantidades de datos de entrenamiento, mejorando cuando se entrenan con millones o miles de millones de ejemplos anotados. Para ayudar a cerrar esta brecha en los datos, los investigadores se han enfocado en desarrollar modelos de propósito general, utilizando enormes cantidades de texto sin etiquetar con la finalidad de obtener representaciones para múltiples tareas. No obstante, fue hasta la llegada de los modelos de Transformers que este tipo de modelos comenzó a obtener mejores resultados.

BERT es un modelo de representación de lenguaje que significa "*Bidirectional Encoder Representations from Transformers*" (Peters et al., 2018; He et al., 2020). Este modelo es un modelo pre-entrenado que utiliza representaciones bidireccionales profundas a partir de texto sin etiquetar y tomando en cuenta el contexto de cada palabra tanto a la izquierda como a la derecha, dando como resultado un modelo que puede ser ajustado con una sola capa de salida adicional a una amplia gama de tareas, como clasificación de textos, respuesta de preguntas, inferencia de lenguaje, etc. sin la necesidad de modificaciones sustanciales en la arquitectura (Devlin et al., 2018). Para la clasificación de textos estos modelos solo necesitan dos componentes adicionales, una red neuronal *feed-forward* y una capa *softmax* como se muestra en la Figura 2.7.

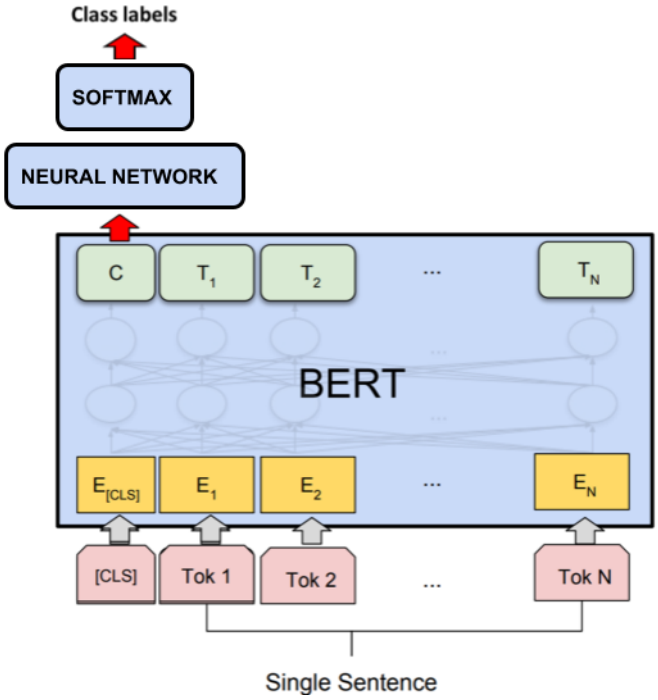

Figura 2.7: Arquitectura BERT para clasificación de texto. Figura basada en el trabajo de Devlin et al. (2018).

Además de esto, una de las ventajas de BERT es que puede representar sin ambigüedad tanto una sola oración como un par de oraciones en secuencias de tokens. Para generar los tokens, este modelo utiliza *embeddings* de WordPiece (Wu et al., 2016) que contiene un vocabulario de más 30,000 tokens. Donde el primer token de cada secuencia es siempre un token de clasificación especial $[CLS]$, seguido de la representación de entrada [Tok 1], [Tok 2], . . . , [Tok N].

Por otra parte debido a la arquitectura con la que fue desarrollado BERT, esto nos permite realizar una clasificación de pares de oraciones ya que BERT puede representar dos oraciones sin ambigüedad, para ello las agrupa en una sola secuencia utilizando un token especial $[SEP]$, seguido de la representación de la segunda entrada [Tok 1], [Tok 2], . . . , [Tok M] tal como se puede ver en la Figura 2.8. Una vez que se cuenta con el vector de entrada, el modelo obtienen tres representaciones diferentes, un *embedding* de token, un *embedding* de posición y un *embedding* de

segmento. Finalmente estas representaciones son sumadas para generar un nuevo vector que contendrá toda esta información, tal como se puede ver en la Figura 2.9.

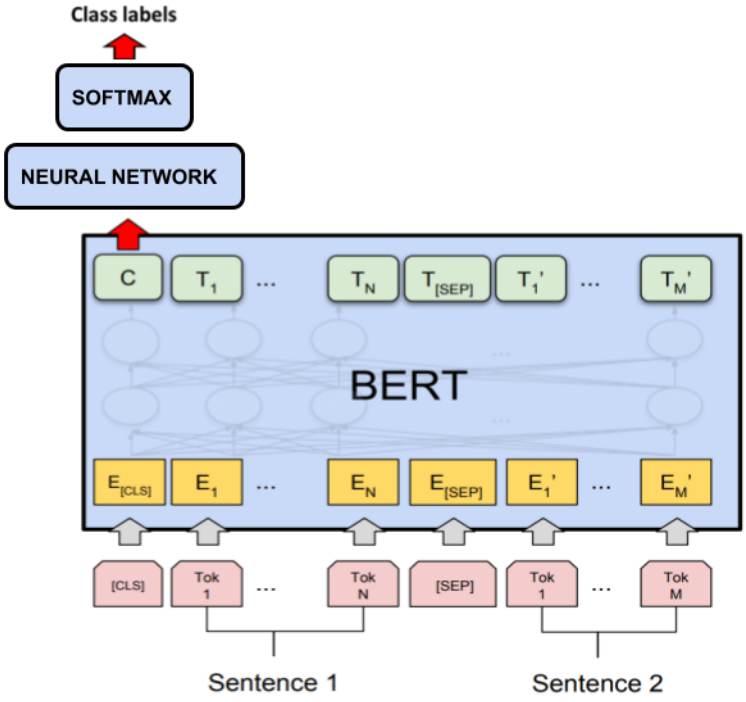

Figura 2.8: Arquitectura de BERT para clasificación de pares de oraciones. Figura basada en el trabajo de Devlin et al. (2018).

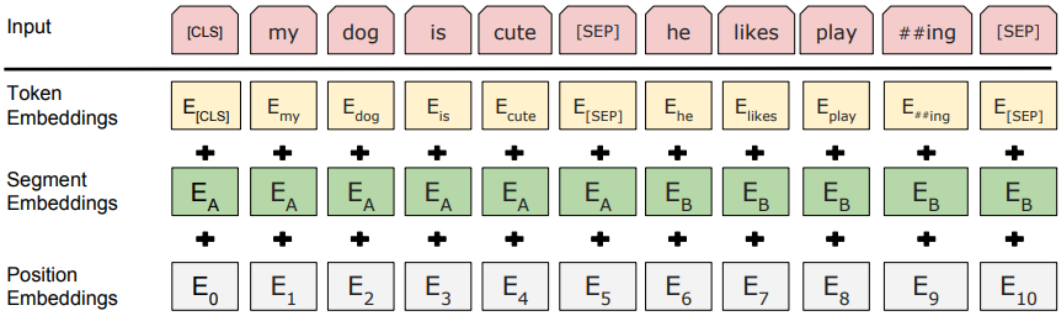

Figura 2.9: Representación de la entrada de BERT. Figura obtenida de Devlin et al. (2018)

2.3.5 Medidas de evaluación

Antes de su uso, todo modelo creado debe ser evaluado para medir su calidad. Esta tarea de evaluación no es trivial, ya que puede depender de varios criterios. En el contexto de un clasificador de textos es necesario medir su capacidad de predicción sobre nuevas instancias, para ello es necesario contar con un conjunto de ejemplos distinto al conjunto con el que se entrenó el clasificador, a esto se le llama "conjunto de evaluación". Para un clasificador de textos las medidas de evaluación comúnmente utilizadas son Precisión (Precision), Recuerdo (Recall) y Puntuación F-1 (F1-Score). Para calcular dichas métricas, se realiza una comparación entre las etiquetas correctas de los documentos y las etiquetas predichas por el modelo, lo que genera una matriz de confusión con cuatro resultados posibles:

- **VP (Verdadero Positivo)**: Son las instancias positivas etiquetadas por el clasificador correctamente.
- **FP (Falso Positivo)**: Son las instancias positivas etiquetadas por el clasificador incorrectamente.
- **VN (Verdadero Negativo)**: Son las instancias negativas etiquetadas por el clasificador correctamente.
- **FN (Falso Negativo)**: Son las instancias negativas etiquetadas por el clasificador incorrectamente.

Estos valores obtenidos de la matriz de confusión servirán para calcular el valor de cada medida de evaluación.

Precisión

La precisión es una métrica que cuantifica el número de predicciones positivas correctas realizadas. La precisión, por lo tanto, calcula la proporción de etiquetas cor-

rectamente etiquetadas por el clasificador, divididas por el número total de etiquetas correctas. Para obtener dicho valor se utiliza la siguiente fórmula.

$$\text{Precision} = \frac{VP}{VP + FP} \tag{2.17}$$

Recuerdo

El recuerdo es la relación entre el número de datos positivos correctamente etiquetados y el número total de muestras positivas. Esta métrica evalúa la capacidad del modelo para detectar los documentos positivos. Para obtener dicho valor se utiliza la siguiente fórmula.

$$Recall = \frac{VP}{VP + FN} \tag{2.18}$$

Puntuación F-1

La medida F1 es la media armónica de la precisión y el recuerdo. Por lo tanto, esta medida tiene en cuenta tanto los FP como los FN y a diferencia de la precisión, ésta medida suele ser más útil si se cuenta con una distribución de datos desequilibrados. Para obtener dicho valor se utiliza la siguiente fórmula.

$$F1 - Score = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}} \tag{2.19}$$

2.4 Recuperación de Información

El significado del término Recuperación de Información puede ser muy amplio. Sin embargo, como campo académico de estudio, la recuperación de información consiste en encontrar material (generalmente documentos) de naturaleza no estructurada (generalmente texto) que satisface una necesidad de información dentro de grandes colecciones de datos. En pocas palabras lo podemos ver como el proceso de acceder y recuperar la información más relevante dada una consulta. Para ello es necesario una representación de la información (comúnmente llamada indexación) que facilite la obtención de estos datos. Dicho de otra manera, a este proceso se le puede denominar como Sistema de Recuperación (SR) de información textual. El objetivo principal del SR es desarrollar un modelo para recuperar información de grandes conjuntos de documentos tomándolo como un problema de recuperación ad-hoc. En la recuperación ad-hoc, el usuario debe ingresar una consulta en lenguaje natural que describa la información requerida. Luego, el SR devolverá como salida los documentos que considere más relevantes con la información proporcionada por el usuario.

La complejidad asociada al lenguaje natural es especialmente clave a la hora de recuperar información textual para satisfacer las necesidades de información de un usuario (Baeza-Yates et al., 1999). Es por esto que en la recuperación de información textual se suelen utilizar técnicas de procesamiento de lenguaje natural tanto para facilitar las descripciones del contenido del documento como para presentar la consulta del usuario. Todo ello con el objetivo de comparar ambas descripciones y presentarle al usuario los documentos que mejor satisfagan sus necesidades de información. En resumen, un sistema de recuperación realiza las siguientes tareas en respuesta a la consulta de un usuario:

1. Indexación de la colección de documentos: en esta fase se aplican técnicas de PLN para generar un índice que contenga las descripciones de los documentos.

Normalmente cada documento se describe a través de un conjunto de términos que, en teoría, representa mejor su contenido. En esta indexación las representaciones más utilizadas son Bolsa de Palabras, *embeddings* de palabras y n-grama de palabras.

2. Cuando un usuario formula una consulta, el sistema lo analiza y, si es necesario, la transforma con la esperanza de representar las necesidades de información del usuario de la misma forma que se representa el contenido del documento.

3. El sistema compara la descripción de cada documento con la de la consulta, y presenta al usuario aquellos documentos cuyas descripciones se acercan más a la descripción de la consulta.

4. Los resultados se enumeran en orden de relevancia, es decir, por el nivel de similitud entre las descripciones del documento y la consulta.

2.4.1 Indexado de documentos

En el contexto de recuperación de información para que los documentos de un conjunto D puedan ser procesados por un sistema de recuperación, estos deben llevarse a una representación adecuada, este procedimiento se conoce como indexado de documentos (Sebastiani, 2005). Como paso previo al indexado de documentos, es necesario pre-procesarlos, eliminando toda la información no útil que depende de la naturaleza de los documentos y puede consistir en etiquetas de meta-texto, comentarios, símbolos y caracteres no alfabéticos. También es necesario sustituir letras mayúsculas y en muchas ocasiones lematizar los términos sustituyéndolos por su lema o raíz eliminando sufijos. Adicionalmente suelen retirarse de los documentos palabras neutrales o palabras que no aportan información sobre su naturaleza como los artículos, preposiciones y conjunciones.

El Indexado consiste en representar cada documento d_i como un vector de características o atributos $[t_{i,1}, t_{i,2}, \ldots, t_{i,|T|}]$ donde cada atributo t representa uno de los términos o palabras que aparecen en la colección de los documentos D. Al conjunto de términos T se le conoce como vocabulario de colección. El peso $w_{i,j}$ asignado a cada uno de los términos en un documento puede determinarse de varias maneras y puede incluir diferentes tipos de información acerca de los documentos. Los esquemas de pesado más utilizados son el Binario, TF y TF-IDF descritos en la sección 2.1.

Cuando el conjunto de documentos ha sido indexado se tiene una matriz cuyos elementos representan el peso de cada término en cada uno de los documentos. Debido a que el tamaño del vocabulario de una colección de documentos suele ser muy grande (miles, decenas o centenas de miles) se emplean estrategias de selección de atributos para reducir la dimensionalidad de los vectores que los representan. Esto permite remover atributos que no aportan información y seleccionar a aquellos con características relevantes para la recuperación lo que permite discriminarlos del resto de los documentos. Para ello, existen diferentes estrategias de selección de atributos como ganancia de información, frecuencia en documentos o métodos basados en información estadística que han obtenido buenos resultados.

2.4.2 Medida de similitud

Para evaluar la relevancia entre consulta y documento es necesario utilizar una medida de similitud. Las medidas de similitud son capaces de expresar una cantidad numérica entre 0 y 1. Lo que significa que para una consulta q, el documento d_i con la puntuación de similitud más alta será más relevante que un documento d_j con una puntuación de similitud más baja. Existen diferentes tipos de de medidas de similitud, sin embargo, aquí solo mostraremos la utilizada en nuestro trabajo.

Similitud Coseno

La similitud coseno es una de las similitudes más utilizadas y se mide por el ángulo entre dos vectores. Si son paralelos, la distancia del coseno es igual a uno y se dice que los vectores son iguales. Si los vectores son perpendiculares, se dice que son diferentes (sin relación entre sí). La fórmula para obtener el valor de similitud se representa de la siguiente manera.

$$similitud(q, d_i) = \frac{(q \cdot d_i)}{(||q|| \times ||d_i||)} = \frac{\sum_{i=1}^{n} q \times d_i}{\sqrt{\sum_{i=1}^{n}(q)^2} \times \sqrt{\sum_{i=1}^{n}(d_i)^2}} \tag{2.20}$$

donde q corresponde a la consulta y d_i a los documentos dentro del sistema.

2.4.3 Medidas de evaluación

De igual manera que un clasificador, el sistema de recuperación tiene que ser evaluado para medir la calidad al recuperar los documentos. El enfoque estándar utilizado para evaluar un sistema de recuperación se basa en los documentos relevantes recuperados para el usuario. Para ello el sistema toma cada documento de la colección de prueba como relevante o no relevante. A esta decisión se le conoce como el estándar de oro o el juicio de relevancia de la verdad fundamental. Para medir la efectividad de la recuperación de información ad-hoc de manera estándar, necesitamos una colección de prueba que consta de tres cosas: una colección de documentos, un conjunto de prueba con consultas y un conjunto de juicios de relevancia. En una evaluación binaria, el juicio de relevancia será relevante o no relevante para cada par consulta-documento. La recopilación de documentos de prueba y el conjunto de necesidades de información deben tener un tamaño razonable para promediar el rendimiento en un conjunto de pruebas bastante grandes, ya que los resultados son variables en diferentes documentos y necesidades de información. Como regla gen-

eral, se ha determinado que 50 necesidades de información son un mínimo suficiente. Las métricas deben ser capaces de evaluar la importancia de los documentos recuperados y la posición en la que el sistema lo está recuperando. Al igual que en la clasificación de textos las medidas más utilizadas son *Precision*, *Recall* y *F1-score*. Sin embargo, debido a que estás métricas no toman en cuenta la posición relativa de las instancias recuperadas correctamente, que es importante para esta tarea y no necesario en la clasificación para evaluar el sistema de recuperación, en este trabajo utilizaremos las métricas Rango Reciproco Medio (MRR) y precisión a k (P@k), ya que consideran estas variaciones.

Presición a K

La Precisión a K es una medida estadística que evalúa los primeros K elementos recuperados para una consulta e ignora el resto de los documentos por debajo de dicho valor. La idea detrás de esta medida de evaluación es simple, ver cuantos de los K primeros documentos recuperados para cada consulta son correctos. La fórmula para obtener este valor es la siguiente.

$$P@K = \frac{1}{|Q|} \sum_{i=1}^{|Q|} prec_k(q_i) \tag{2.21}$$

donde $prec_k$ indica la precisión de los documentos relevantes en las primeras K posiciones para la consulta q_i. A manera de ejemplo, para entender esta medida de evaluación tomemos una consulta q_1 y un conjunto de documentos recuperados $\{d_5, d_3, d_2, d_1, d_{11}, d_9\}$. Para esta consulta elegiremos un valor de $k = 3$, lo que significa que solo tomaremos los primeros tres documentos (d_5,d_3,d_2), y consideraremos que el conjunto de documentos relevantes es $\{d_5$, d_2 y $d_9\}$. Sin embargo, podemos notar que d_9 se encuentra en la posición número seis por lo que con una $P@3$ este documento quedaría fuera de la evaluación, obteniendo un resultado de $P@3 = 2/3$.

Rango Reciproco Medio

El Rango Reciproco Medio o *Mean Reciprocal Rank (MRR)* es una medida estadística para evaluar cualquier proceso que genera una lista de posibles respuestas a una muestra de consultas Q ordenada por probabilidad de relevancia. El rango recíproco de una respuesta a una consulta es el inverso multiplicativo de la fila de la primera respuesta correcta. Es decir, es el promedio de los rangos recíprocos de resultados para una muestra de consultas. El rango de valores de esta medida va desde 0 hasta 1 donde la posición del documento relevante en la lista de documentos recuperados es importante. La fórmula para obtener este valor es la siguiente.

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{rank_i} \tag{2.22}$$

donde $rank_i$ es la posición en la lista de respuestas en la que se encontró el documento relevante para la i-ésima consulta. Por ejemplo si para una consulta q_1 el sistema recupera los siguientes documentos $\{d_5, d_3, d_2, d_1, d_{11}, d_9\}$ y el primer documento relevante en la lista es d_3 que se encuentra en la posición dos, entonces el valor sería $RR = 1/2$.

2.5 Normalización de reacciones adversas a términos médicos

La tarea de asignar un nombre o concepto a un síntoma, signo, enfermedad, diagnóstico, indicación terapéutica o procedimiento quirúrgico o médico a un vocabulario controlado dentro de un diccionario estándar en el Sistema de Lenguaje Médico Unificado (UMLS) se conoce como normalización. La normalización de reacciones adversas en redes sociales es una tarea propuesta por el foro Social Media Mining

for Health (2022) (SMM4H) con la finalidad de identificar las reacciones asociadas a los medicamentos y así construir un sistema que sea capaz de mantener un monitoreo activo. La normalización de estas reacciones es una tarea desafiante debido al tipo de vocabulario utilizado en redes sociales y la terminología médica entre los profesionales de la salud. Esto se debe principalmente a la falta de conocimiento de los usuarios al momento de expresar su malestar. Por lo que para cerrar esta brecha, se han creado diferentes diccionarios que contienen reacciones adversas con su terminología médica y las diferentes formas en como se refieren a ellas los pacientes.

MedDRA es uno de los diccionarios más utilizados hoy en día, creado en la cuarta Conferencia Internacional sobre Armonización, conformada por un grupo de expertos en medicina.El término MedDRA significa Diccionario Médico para Actividades Reguladoras (Brown et al., 1999). El objetivo de este diccionario es producir una terminología médica única e internacionalmente aceptada para su uso en las fases previas y posteriores a la comercialización del proceso regulatorio de medicamentos.

Figura 2.10: Estructura jerárquica de MedDRA. Todas las siglas fueron tomadas del artículo original para facilitar la explicación. Figura obtenida de Brown et al. (1999).

Este diccionario contiene una estructura jerárquica de cinco niveles como se puede ver en Figura 2.10 donde los "Términos de bajo nivel" (LLT por sus siglas en inglés) constituyen el nivel más bajo de la terminología y se encuentran vinculados a un "Término Preferents" (PT por sus siglas en inglés). Dentro de esta jerarquía se estipula que cada término LLT puede tener alguna de las siguientes relaciones con sus padres PT del tipo:

- **Sinónimos:** Diferentes términos para el mismo concepto inherente al PT.
- **Variantes léxicas:** Distintas formas de palabras para una misma expresión. Estos incluyen abreviaciones, nombres completos y términos médicos con diferente orden de palabras.
- **Cuasi-sinónimos:** Estos son términos que no tienen exactamente el mismo significado, pero se tratan como sinónimos dentro de una terminología dada. Estos incluyen descripciones del sitio y la lateralidad (por ejemplo el término PT: "Otitis externa" con el término LLT: "Bilateral otitis externa").
- **Sub-concepto:** Los sub-conceptos (del concepto PT principal) están representados por un LLT con información más detallada, como la especificidad anatómica (por ejemplo témino PT:"Contusion" con LLT: "Bruising of face" o LLT: "Bruising of leg").
- **LLT idéntico:** Cada LLT contará con un PT idéntico.

Los **Términos Preferentes** son descriptores distintos (concepto médico único) para un síntoma, signo, enfermedad, diagnóstico, indicación terapéutica, investigación, procedimiento quirúrgico o médico. Estos deben ser términos inequívocos, específicos y auto-descriptivos. La especificidad de este nivel debe ser tal que los calificadores clínico-patológicos o etiológicos de los descriptores están representados en el nivel de PT. Por ejemplo, existe una variedad de "rhinitis" y "meningitis"

como entidades separadas en este nivel como PT: “Rhinitis perennial”, PT: “Rhinitis ulcerative”, PT: “Rhinitis atrophic”, PT: “Meningitis aseptic”, PT: “Meningitis cryptococcal”, PT: “Meningitis viral”, PT: “Meningitis bacterial”, etc. Este nivel de especificidad en los PT garantiza que la naturaleza multi-axial de la terminología pueda aprovecharse al máximo. En este nivel cada PT debe estar vinculado al menos a un órgano en el nivel SOC (Sistema de Clasificación de Órgano) por lo que debe esta seguir una ruta del tipo:

$$PT \rightarrow HLT \rightarrow HLGT \rightarrow SOC.$$

Los **Términos de Alto Nivel (HLT por sus siglas en inglés)** son descriptores de orden superior para los PT vinculados a él. Esta es una categoría inclusiva que vincula los PT relacionados con ella por anatomía, patología, fisiología, etiología o función. Los términos HLT están destinados a fines de recuperación y presentación de datos; son un nivel de agrupación y no pretenden ser un nivel de codificación. Al igual que los PT cada HLT se encuentra vinculado con al menos un SOC y sigue una ruta del tipo:

$$HLT \rightarrow HLGT \rightarrow SOC.$$

Los **Grupo de Términos de Alto Nivel (HLGT por sus siglas en inglés)** son descriptores superiores para uno o más términos HLT relacionados por anatomía, patología, fisiología, etiología o función. Los términos HLGT agrupan los términos HLT para facilitar la recuperación por conceptos más amplios y cada HLGT debe estar vinculado al menos a un SOC y al menos a un HLT.

Como último nivel se encuentra el **Sistema de Clasificación de Órganos (SOC)** que agrupa cada término por:

- Etiología (Por ejemplo “Infecciones e infestaciones”)
- Sitio de manifestación (Por ejemplo “Trastornos gastrointestinales”)

- Finalidad (Por ejemplo “Procedimientos quirúrgicos y médicos”)

Una descripción más detallada de la jerarquía puede ser encontrar en artículo de Brown et al. (1999).

2.5.1 Alcances y Limitaciones

- MedDRA cubre diagnósticos, síntomas y signos, reacciones adversas a medicamentos e indicaciones terapéuticas, los nombres y resultados cualitativos de investigaciones, procedimientos quirúrgicos y médicos, e historial médico/social. En general, solo se incluyen términos relevantes para los asuntos regulatorios farmacéuticos.

- MedDRA no comprende una nomenclatura de medicamentos o dispositivos y no contiene términos que cubran el diseño del estudio, la demografía del paciente, los valores numéricos o calificadores como los que describen la gravedad o la frecuencia de la enfermedad.

- MedDRA está destinado a cubrir los productos medicinales de origen biológico y los efectos sobre la salud de los productos sanitarios, y puede utilizarse para registrar los acontecimientos adversos e historial médico en los ensayos clínicos.

- MedDRA no incluye definiciones de términos.

Debido al extenso contenido del diccionario MedDRA en nuestro trabajo nos enfocaremos solo en los datos recomendados por el foro SMM4H que consiste en utilizar solamente los términos etiquetados como PT y LLT. La Tabla 2.2 muestra un ejemplo de un término médico donde cada término médico dentro del diccionario MedDRA se compone por cuatro columnas. La primera corresponde a un código único para el término, la segunda columna corresponde a la relevancia jerárquica, la

tercera columna corresponde a un identificador único de cada término y sus variaciones y la cuarta columna corresponde a los términos médicos (reacciones adversas).

Tabla 2.2: Términos médicos de reacciones adversas en MedDRA.

id_unico	relevancia	id_término	término
C0018681	PT	10019211	Headache
C0018681	LT	10019211	Headache
C0018681	LT	10019218	Headache NOS
C0018681	LT	10019198	Head pain
C0018681	LT	10008013	Cephalgia
C0018681	LT	10033405	Pain head
C0009676	PT	10010305	Confusional state
C0009676	LT	10010305	Confusional state
C0009676	LT	10010304	Confusion state
C0009676	LT	10010300	Confusion
C0009676	LT	10010298	Confused
C0009676	LT	10027350	Mental confusion
C0009676	LT	10050462	Feeling dazed

CAPÍTULO 3

TRABAJO RELACIONADO

La extracción automática de reacciones adversas a medicamentos no es un problema nuevo, sin embargo, pese a los esfuerzos realizados por diferentes investigadores y organizaciones aún queda mucho camino para resolver esta problemática. Diferentes trabajos se han enfocado en aportar información adicional que ayude en esta labor como es el caso de SIDER que es una base de datos que cuenta con reacciones adversas asociadas a diferentes medicamentos extraídas de la literatura (Kuhn et al., 2010). Otro recurso enfocado en ayudar en esta labor es MetaMap que es un sistema principalmente léxico utilizado para mapear conceptos en texto biomédico a conceptos en UMLS Meta-thesaurus (Aronson, 2001) y el trabajo realizado por Jagannatha et al. (2019) quienes construyeron un conjunto de datos que contiene de igual manera el medicamento y sus reacciones adversas de registros médicos electrónicos. Sin embargo, estas fuentes de datos se enfocan en textos biomédicos, registros clínicos o informes de noticias, dejando de lado las redes sociales. Hoy en día los trabajos enfocados en redes sociales se basan principalmente en analizar las interacciones sociales y obtener conjuntos de datos por lo que los trabajos enfocados en la identificación y normalización de reacciones adversas son limitados. Por su parte el foro SMM4H se ha enfocado en solucionar esta problemática creando conjuntos de datos etiquetados en Twitter, y proponiendo un enfoque de solución que consta de una clasificación de texto, identificación del tramo del texto que contiene dicha mención y posteriormente

la normalización de la mención a un diccionario médico en diferentes idiomas para que participantes de diferentes partes del mundo puedan proponer e implementar diferentes soluciones.

3.1 Identificación de reacciones adversas en historiales médicos

Diferentes estudios se han realizado en el ámbito de historias clínicas de pacientes, tal es el caso de trabajo realizado por Foreman et al. (2020) quien realizó un estudio manual de historiales de pacientes en Australia donde observaron que la mayoría de las RA son debido a alergias o intolerancia. Por otra parte Harpaz et al. (2013) enfocó su trabajo en analizar la importancia de utilizar múltiples fuentes de información como el sistema de notificación de reacciones adversas (AERS) de la Food and Drug Administration (FDA) e historiales clínicos de pacientes electrónicas (HCE) para analizar si mejoraba la identificación de RA. De la misma manera Bean et al. (2017) realizó un estudio de historiales médico de pacientes en Reino Unido mediante técnicas de aprendizaje computacional como regresión logística, árboles de decisión y máquina de vector de soporte para identificar las reacciones adversas o el trabajo de Casillas et al. (2016) quien mediante un sistema híbrido con un analizador morfosintáctico y semántico y técnicas de aprendizaje profundo Henry et al. (2020) identifica reacciones adversas. Finalmente Christopoulou et al. (2020) propuso un enfoque de extracción de información basado en reconocimiento de entidades nombradas y el uso de redes neuronales bidireccionales a corto plazo y campos aleatorios condicionales.

3.2 Foro de Minería de redes sociales para la salud

El taller de Minería de redes sociales para la salud (SMM4H) es un foro realizado por el laboratorio de Procesamiento del Lenguaje de la Salud del Instituto de Informática Biomédica de la Facultad de Medicina de la Universidad de Pensilvania que se utiliza para reunir a investigadores interesados en métodos automáticos para la recopilación, extracción, representación, análisis y validación de datos de redes sociales (p. ej., Twitter, Facebook)[1]. El objetivo de este foro es compartir tareas relacionadas a la salud, así como los conjuntos de datos y propuestas de evaluación. Algunos de los temas de interés incluyen, pero no se limitan a:

- Métodos para la detección y extracción automática de menciones de conceptos relacionados con la salud en redes sociales
- Mapeo de menciones relacionadas con la salud en las redes sociales a vocabularios estandarizados
- Obtención de tendencias relacionadas con la salud a partir de las redes sociales
- Métodos de recuperación de información para obtener datos relevantes de redes sociales
- Inferencia de datos geográficos o demográficos a partir del discurso de las redes sociales
- Monitoreo de propagación de virus usando las redes sociales
- Estudios de incidencia de enfermedades usando redes sociales
- Clasificación de los mensajes relacionados con la salud en las redes sociales
- Análisis automático de mensajes de redes sociales para vigilancia de enfermedades y educación del paciente.

[1] https://healthlanguageprocessing.org/

La finalidad de este foro es mejorar la prestación y los resultados de la atención médica, el control y la vigilancia de la salud pública a través de innovaciones en el procesamiento automatizado del lenguaje. Este foro es importante para nuestro trabajo debido a que se ha enfocado en solucionar la problemática de identificación de RA en Twitter, proponiendo un enfoque de tres etapas que consiste en una clasificación de tweets que mencionan al menos una reacción adversa a medicamentos, seguido de una identificación y extracción del tramo del texto que menciona dicha reacción para finalmente normalizarlos a un vocabulario médico. Adicional a esto, este foro ha compartido el conjunto de datos y el sistema de evaluación con el que trabaja la mayoría de los trabajos del estado del arte.

3.2.1 Clasificación de tweets

Se han propuesto diferentes métodos para la clasificación de tweets que reportan alguna RA. Estos métodos se pueden clasificar en términos generales en tres grupos. El primer grupo considera un enfoque/metodología de aprendizaje automático tradicional; empleando principalmente características definidas manualmente basadas en una representación de bolsa de palabras, extendida con etiquetas de *part-of-speech* en algunos casos, y utilizando como clasificador una Máquina de Vectores de Soporte (SVM) (Ruchay and Kober, 2017; Sarker et al., 2015; Liza, 2020). Sin embargo los resultados utilizando este tipo de clasificadores suelen ser deficientes en comparación con modelos de redes neuronales. Los autores en estos trabajos comentan que uno de los principales problemas que encontraron fue la falta de datos y el desequilibrio de los mismos, lo que provoca que los clasificadores se equivoquen. El segundo grupo considera representaciones más avanzadas como el uso de *embeddings* de palabras y utiliza enfoques de redes neuronales como redes neuronales convolucionales (CNN) y redes neuronales recurrentes (RNN) para aprender la representación de los tweets y a partir de ellas clasificarlas utilizando NN o SVM (Aduragba et al.,

2020; Wu et al., 2019; Miranda, 2018). Este enfoque logra mejorar los resultados del enfoque tradicional (BOW-SVM), sin embargo, los resultados no son alentadores. Esto se debe principalmente a su inadecuación para tratar colecciones con pocos datos y alto desequilibrio de clases, así como con el uso de lenguaje coloquial, abreviaturas y errores ortográficos y gramaticales, lo que perjudica fuertemente a los *embeddings* de palabras. Finalmente, el tercer grupo se enfoca en el uso de *Transformers* pre-entrenados con diferentes tipos de datos. Estos modelos suelen mostrar un mejor rendimiento que los enfoques anteriores porque permiten considerar el contexto codificando cada palabra de una oración tomando en cuenta el resto de ella. Esta característica le ha permitido a estos modelos obtener mejores resultados que los enfoques previos. Entre los modelos más utilizados en este grupo se encuentra BERT, BERTweet, ChemBERT, RoBERTa, etc. Además de esto, este grupo de trabajo suele enfocarse más en aplicar técnicas de aumento de datos para mitigar el desequilibrio de clases (Ramesh et al., 2021; Sakhovskiy et al., 2021; Pimpalkhute et al., 2021; Aji et al., 2021b; Zhou et al., 2021a; Dima et al., 2021).

La Tabla 3.1 muestra los resultados obtenidos en esta etapa de clasificación. Si bien los resultados no son sobresalientes con respecto a otras tareas, esto se debe en gran medida a la cantidad de datos y el desequilibrio de los mismos como se comentó previamente.

Tabla 3.1: Mejores resultados obtenidos por los participantes en la etapa de clasificación en el foro SMM4H (Sarker and Gonzalez-Hernandez, 2017; Weissenbacher et al., 2018, 2019; Klein et al., 2020).

Participantes	F1	P	R
SMM4H-2017			
NRC_Canada	0.435	0.392	0.488
CSaRUSCNN	0.414	0.437	0.393
NorthEasterNLP	0.412	0.395	0.431
SMM4H-2018			
THU_NGN	0.522	0.442	0.636
IRISA	0.478	0.378	0.649
UHZ	0.445	0.455	0.436
SMM4H-2019			
ICRC	0.6457	0.6079	0.6885
UZH	0.6048	0.6478	0.5671
MIDAS@IIITD	0.5988	0.6647	0.5447
SMM4H-2020			
RoBERTa	0.64	0.62	0.65
EnDR-BERT, ensemble	0.58	0.63	0.54
RoBERTa, SMM4H'17 and SMM4H'19 corpora	0.58	0.52	0.65
SMM4H-2021			
RoBERTa with under and oversampling	0.61	0.52	**0.75**
RoBERTa + ChemBERTa	0.61	0.55	0.68
BERT ensemble with oversampling	0.54	**0.60**	0.49

3.2.2 Extracción de reacciones adversas

Entre los principales enfoques utilizados, notamos que para la identificación y extracción de tramo de texto que mencionan la reacción adversa la mayoría de los enfoques utilizan *Transformers* pre-entrenados con una capa CRF (Conditional Random Field) y BiLSTM-CRF para el Reconocimiento de Entidades Nombradas para etiquetar, segmentar y extraer las reacciones adversas.

Tal es el caso de Yaseen and Langer (2021) donde los autores proponen un enfoque basado en BiLSTM-CRF que recibe como entrada diferentes embeddings apilados, con lo que buscan resaltar la información contenida a nivel de sub-palabras por medio de *Byte-pair embeddings* y utilizando los embeddings generados por BERT en su capa de entrada. Sin embargo, este trabajo no obtuvo buenos resultados debido a las palabras fuera del vocabulario en los embeddings. Otro trabajo enfocado en transformers fue realizado por Sakhovskiy et al. (2021), donde los autores utilizaron una arquitectura de transformer con una capa CRF como salida para realizar la identificación de reacciones adversas. Su aporte consistió en utilizar dos conjuntos de datos obtenidos de CADEC y COMETA como entrenamiento, obteniendo una mejora de tres puntos porcentuales con respecto a trabajos que utilizan un modelo similar. Por otro lado Ji et al. (2021) utilizaron Neural Transition-base Model (NTM) para reconocimiento de entidades nombradas. NTM es un método basado en transiciones para modelar la estructura anidada de menciones siguiendo la idea de la red LSTM. Para ello utilizaron como entrada una representación a nivel de carácter obtenida con una red CNN, *embeddings* de palabras obtenidos de Glove y una representación de palabras contextual usando ELMo, sin embargo, debido a la naturaleza de los conjuntos de datos, solo obtuvieron un pequeño porcentaje de mejora. Por otra parte Zhou et al. (2021b) utilizaron un enfoque basado en BERTweet con una capa CRF para identificar las RA.

Adicional a esto, para realizar la normalización de los tramos de texto a términos médicos, la gran mayoría de los trabajos opta por un enfoque simple que consiste en convertir mención y término médico en vectores y mediante una medida de similitud normalizarlos. La construcción de dicha representación vectorial puede variar en cada, utilizando codificadores basados en redes neuronales como CNN, NN, *embeddings* de palabras u oración y uso de Transformers como BERTweet, BERT, RoBERTa, etc.

Tabla 3.2: Mejores resultados obtenidos por los participantes en la etapa de identificación, extracción y normalización en el foro SMM4H (Sarker and Gonzalez-Hernandez, 2017; Weissenbacher et al., 2018, 2019; Klein et al., 2020).

Modelo	F1	P	R
SMM4H-2019			
KFU NLP	0.432	0.362	0.535
myTomorrows-TUDelft	0.345	0.336	0.355
TMRLeiden	0.312	0.370	0.270
SMM4H-2020			
EnDR-BERT, dictionary, BERT-based similarity metrics, CADEC	0.46	0.48	0.45
BERT, CADEC, SMM4H'17 corpus	0.38	0.34	0.44
RoBERTa, multi-task learning	0.35	0.33	0.38
SMM4H-2021			
EnDR-BERT with data from CADEC and COMETA corpora	0.29	0.301	0.275
ELMO, CharCNN and Glove in trained jointly	0.24	0.317	0.196
BERT with joint NER and Normalization	0.24	0.371	0.178

La Tabla 3.2 muestra los mejores resultados de los participantes en el foro SMM4H en la fase de identificación, extracción y normalización de las menciones de reacción adversa. En esta tabla podemos notar que el rendimiento de los modelos conforme a los años ha bajado considerablemente. Sin embargo, esto puede deberse principalmente al error en cascada obtenido por las dos etapas anteriores (clasificación y extracción), lo que ha llevado a los modelos a obtener un rendimiento inferior al 0.50 en la medida F1. En las siguiente subsecciones se detallarán las soluciones propuestas en cada una de estas subtareas, describiendo no solo sus principales características, sino ventajas y desventajas.

3.3 Discusión

Uno de los principales problemas que notamos en los trabajos realizados en la etapa de clasificación es el problema de los modelos para entrenarse utilizando conjuntos de datos altamente desbalanceados, lo que ha provocado que la gran mayoría no obtenga buenos resultados. Por lo que, debido a esto modelos como los Transformers han logrado obtener un mejor rendimiento al ser modelos pre-entrenados y a su facilidad de generalización. Esto los ha convertido en modelos de última generación obteniendo los mejores resultados en la clasificación, identificación y normalización de RA. Sin embargo, si bien estos modelos obtienen un mejor resultado es importante abordar el problema de desbalance de datos ya que esto afecta al modelo. Por otra parte, los trabajos enfocados en la identificación y extracción del tramo de texto no cuentan con un conjunto de datos de entrenamiento por lo que de la misma manera los modelos propuestos no pueden generalizar las menciones coloquiales como reacción adversa lo que dificulta la extracción, por lo que es necesario buscar una alternativa de enfoque. Finalmente para la normalización de reacciones adversas a términos coloquiales notamos que la mayoría de los trabajos tienen dificultades para obtener representaciones de *embeddings* que logren una buena asociación entre

término coloquial y término médico. Esto puede deberse en gran medida a los errores ortográficos, de escritura y el lenguaje coloquial.

Tomando en consideración lo antes mencionado, para resolver esto en nuestro trabajo proponemos un método que consta de dos etapas. Para la primera etapa de clasificación abordaremos el problema mediante un modelo pre-entrenado que no necesita grandes cantidades de datos para ajustarse a una tarea objetivo como lo son los modelos de Transformers. Además utilizaremos un enfoque de pares de oraciones al considerar algunas oraciones auxiliares derivadas de los datos de entrada como información contextual adicional. Como segunda etapa, nos enfocaremos directamente en la normalización de las menciones a vocabulario médico, mediante un sistema de recuperación de información que a diferencia de los métodos propuestos no necesita un conjunto de entrenamiento. Además al utilizar un enfoque de dos etapas el error en cadena puede ser reducido con respecto al enfoque propuesto en el estado del arte.

CAPÍTULO 4

MÉTODO PROPUESTO

En este capítulo se explica a detalle el método propuesto para la normalización de reacciones adversas a medicamentos en Twitter. En la primera sección se describe de manera general el método propuesto. En la segunda sección se describe el módulo de clasificación a detalle, explicamos a detalle nuestro módulo de clasificación de dos oraciones, nuestro enfoque de generación de oraciones auxiliares y el funcionamiento de cada componente. En la tercera sección se describe nuestro módulo de normalización, explicamos a detalle nuestro enfoque utilizando un sistema de recuperación con diferentes representaciones y se describen a detalle las consideraciones tomadas para los términos médicos.

4.1 Descripción general del método

El método propuesto para la identificación de reacciones adversas a medicamentos se muestra en la Figura 4.1 donde podemos ver que nuestro método consta de dos etapas principales. La primera etapa es de clasificación, su objetivo es detectar los tweets que reportan una reacción adversa para ello consta de un clasificador que asignará a cada tweet una etiqueta dependiendo de su contenido.

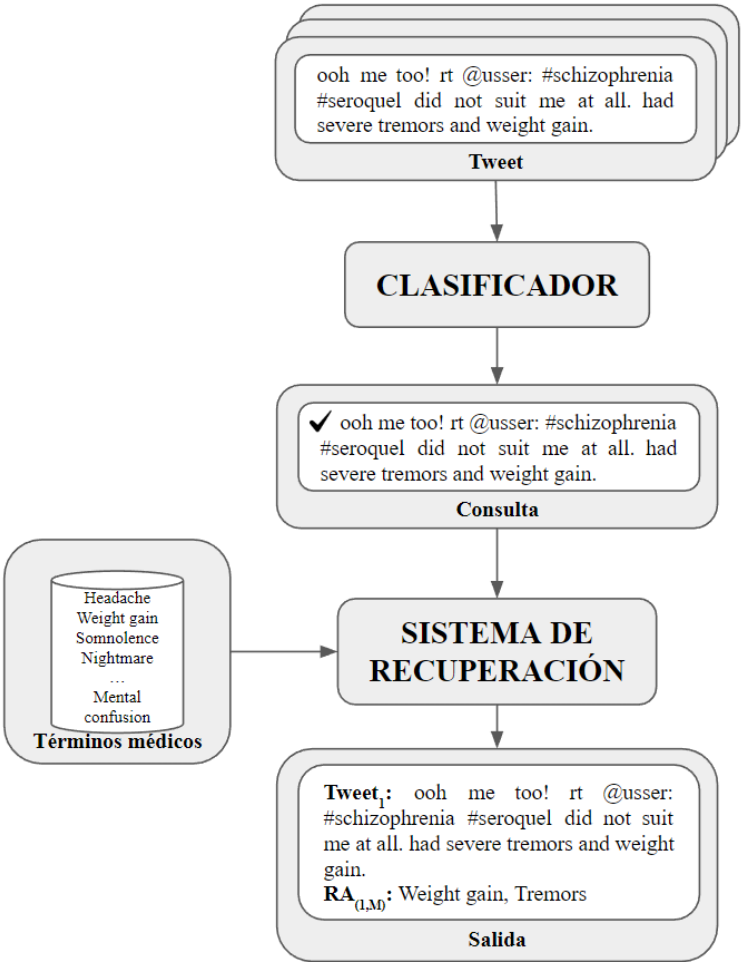

Figura 4.1: Método general para la clasificación y normalización de reacciones adversas a medicamentos.

Estos tweets se etiquetarán como *Ra* si contienen por lo menos una reacción adversa a medicamentos o como *NoRa* en caso contrario. Como segunda etapa contamos con un sistema de recuperación que utilizará los tweets etiquetados como Ra por el clasificador y normalizaremos las menciones de reacciones adversas a una terminología médica. Para ello, en esta etapa los *tweets* son considerados como consultas y los términos médicos obtenidos de MedDRA son considerados como documentos.

Finalmente como salida del método propuesto, se espera recuperar para cada tweet un conjunto de términos médicos que corresponden a las menciones de RA dentro del tweet.

$$\{\text{Tweet}_1, RA_1, RA_2, \ldots, RA_N\}$$
$$\{\text{Tweet}_2, RA_1, RA_2, \ldots, RA_N\}$$
$$\vdots$$
$$\{\text{Tweet}_m, RA_1, RA_2, \ldots, RA_N\}$$

4.2 Módulo de clasificación de tweets

Para nuestro módulo de clasificación optamos por utilizar un modelo de Transformer pre-entrenado debido a la cantidad de datos y a sus ventajas para representar el texto. Además de esto, aprovecharemos la segunda entrada con la que cuenta el modelo para convertir la tarea en una clasificación de pares de oraciones por medio de un generador de oraciones. Para finalmente, obtener una nueva representación y con ayuda de una capa de clasificación asignarle una etiqueta. La Figura 4.2 muestra módulo de clasificación propuesto.

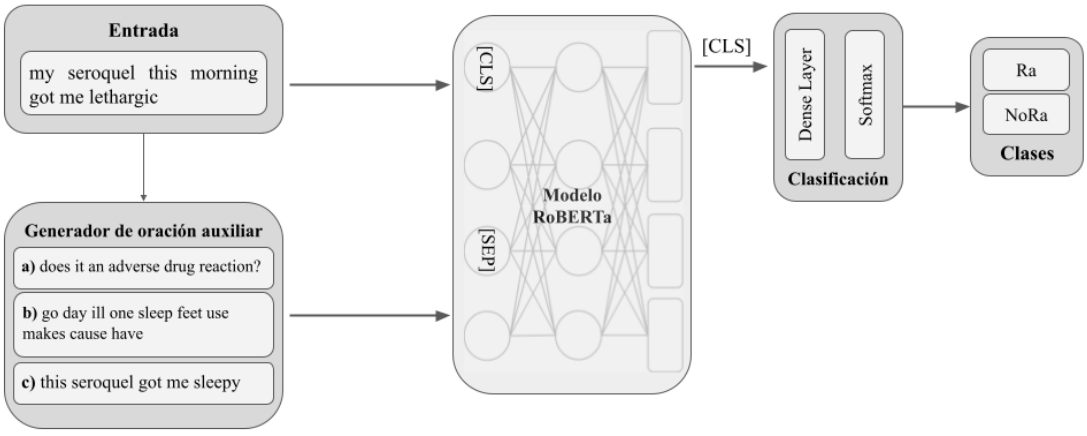

Figura 4.2: Módulo de clasificación de pares de oraciones utilizando $RoBERTa_{LARGE}$

Para elegir el mejor modelo para esta tarea se realizaron diferentes pruebas utilizando BioBERT, BERTweet, RoBERTa, ChemBERTa y BERT, sin embargo, el modelo RoBERTa$_{LARGE}$ fue el modelo que mejores resultados presentó. El modelo RoBERTa fue entrenado utilizando conjuntos de datos obtenidos de BookCorpus, CC-News, OpenWebText, y Stories (Liu et al., 2019). Para re-entrenar el modelo utilizamos un conjunto de datos específico de nuestra tarea. Aprovechando la arquitectura con la que cuenta RoBERTa, en nuestro trabajo utilizamos las dos entradas del modelo. Para la segunda entrada, contamos con un módulo generador de oraciones auxiliares que toma en cuenta la información contenida en el conjunto de entrenamiento para agregar información contextual adicional que ayude en la clasificación, donde el tweet y el generador de oraciones servirán como entrada al modelo. Para generar la oración auxiliar nuestro generador utiliza el tweet de entrada y mediante uno de los tres enfoques propuestos agrega la nueva oración. Una vez que se cuenta con el tweet y la oración auxiliar, se crea un vector de entrada que constará de un token inicial [*CLS*], el tweet de entrada, un token de separación [*SEP*] y la oración auxiliar, obteniendo un nuevo vector [[*CLS*], tweet, [*SEP*], oración auxiliar]. Finalmente el modelo regresará un vector [CLS] contextualizado y mediante una capa densa y una función *softmax* asignará la probabilidad de pertenecer a cada clase (Ra o NoRa).

Para el proceso de generación de oraciones auxiliares como se comentó anteriormente utilizamos tres enfoques diferentes como se puede ver en la Figura 4.2. Los enfoques "Pregunta simple" y "Tweet similar" fueron retomados y adaptados de trabajos anteriores (Sun et al., 2019; Sánchez-Vega and López-Monroy, 2021) y "Palabras relevantes" es un enfoque propuesto por nosotros que se basa en utilizar las palabras con mayor frecuencia dentro de la clase positiva del conjunto de entrenamiento. El objetivo de utilizar una oración auxiliar como entrada adicional es ayudar a contextualizar los tweets y guiar a la red para ajustar mejor sus pesos. A continuación se detalla cada uno de los enfoques.

- **Pregunta Simple**. Motivados por los excelentes resultados de BERT en la tarea de Preguntas y Respuestas (BERT QA) (Sun et al., 2019) propuso utilizar una pregunta relacionada al tema de interés como oración auxiliar. Siguiendo esta misma idea, se planteó utilizar la siguiente pregunta "*Does it report an adverse drug reaction?*" como oración auxiliar.

- **Tweet Similar**. Con base en la aplicación de BERT en tareas de similitud textual semántica, Sánchez-Vega and López-Monroy (2021) propusieron utilizar un texto similar a la entrada como oración auxiliar. Siguiendo esta idea, se planteó identificar dentro del conjunto de entrenamiento el tweet con mayor similitud al tweet de entrada. Para este experimento, la clase del tweet con mayor similitud no fue tomado en cuenta, esto con la idea de ayudar al clasificador en el aprendizaje al ofrecerle un punto de comparación entre dos oraciones similares que no necesariamente pertenecen a la misma clase. Para la implementación de esta idea se utilizó una representación de bolsa de palabras con un pesado TF-IDF y mediante la distancia coseno se midió la similitud de ambos vectores.

- **Palabras Relevantes**. La idea de este nuevo enfoque es proporcionar a la red un conjunto de palabras altamente relacionadas con la descripción de RA para que pueda evaluar mejor la relación del tweet de entrada con la categoría objetivo. Para ello, se consideró utilizar un umbral N de palabras más frecuentes dentro de la clase positiva. Para elegir dichas palabras no se tomaron en cuenta las *stop words*[1]. Dentro de nuestros experimentos el valor de nuestro umbral que mejores resultados nos dio fue $N = 10$.

[1]**Stop words** palabras sin significado como artículos, pronombres, preposiciones, etc. que son filtradas antes o después del procesamiento de datos en lenguaje natural.

En la Tabla 4.1 se muestra la generación de oraciones auxiliares para tweets etiquetados como Ra y NoRa. Para los enfoques de pregunta simple y palabras relevantes podemos notar que la oración generada es la misma sin importar la clase. Mientras que para el tweet similar este oración va cambiando.

Tabla 4.1: Ejemplo del generador de oraciones auxiliares para tweets etiquetados como Ra y NoRa.

	Ra	**NoRa**
Tweet de entrada	*my seroquel this morning got me lethargic*	*Does anyone have a lozenge?*
Pregunta Simple	Does it report an adverse drug reaction?	Does it report an adverse drug reaction?
Tweet Similar	this seroquel got me sleepy	I need a lozenge. Does anyone have a lozenge ?
Palabras Relevantes	go day ill one sleep feet use makes cause have	go day ill one sleep feet use makes cause have

4.3 Módulo de recuperación de reacciones adversas

Para nuestro módulo de recuperación de reacciones adversas contamos con cuatro componentes principales, los documentos, la consulta, la representación del texto y la medida de similitud. La Figura 4.3 muestra la arquitectura general de nuestro sistema de recuperación.

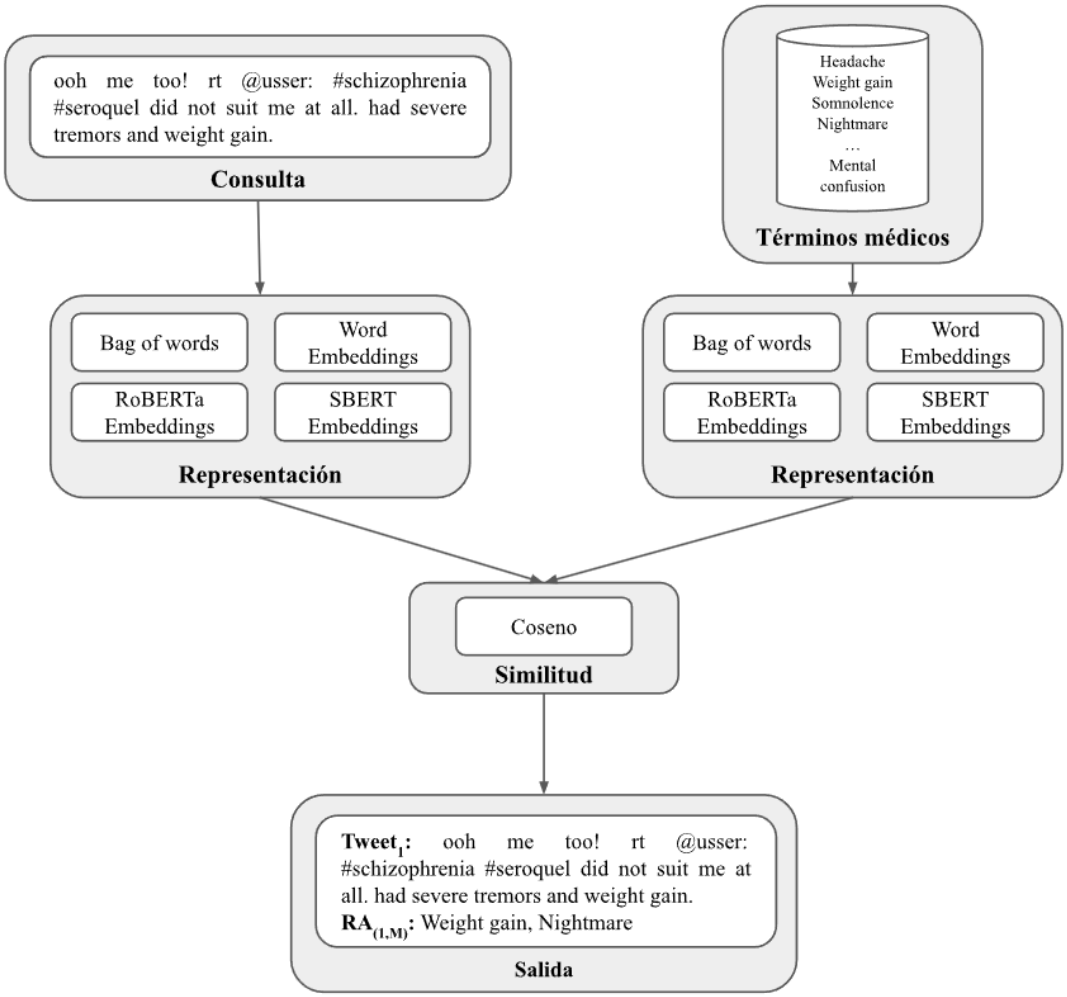

Figura 4.3: Arquitectura general del módulo de recuperación de información.

Para construir nuestro sistema de recuperación de información el primer elemento son los documentos donde se realizará la búsqueda dada una consulta, estos serán los términos médicos obtenidos del conjunto de datos de MedDRA. Éste conjunto de datos contiene un total de 95,911 instancias, sin embargo, dada la estructura del diccionario de MedDRA se realizó una reducción del conjunto, por lo que solamente se tomaron en cuenta los términos etiquetados como Términos Preferentes (PT) y los Términos de Bajo Nivel (LT). Dado que muchos de los términos médicos están relacionados con problemas congénitos, enfermedades, diagnósticos, proced-

imiento quirúrgico o médicos. Realizamos una reducción de instancias, para ello se realizó un análisis estadístico y elegimos un conjunto de palabras que no tienen relación con reacciones adversas como "Test", "Theraphy", "Transplant", "Virus", "Tumor", "Lymphoma", "Congenital", "Neonatal", etc. Sin embargo, para realizar una mejor reducción es importante la ayuda de un experto. Una vez aplicado este filtro el número de términos médicos resultante, tomando en cuenta sinónimos y variantes léxicas fueron 22,341.

Por otra parte dado que los términos médicos se encuentran asociados mediante un código único como es el caso del término "Cephalgia" que se asocia con "Headache", "Pain head", "Head pain" mediante el código "C0018681". Para construir el conjunto de documentos utilizamos dos enfoques diferentes llamados "Términos separados" y "Términos juntos". Para el primer enfoque decidimos utilizar cada término médico dentro del diccionario como documento separado de tal manera que "Cephalgia", "Headache", "Pain head", "Head pain" se cuentan como documentos separados. Para el segundo enfoque agrupamos los términos médicos por código creando un solo documento de tal manera que "Cephalgia Headache Pain head Head pain" cuenta como un solo documento.

Como segundo elemento tenemos las consultas, estas serán los tweet etiquetados como Ra en la etapa de clasificación. Como tercer elemento tenemos la representación del texto y las consultas, para ello probaremos diferentes representaciones de texto como bolsa de palabras, embeddings de palabras y *embeddings* de oraciones obtenidos de diferentes modelos de BERT. Para la representación de bolsa de palabras, utilizaremos un esquema de pesado IDF y TF-IDF con ayuda de n-gramas de palabras. Para los *embeddings* de palabras utilizaremos *embeddings* pre-entrenados de Glove y crearemos *embeddings* a partir de los conjuntos de datos existentes. Además, utilizaremos los *embeddings* generados por RoBERTa en la etapa de clasificación. Por otra parte, para los *embeddings* de oraciones utilizaremos *embeddings* generados por el modelo SBERT que fue entrenado específicamente para generar

este tipo de representación. Una vez que contamos con la representación del texto, esto generará vectores que contendrán la representación de los tweets y términos médicos.

Como cuarto elemento tenemos la medida de similitud para ello utilizaremos el vector de la consulta y el vector de cada documento y mediante la similitud coseno se le asignará a cada término médico la relevancia dado un tweet. Esto dará como resultado un conjunto de términos médicos recuperados en orden de relevancia. Finalmente como salida se espera que nuestro sistema sea capaz de asociar para cada tweet los términos médicos con mayor relevancia mencionados de manera implícita. Por ejemplo, para el siguiente tweet *"rt @user: my philly dr prescribed me trazodone,1pill made me so fkn sick, couldnt move 2day.xtreme migraine, puke, shakes. any1else"* los términos médicos esperados son los siguientes ("*C0857027: Malaise*", "*C0149931: Migraine*", "*C0042963: Emesis*", "*C0040822: Tremor*").

CAPÍTULO 5

EXPERIMENTOS

En este capítulo se describe de manera detallada los conjuntos de datos utilizados, el pre-procesamiento de los datos aplicado y las configuraciones utilizadas para nuestro método propuesto, con la finalidad de facilitar la reproducibilidad de los resultados. Adicional a esto, se muestran los resultados obtenidos por nuestro método en sus dos etapas, una comparación con el estado del arte y un análisis de los resultados.

5.1 Datos

El conjunto de datos utilizado para evaluar nuestro método fue proporcionado por el foro de evaluación Social Media Mining for Health (SMM4H) en su edición 2021[1]. Para obtener estos datos fue necesario realizar un registro previo. La Tabla 5.1 muestra la distribución general del conjunto de datos. En esta tabla podemos notar que el conjunto de entrenamiento contiene un gran desequilibrio entre las clases Ra y NoRa, donde solo el 7% de los tweets de entrenamiento y validación corresponden a la clase positiva. Para el conjunto de Prueba los organizadores decidieron mantener las etiquetas de manera privada con la finalidad de garantizar una evaluación justa. Por lo que para evaluar nuestros modelos, las predicciones deben ser enviadas a

[1] https://healthlanguageprocessing.org/

través de la plataforma Codalab[2] (Pavao et al., 2022) para poder compararnos con el estado del arte.

Tabla 5.1: Distribución de los conjuntos de datos obtenidos del foro SMM4H.

Partición	Ra	NoRa	Todos
Entrenamiento	1,258	16,449	17,707
Validación	65	884	913
Prueba	-	-	10,909

5.2 Configuración experimental

5.2.1 Pre-procesamiento de texto

Uno de los procesos más importantes previos a una clasificación o extracción de características es el pre-procesamiento de los datos. Debido a que el tipo de lenguaje utilizado en redes sociales es coloquial, éste presenta varios retos al momento de trabajar con ellos, como el uso excesivo de abreviaturas, errores ortográficos y gramaticales, uso de emojis, hashtags, urls, etc. Tener un control sobre estas palabras y expresiones resulta de gran ayuda para evitar el ruido en las cadenas de texto. Para la etapa de clasificación, el pre-procesamiento utilizado fue el siguiente:

- Se llevó a cabo un enmascaramiento de los usuarios y ligas de internet, la cual consistió en reemplazar todos los nombres de usuarios y ligas de internet por las palabras "USER_" y "URL_", esto con la finalidad de evitar sesgos en la clasificación. El enmascaramiento fue realizado con ayuda de la librería RegEx[3], enfocada en el manejo de expresiones regulares.

[2]https://competitions.codalab.org/

[3]https://docs.python.org/3/library/re.html

- Para la expansión de hashtag se utilizó la librería tweet-preprocessor [4]

De la misma manera, para el sistema de recuperación se aplicó un pre-procesamiento diferente, esto con la finalidad de reducir el tamaño de nuestros vectores y resaltar la información importante. El pre-procesamiento utilizado fue el siguiente:

- Todo el texto fue convertido a minúsculas
- Se aplicó una expansión de hashtags utilizando la librería tweet-preprocessor
- Se removieron todos los emoticones utilizando la librería ekphrasis[5]
- Se removieron todos los símbolos y números mediante expresiones regulares
- Se removieron las *stop words*. Estas son palabras frecuentes que no llevan a ninguna información como pronombres, preposiciones, conjunciones, etc. Para ello se utilizó la librería NLTK[6]
- Se aplicó lematización al texto. Para ello se utilizó la librería NLTK

5.2.2 Configuración de RoBERTa

Para elegir la configuración de los hiper-parámetros se realizaron pruebas con diferentes configuraciones tomando en cuenta los trabajos relacionados (Ramesh et al., 2021; Yaseen and Langer, 2021). Para cada prueba se re-entrenó el modelo $RoBERTa_{LARGE}$ utilizando los siguientes hiper-parámetros: batch size de 32, tamaño máximo de token de 250, número total de épocas de 15, tasa de aprendizaje de $1e^{-5}$, optimizador Adam y dropout de 0.1. Para el conjunto de validación y entrenamiento se decidió tomar una partición de 80-20 respectivamente.

[4] https://pypi.org/project/tweet-preprocessor/
[5] https://github.com/cbaziotis/ekphrasis
[6] https://www.nltk.org/

5.2.3 Sistema de recuperación

Los experimentos realizados con el sistema de recuperación consistieron en utilizar los enfoques llamados "Términos separados" y "Términos juntos" para construir el conjunto de documentos. Además de los diferentes tipos de representaciones del texto y diferentes tipos de pre-procesamiento del mismo.

Sistema de recuperación básico

Para este enfoque utilizamos un sistema de recuperación clásico que consiste en utilizar una representación de Bolsa de palabras con una ponderación IDF y TF-IDF. Además de esto, realizamos pruebas de unigramas, bigramas y trigramas para generar nuestros vectores de documentos y consultas. Finalmente, con ayuda de la similitud coseno se obtuvo la relevancia de los documentos dada una consulta.

Sistema de recuperación con word embeddings

Para éste sistema se utilizó una representación de *embeddings* de palabras obtenidos con ayuda de Word2Vec y Glove. Con Word2Vec los *embeddings* fueron generados a partir de los conjuntos de entrenamiento, validación y prueba, mientras que con Glove[7] los *embeddings* fueron pre-entrenados con más de 2,000 millones de tweets. Para los *embeddings* generados con Word2Vec se realizaron pruebas con diferentes longitudes (50,100 y 200); de la misma manera, para los *embeddings* obtenidos de Glove se probaron longitudes de 100 y 200. Para ambos casos la longitud que mejor resultado mostró fue 200. Para generar los vectores de documento y consulta, se realizó una suma de vectores de palabras, generando un vector de oración de la misma longitud (200). Finalmente, mediante la similitud coseno se obtuvo la relevancia de los documentos dada una consulta.

[7] https://huggingface.co/fse/glove-twitter-200

Sistema de recuperación con BERT embeddings

Para este sistema se utilizó una representación basada en los *embeddings* generados por BERT. Para ello se utilizó el modelo RoBERTa utilizado en la etapa de clasificación. El procedimiento seguido fue ingresar como entrada todos los documentos para obtener su representación en el SR. Una vez que contamos con el sistema se obtuvo la representación de la consulta y mediante la similitud coseno se obtuvo la relevancia de los documentos. De la misma manera se utilizó el modelo SBERT entrenado para generar *embeddings* de oraciones con más de mil millones de ejemplos. Decidimos utilizar SBERT debido a que el modelo promete que los *embeddings* de oraciones generados son de propósito general.

5.3 Resultados de la clasificación

Para elegir el modelo, primero fue necesario comparar diferentes modelos y arquitecturas. La Tabla 5.2 muestra los resultados obtenidos en la clase de interes (Ra) por cada modelo, donde podemos notar que el uso de una red CNN con *embeddings* de palabras fue el modelo con el rendimiento más bajo. Esto puede deberse al tipo de lenguaje utilizado en los tweets, ya que los errores ortográficos pueden provocar que existan palabras fuera del vocabulario dentro de los *embeddings* pre-entrenados. Por otra parte, notamos que utilizar GRU y ATT puede mejorar el rendimiento. Sin embargo, el problema de las palabras fuera del vocabulario se mantiene. Por otra parte un modelo de *transformer*, debido a que separa cada palabra por tokens, puede tener una ventaja sobre modelos que utilizan solamente *embedings* como se puede ver en la tabla. Se realizaron pruebas con tres modelos BERT diferentes, el primero es BioBERT que fue entrenado en documentos de cohorte biomédico, utilizando como fuente de datos Artículos de PubMed, Wikipedia y BookCorpus. El segundo fue BERTweet que fue entrenado utilizando tweets de dominio general y

por último tenemos RoBERTa que fue entrenado con BookCorpus, CC-News, OpenWebText y Stories. Como podemos ver en la tabla, el modelo que mejor rendimiento presentó fue RoBERTa$_{LARGE}$ superando al modelo entrenado con tweets y al modelo entrenado con artículos de PubMed.

Tabla 5.2: Comparación de resultados utilizando diferentes enfoques de referencia. Las métricas de evaluación son la medida F1, precisión y recuerdo (R) sobre la clase Ra.

Modelo	F1	P	R
CNN	0.20	0.24	0.17
CNN+GRU+ATT	0.39	0.50	0.33
BioBERT	0.43	0.41	0.46
BERTweet$_{BASE}$	0.41	0.43	0.49
RoBERTa$_{BASE}$	0.46	0.45	0.47
RoBERTa$_{\text{LARGE}}$	**0.51**	**0.55**	**0.49**

Como experimento adicional se optó por realizar pruebas utilizando la segunda entrada del modelo RoBERTa, esto con el fin de agregar información adicional al momento de la clasificación. La Tabla 5.3 muestra los resultados del modelo clásico y el modelo que utiliza pares de oraciones. En esta tabla podemos notar que el uso de una oración auxiliar, independientemente de cómo se generó, tiene un impacto positivo en el desempeño de RoBERTa. Sin embargo, a pesar de la mejora que este enfoque representa podemos ver que el modelo que utiliza una *pregunta simple* fue el mejor. Esto puede explicarse por la aplicación exitosa anterior de esta arquitectura en tareas de respuesta a preguntas. No obstante, es importante prestar principal atención en los otros dos enfoques ya que si bien mejoran el modelo clásico, al ser superados por una pregunta simple nos indican que la información agregada con la segunda oración no fue suficiente para ayudar al modelo por lo que buscar diferentes maneras de generar la segunda oración es importante.

Tabla 5.3: Comparación de resultados utilizando el modelo clásico de RoBERTa y el modelo de pares de oraciones. Las métricas de evaluación son la medida F1, precisión y recuerdo (R) sobre la clase Ra.

Modeo	F1	P	R
RoBERTa	0.51	0.55	0.49
RoBERTa (pregunta simple)	**0.60**	0.58	**0.63**
RoBERTa (tweet similar)	0.53	0.59	0.49
RoBERTa (palabras relevantes)	0.58	**0.60**	0.56

Debido a que los datos con los que contamos para entrenar nuestro modelo se encuentran desbalanceados se optó por evaluar la relevancia de aplicar estrategias de balanceo de datos como *Oversampling* y *Undersampling*. La Tabla 5.4 muestra los resultados obtenidos aplicando dichas estrategias. Para ambos casos, se aplicaron diferentes proporciones de aumento o disminución de datos. Para el caso de aumento de datos, el porcentaje indica en que proporción de datos se hizo crecer la clase minoritaria (Ra) con respecto a la clase mayoritaria (NoRa), por ejemplo, para un valor de 50%, esto indica que las instancias de la clase Ra se duplicaron hasta llegar a ser el 50% del tamaño de la clase NoRa. Mientras que para la disminución de datos, esta relación indica el porcentaje de instancias que se mantuvieron de la clase mayoritaria, por lo que un valor de 50% indica que se eliminaron el 50% de las instancias de la clase mayoritaria. Para el aumento de instancias estas fueron duplicadas hasta llegar a ser el porcentaje indicado y para la disminución de instancias esto se hizo de manera aleatoria. En la Tabla 5.5 se muestra un ejemplo de lo antes mencionado donde el valor 6,967 corresponde a la cantidad de datos necesarios para llegar al 50% de la clase NoRA y el valor 8,225 corresponde a la cantidad de datos a eliminar para llegar al 50%.

Tabla 5.4: Comparación de resultados utilizando técnicas de muestreo. Las métricas de evaluación son la medida F1, precisión y recuerdo (R) sobre la clase Ra.

Modelo	F1	P	R	F1	P	R
	Oversampling			Undersampling		
RoBERTa 0%	0.60	0.58	0.59	-	-	-
RoBERTa 25%	0.61	0.51	**0.73**	0.56	0.46	**0.71**
RoBERTa 50%	**0.64**	**0.59**	0.69	0.53	0.58	0.48
RoBERTa 75%	0.61	**0.59**	0.63	0.58	**0.59**	0.56
RoBERTa 100%	0.62	0.58	0.67	**0.60**	0.58	0.59

Los resultados obtenidos al aplicar estas técnicas de balanceo de datos se puede observar en la Tabla 5.4, donde podemos notar que aplicar un aumento de datos (*oversampling*) mejora el rendimiento de nuestro modelo, sin embargo, al sobrepasar el 50% el rendimiento decae. Por otra parte eliminar instancias no logra mejorar el rendimiento del modelo, lo que nos indica la importancia de trabajar en enfoques que nos ayuden a aumentar los datos de la clase Ra. Por otra parte, podemos atribuir estos resultados al hecho de que estas técnicas no agregaron variabilidad léxica y sintáctica al conjunto de entrenamiento.

Tabla 5.5: Ejemplo de la distribución de los conjuntos de datos utilizando técnicas de balanceo de datos al 50%.

Partición	Ra	NoRa	Todos
Entrenamiento	1,258	16,449	17,707
Entrenamiento con oversampling	1,258+6,967	16,449	24,674
Entrenamiento con undersampling	1,258	16,449-8,225	9,483

Tabla 5.6: Comparación de resultados con el estado del arte (Magge et al., 2021). Las métricas de evaluación son la medida F1, precisión y recuerdo (R) sobre la clase Ra.

Modelo	F1	P	R
RoBERTa with under and oversampling	0.61	0.52	**0.75**
RoBERTa + ChemBERTa	0.61	0.55	0.68
BERT ensemble with oversampling	0.54	**0.60**	0.49
BERTweet with (pseudo) data	0.49	0.59	0.42
BERT trained with class weights	0.46	0.47	0.46
BERTweet with class weights	0.46	0.52	0.41
RoBERTa with data augmentation	0.40	0.41	0.40
BERT	0.23	0.14	0.73
RoBERTa pregunta simple con oversampling	**0.64**	0.59	0.69

Finalmente, la Tabla 5.6 muestra una comparación de nuestro mejor modelo con el estado del arte. En esta tabla podemos observar dos cosas, la primera es que todos los modelos utilizan un modelo de *transformer* y la segunda es que la mayoría de modelos utilizaron diferentes enfoques para aumentar los datos. Además, también podemos notar que los mejores resultados fueron obtenidos por el modelo RoBERTa. Sin embargo, la mejora obtenida fue mínima entre cada modelo. Lo que nos indica que si bien los métodos de muestreo pueden mejorar el rendimiento, éstos no son suficientes, por lo que buscar diferentes alternativas de aumento de datos que agreguen diversidad léxica es una mejor opción.

El análisis de errores mostró que los tweets clasificados erróneamente por las diferentes configuraciones de nuestro método, tanto falsos positivos como negativos, son muy diversos en contenido y estilo.

Sin embargo, después de un cuidadoso análisis manual, distinguimos las siguientes tres causas frecuentes:

- Tweets que reportan una RA utilizando un estilo humorístico o sarcástico. Por ejemplo, el tweet "*Cool, the Humira commercial just made me lose my appetite*".

- Tweets que informan una RA utilizando lenguaje coloquial e incluso blasfemias. Por ejemplo, el tweet "*crap. forgot to take my doubled dose of fluoxetine and now i'm busy wanting to die a hell of a lot*".

- Tweets que narran una situación completa y no mencionan directamente la reacción adversa, por ejemplo, "*fell asleep and woke back up again. trying a quarter of a seroquel instead of half this time. i'd like to get out of bed tomorrow*".

- Tweets que contienen varios errores gramaticales, abreviaturas o términos médicos, por ejemplo, "*@user I've had Cipro before. Luckily for me, the only side fx I tend to get from AB is gastic upset. But I RARELY use AB*".

Estos errores nos muestran que el pre-procesamiento es una parte fundamental antes de la clasificación y se debe prestar principal atención en tratar los errores ortográficos y abreviaturas, ya que pueden generar confusión al modelo.

5.4 Resultados del sistema de recuperación

Los experimentos iniciales se realizaron utilizando el conjunto de Validación debido a que no contamos con las etiquetas del conjunto de Entrenamiento y evaluarnos en el sistema proporcionado por el foro SMM4H es un proceso lento. Esto se debe principalmente a que el foro tiene un número limitado de tres envíos por día. Una vez elegido el mejor método éste será evaluado sobre el conjunto de Prueba para podernos comparar con el estado del arte.

Tabla 5.7: Resultados experimentales utilizando un sistema de recuperación básico. Los resultados fueron obtenidos utilizando el conjunto de validación.

Sistema de recuperación básico	P@1	P@2	P@3	P@5	P@10	MRR
Términos juntos (TF-IDF + 1-grama)	0.09	0.12	0.13	0.17	0.29	0.15
Términos juntos (TF-IDF + 12-grama)	0.09	0.13	0.17	0.20	0.33	0.16
Términos juntos (TF-IDF + 123-grama)	0.09	0.12	0.16	0.21	0.33	0.15
Términos juntos (IDF + 1-grama)	0.08	0.12	0.14	0.21	0.31	0.14
Términos juntos (IDF + 12-grama)	0.08	0.12	0.15	0.20	0.29	0.14
Términos juntos (IDF + 123-grama)	0.07	0.12	0.16	0.20	0.29	0.13
Términos separados (TF-IDF + 1-grama)	0.12	0.21	0.24	0.28	0.41	0.21
Términos separados (TF-IDF + 12-grama)	**0.16**	**0.31**	**0.33**	**0.37**	**0.48**	**0.27**
Términos separados (TF-IDF + 123-grama)	**0.16**	**0.31**	**0.33**	**0.37**	**0.48**	**0.27**
Términos separados (IDF + 1-grama)	0.15	0.26	0.29	0.33	0.47	0.25
Términos separados (IDF + 12-grama)	**0.16**	**0.31**	**0.33**	**0.37**	**0.48**	**0.27**
Términos separados (IDF + 123-grama)	**0.16**	**0.31**	**0.33**	**0.37**	**0.48**	**0.27**

Dicho lo antes mencionado, dado que nuestro sistema de recuperación obtiene un conjunto de términos médicos dado un tweet, las métricas utilizadas para evaluar cada sistema de recuperación fueron precisión a N (P@N) y Rango Reciproco

Medio (MRR). La Tabla 5.7 muestra los resultados obtenidos para el sistema de recuperación básico utilizando diferentes configuraciones. La razón de utilizar P@N con diferentes valores se debe a que notamos que cada tweet puede contener más de una reacción adversa. Como podemos ver en la tabla hay una clara diferencia entre los sistemas que utilizan los *Términos Juntos* y *Términos separados*. Esto se debe principalmente a los términos médicos específicos que no suelen ser mencionados en los tweets, lo que provoca que al ser agregados a un mismo documento (Términos juntos) la similitud asignada sea menor, mientras que a los términos que contienen palabras comunes se les asigna una mayor similitud, lo que provoca que el SR agregue una mayor relevancia a los documentos con palabras más comunes. De igual manera el mantener documentos cortos con variantes léxicas que suelen utilizarse más por los pacientes benefician más al sistema. Sin embargo, también notamos que el uso de n-gramas fue benéfico teniendo una repercusión en el aumento de la precisión. Sin embargo, el uso de una representación de bolsa de palabras limita al sistema a una identificación a nivel de palabra por lo que errores ortográficos, uso de sinónimos o maneras coloquiales pueden afectar al momento de recuperar los términos médicos.

Tabla 5.8: Resultados experimentales utilizando un sistema de recuperación con *embeddings* de palabras. Los datos fueron obtenidos utilizando el conjunto de validación.

Sistema de recuperación con *embeddings* de palabras	P@1	P@2	P@3	P@5	P@10	MRR
Términos juntos (Glove)	0.01	0.01	0.02	0.02	0.03	0.02
Términos separados (Glove)	0.00	0.01	0.01	0.02	0.02	0.01
Términos juntos (Word2Vec)	0.01	0.02	0.02	0.04	0.06	0.03
Términos separados (Word2Vec)	0.01	0.03	0.05	0.06	0.07	0.04

La Tabla 5.8 muestra los resultados obtenidos con nuestro sistema de recuperación utilizando *embeddings* de palabras donde podemos ver que los resultados son inferiores al sistema de recuperación básico. Esto se debe a que el sistema de recuperación va sumando todos los *embeddings* de palabras para generar el vector de representación, lo que provoca que los términos médicos con más palabras se les asigne una mayor puntuación y por lo tanto se recupere una cantidad mayor de términos equivocados para un tweet. En la tabla también podemos notar que entre estos dos sistemas de recuperación el mejor resultado se obtuvo con los *embeddings* pre-entrenados con los tweets y términos médicos, sin embargo, éstos no son suficientes para obtener *embeddings* de calidad y los problemas asociados con los errores ortográficos persisten.

Tabla 5.9: Resultados experimentales utilizando un sistema de recuperación con BERT embeddings. Los datos fueron obtenidos utilizando el conjunto de validación.

Sistema de recuperación con BERT embeddings	P@1	P@2	P@3	P@5	P@10	MRR
Términos juntos (RoBERTa)	0.00	0.00	0.01	0.02	0.02	0.02
Términos separados (RoBERTa)	0.00	0.00	0.02	0.02	0.05	0.02
Términos juntos (SBERT)	0.00	0.00	0.00	0.01	0.02	0.02
Términos separados (SBERT)	0.00	0.00	0.01	0.02	0.05	0.02

Para nuestro Sistema de recuperación con BERT tomamos en cuenta dos modelos diferentes. El primero es el modelo utilizado en la etapa de clasificación RoBERTa$_{LARGE}$ y el segundo es el modelo SBERT[8] enfocado en la similitud de oraciones, entrenado con más de mil millones de pares de oraciones. Los resultados se muestran en la Tabla 5.9 donde podemos ver que utilizar esta estrategia al igual que utilizar *embeddings* de palabras no obtiene buenos resultados.

[8]https://www.sbert.net/docs/pretrained_models.html

Esto puede deberse a que los tweets y los términos médicos tienen una longitud diferente. Además de esto, los modelos no fueron entrenados para este tipo de tareas y la falta de ejemplos de tweets con RA coloquiales para cada término médico dificulta que el modelo pueda realizar una mejor relación entre término médico y tweet.

Como se observa en los experimentos realizados, utilizar un sistema de recuperación básico con bolsa de palabras y utilizar recursos que cuenten con términos alternos a los médicos ayuda a obtener mejores resultados. Sin embargo, buscar alternativas robustas para corregir los errores ortográficos y buscar alternativas que ayuden a comparar dos oraciones con diferentes longitudes puede ser beneficioso para el sistema. Finalmente, debido a que los trabajos del estado del arte se evalúan utilizando las métricas de *precisión*, *recuerdo* y *F1*, para evaluar nuestro modelo se optó por enviar las primeras posiciones obtenidas de nuestro sistema de recuperación. Es importante mencionar que en el caso de que el tweet contenga más de una reacción adversa nuestro sistema solo será capaz de acertar en una de las reacciones adversas y todas las demás serán consideradas como errores.

Tabla 5.10: Comparación de resultados obtenidos con el estado del arte. Las métricas de evaluación son la medida F1, precisión y recuerdo (R) sobre la clase Ra. El símbolo $\star$ indica nuestro enfoque propuesto.

Modelo	F1	P	R
nDR-BERT + CADEC and COMETA data	**0.29**	0.30	0.28
$\star$ **Sistema de Recuperación (una sola respuesta)**	0.25	0.24	0.26
ELMO, CharCNN and Glove in trained jointly	0.24	0.32	0.20
BERT with joint NER and Normalization	0.24	**0.37**	0.20
BERTweet and similarity measures	0.20	0.14	**0.34**
Multi-task learning with selective oversampling	0.16	0.16	0.17

Para evaluarnos con el estado del arte, utilizamos el sistema de recuperación básico con los Términos separados y con un pesado IDF con unigrama, bigrama y trigrama para obtener una mejor representación, elegimos esta configuración ya que consideramos que la frecuencia de términos (TF) no es relevante para nuestra tarea. La Tabla 5.10 muestra los resultados del estado del arte donde podemos notar que todos los trabajos utilizan modelos de Transformers. Sin embargo, debido a que no se cuenta con un conjunto de entrenamiento para esta tarea y a que los resultados son dependientes de la etapa de clasificación y la etapa de identificación y extracción del tramo de la mención; podemos notar que el rendimiento de los modelos es bajo. Además de esto, podemos notar que el sistema de recuperación al no depender de un conjunto de entrenamiento logra obtener resultados equiparables a los obtenidos en el estado del arte, obteniendo el segundo lugar en la tabla. Lo que nos puede indicar la importancia de tener un conjunto de datos que ayude al modelo BERT.

Tabla 5.11: Comparación de resultados utilizando P@1 y P@3. Las métricas de evaluación son la medida F1, precisión y recuerdo (R) sobre la clase Ra. El símbolo ⋆ indica nuestro enfoque propuesto.

Modelo	F1	P	R
nDR-BERT + CADEC and COMETA data	0.29	0.30	0.28
⋆Sistema de Recuperación (una respuesta)	0.25	0.24	0.26
⋆Sistema de Recuperación (tres respuestas)	0.44	0.47	0.46

Adicional a esto, realizamos un experimento que consiste en evaluar nuestro sistema de recuperación tomando en cuenta las tres primeras posiciones, es decir, los tres primeros términos médicos recuperados. Esto se debe a que en el conjunto de Validación observamos que una porción de los tweet contenían más de una reacción adversa. Para ello obtuvimos las matrices de confusión de tres envios del sistemas de recuperación enviando una sola respuesta (solo la primera posición, solo la segunda

posición y solo la tercera posición) y sumamos los TP y TN para generar una nueva matriz de confusión y así obtener las métricas *F1*, *P* y *R* mostradas en la Tabla 5.11, donde podemos notar que utilizar un sistema de recuperación simple con variantes léxicas de los términos médicos puede obtener resultados equiparables con el estado del arte en comparación a arquitecturas de Redes Neuronales, no obstante el utilizar un sistema basado en similitud a nivel de palabras no soluciona la problemática que se tiene con los tweets donde la reacción adversa no contiene palabras en común. Esto se puede ver en la Tabla 5.12 donde entender el contexto es necesario para recuperar la reacción adversa correcta, por lo que los sistemas de recuperación que utilizaron bolsa de palabras no son capaces de recuperarlos, mientras que los que utilizaron *embeddings* los recuperaron en posiciones inferiores a mil.

Tabla 5.12: Ejemplos de menciones coloquiales de reacciones adversas.

Tweet	**Reacción adversa**
the pathway of cipro is definitely murdering my mind right now ? donedonedone	Mental impairment
cant cope coming off venlafaxine feel like iv got a hangover mixed with flu symptoms. total nausea & i cant keep still. please hurry & pass	Drug withdrawal syndrome, Agitation
just woke up. since i started on the higher dose of quetiapine i'm sleeping even more & i feel knackered when i wake.	Somnolence

Por otra parte, otro de los principales problemas que notamos al momento de recuperar los términos médicos para cada tweet fue la variabilidad léxica y los errores ortográficos que provocaron que no se realizara una correcta similitud. Esto se ilustra en la Figura 5.1, en donde podemos observar el poco traslape entre las

menciones de las reacciones adversas dentro de los tweets y los términos médicos esperados, indicando la importancia de utilizar métodos que logren relacionar las palabras debido a su contexto y no solo a nivel de palabra. Adicional a esto, la integración de una persona experta en el área de la salud es de suma importancia, debido a que el diccionario médico utilizado no solo contiene reacciones adversas, lo que provoca que existan términos médicos que sean fácilmente confundidos.

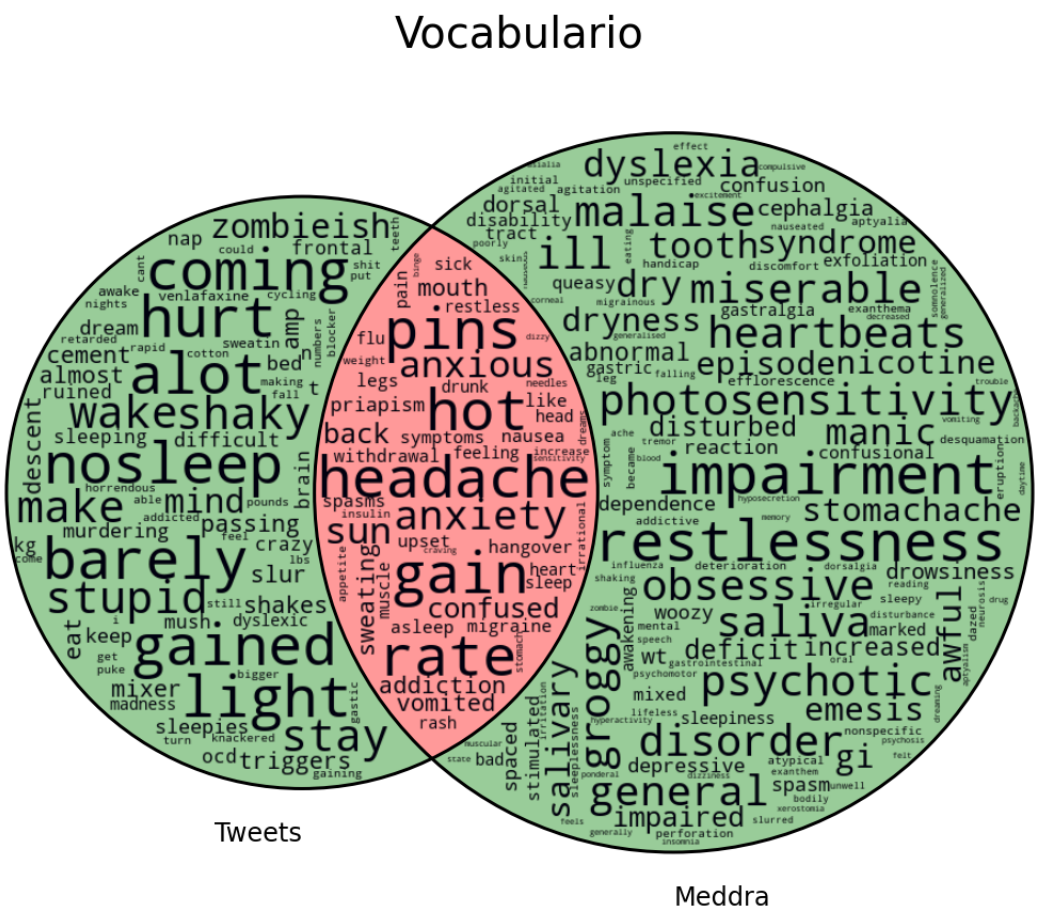

Figura 5.1: Visualización de palabras de los términos médicos esperados y los términos coloquiales dentro de los tweets. Es importante resaltar que el tamaño de las palabras dentro del diagrama de Venn es solo estético y no representa el número de apariciones o relevancia de las palabras.

CAPÍTULO 6

CONCLUSIONES Y TRABAJO FUTURO

La identificación de tweets que informan reacciones adversas a medicamentos en Twitter es una tarea compleja y desafiante debido a múltiples factores, como: el tamaño de los tweets, el lenguaje coloquial, errores ortográficos, gramaticales y el lenguaje médico específico, así como el alto desequilibrio en los conjuntos de datos con los que contamos hasta el momento. En este trabajo, presentamos un método basado en *Transformer*, que aborda el problema de clasificación como una clasificación de pares de oraciones al considerar la generación de oraciones auxiliares a partir de los tweets de entrada como información contextual adicional. En consecuencia, se propusieron tres enfoques para generar oraciones auxiliares a partir de los tweets de entrada, tomando como motivación las aplicaciones para las que se diseñó la arquitectura del *transformer* de dos entradas, tales como respuesta a preguntas y similitud textual semántica.

Los experimentos realizados mostraron que el enfoque de pares de oraciones para la clasificación de textos es un enfoque que mejora el rendimiento general del modelo clásico de $RoBERTa_{LARGE}$ y que la construcción de la segunda oración es un proceso al que se debe prestar una mayor atención debido a que de esta depende la mejora del modelo. El mejor rendimiento obtenido por nuestros clasificadores fue el modelo que agrega una pregunta simple, esto nos indica que el uso de palabras

relevantes y el tweet con mayor similitud a pesar de mejorar el rendimiento general no aportaron la información adicional esperada al momento de la clasificación. Por lo que es importante explorar formas diferentes para generar estas oraciones que ayuden a obtener mejores resultados. Adicional a esto, notamos que el conjunto de tweets etiquetados erróneamente se debe en gran medida al tipo de lenguaje utilizado, errores ortográficos y uso de abreviaturas, por lo que es importante buscar estrategias de pre-procesamiento que ayuden a corregir este tipo de problemas.

Por otra parte, los experimentos realizados con el sistema de recuperación mostraron la complejidad de la tarea de normalización. Los problemas principales que encontramos fue la dificultad de comparar vectores de diferentes tamaños utilizando los *embeddings* de palabras y los *embeddings* generados por BERT, ya que notamos que mientras el término médico contenga más palabras se le asigna una mayor relevancia, lo que provoca que el sistema no recuperen los términos médicos correctos. Por otra parte, consideramos importante entrenar *embeddings* específicos para esta tarea, debido a que al comparar términos como "headpain" y "Headache" que tienen un significado altamente relacionado los *embeddings* asignan una similitud menor al 60%, mientras que comparar ambos términos con un término más específico como "Cephalgia" la similitud es aún menor lo que provoca que el sistema de recuperación recuperen términos médicos incorrectos en las primeras posiciones.

El método propuesto en este trabajo muestra que utilizar un sistema de recuperación para la normalización de términos médicos puede obtener resultados equiparables a los del estado del arte. Adicional a esto, el uso del SR también nos permite realizar la tarea opuesta, es decir, dado un término médico (reacción adversa) de interés, encontrar aquellos tweets que lo mencionan, y con ello brindar otras posibilidades para construir herramientas que apoyen a expertos en el monitoreo continuo y análisis de información. Este trabajo muestra que utilizar un enfoque diferente al uso de las redes neuronales puede obtener un resultado competitivo en la tarea de normalización de reacciones adversas. Sin embargo, sabemos que

el sistema de recuperación tiene sus desventajas frente a modelos más robustos que utilizan representaciones complejas que logran capturar el contexto.

A continuación, se exponen algunas ideas que se proponen como trabajo futuro para la presente investigación:

- Buscar alternativas para generar oraciones auxiliares debido a los bajos resultados obtenidos con las tres propuestas en este trabajo.
- Explorar otras formas de aumento de datos diferentes a las utilizadas, que se enfoquen en agregar información adicional léxica debido a que se cuenta con un conjunto reducido de datos.
- Explorar diferentes enfoques aplicando reconocimiento de entidades nombradas para la extracción de reacciones adversas, utilizando diferentes representaciones y conjuntos de datos.
- Plantear diferentes enfoques de normalización con el uso de representaciones más sofisticadas, basadas en modelos de lenguaje neuronales.

APPENDIX A

ARTÍCULOS PUBLICADOS

Los artículos derivados de esta tesis se listan a continuación.

- *Does This Tweet Report an Adverse Drug Reaction? An Enhanced BERT-Based Method to Identify Drugs Side Effects in Twitter.* José Alberto Fuentes-Carbajal, Manuel Montes-y-Gómez, Luis Villaseñor-Pineda. 14th Mexican Conference on Pattern Recognition, MCPR-2022 Ciudad Juarez, Chihuahua, México, Lectures Notes in Computer Science 13264, Springer, 2022.

Referencias

Aduragba, O. T., Yu, J., Senthilnathan, G., and Crsitea, A. (2020). Sentence contextual encoder with BERT and BiLSTM for automatic classification with imbalanced medication tweets. In *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*, pages 165–167, Barcelona, Spain (Online). Association for Computational Linguistics.

Aizenberg, I., Aizenberg, N. N., and Vandewalle, J. P. (2000). *Multi-valued and universal binary neurons: Theory, learning and applications*. Springer Science & Business Media.

Aji, A. F., Nityasya, M. N., Wibowo, H. A., Prasojo, R. E., and Fatyanosa, T. (2021a). Bert goes brrr: A venture towards the lesser error in classifying medical self-reporters on twitter. In *Proceedings of the Sixth Social Media Mining for Health (# SMM4H) Workshop and Shared Task*, pages 58–64.

Aji, A. F., Nityasya, M. N., Wibowo, H. A., Prasojo, R. E., and Fatyanosa, T. (2021b). BERT goes brrr: A venture towards the lesser error in classifying medical self-reporters on Twitter. In *Proceedings of the Sixth Social Media Mining for Health (#SMM4H) Workshop and Shared Task*, pages 58–64, Mexico City, Mexico. Association for Computational Linguistics.

Aronson, A. R. (2001). Effective mapping of biomedical text to the umls metathe-

saurus: the metamap program. In *Proceedings of the AMIA Symposium*, page 17. American Medical Informatics Association.

Baeza-Yates, R., Ribeiro-Neto, B., et al. (1999). *Modern information retrieval*, volume 463. ACM press New York.

Bean, D. M., Wu, H., Iqbal, E., Dzahini, O., Ibrahim, Z. M., Broadbent, M., Stewart, R., and Dobson, R. J. (2017). Knowledge graph prediction of unknown adverse drug reactions and validation in electronic health records. *Scientific reports*, 7(1):1–11.

Berlin, J. A., Glasser, S. C., and Ellenberg, S. S. (2008). Adverse event detection in drug development: recommendations and obligations beyond phase 3. *American journal of public health*, 98(8):1366–1371.

Brown, E. G., Wood, L., and Wood, S. (1999). The medical dictionary for regulatory activities (meddra). *Drug safety*, 20(2):109–117.

Casillas, A., Pérez, A., Oronoz, M., Gojenola, K., and Santiso, S. (2016). Learning to extract adverse drug reaction events from electronic health records in spanish. *Expert Systems with Applications*, 61:235–245.

Christopher Olah (2015). Understanding lstm networks.

Christopoulou, F., Tran, T. T., Sahu, S. K., Miwa, M., and Ananiadou, S. (2020). Adverse drug events and medication relation extraction in electronic health records with ensemble deep learning methods. *Journal of the American Medical Informatics Association*, 27(1):39–46.

Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Coloma, P. M., Trifirò, G., Patadia, V., and Sturkenboom, M. (2013). Postmarketing safety surveillance. *Drug safety*, 36(3):183–197.

Coppersmith, G., Dredze, M., and Harman, C. (2014). Quantifying mental health signals in Twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 51–60, Baltimore, Maryland, USA. Association for Computational Linguistics.

Dechter, R. (1986). Learning while searching in constraint-satisfaction problems.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dima, G.-A., Cercel, D.-C., and Dascalu, M. (2021). Transformer-based multi-task learning for adverse effect mention analysis in tweets. In *Proceedings of the Sixth Social Media Mining for Health (#SMM4H) Workshop and Shared Task*, pages 44–51, Mexico City, Mexico. Association for Computational Linguistics.

Edo-Osagie, O., De La Iglesia, B., Lake, I., and Edeghere, O. (2020). A scoping review of the use of twitter for public health research. *Computers in biology and medicine*, 122:103770.

Edwards, I. R. and Aronson, J. K. (2000). Adverse drug reactions: definitions, diagnosis, and management. *The lancet*, 356(9237):1255–1259.

Foreman, C., Smith, W. B., Caughey, G. E., and Shakib, S. (2020). Categorization of adverse drug reactions in electronic health records. *Pharmacology Research & Perspectives*, 8(2):e00550.

Fukushima, K. and Miyake, S. (1982). Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In *Competition and cooperation in neural nets*, pages 267–285. Springer.

Ghannay, S., Favre, B., Esteve, Y., and Camelin, N. (2016). Word embedding evaluation and combination. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 300–305.

Ginn, R., Pimpalkhute, P., Nikfarjam, A., Patki, A., O'Connor, K., Sarker, A., Smith, K., and Gonzalez, G. (2014). Mining twitter for adverse drug reaction mentions: a corpus and classification benchmark. In *Proceedings of the fourth workshop on building and evaluating resources for health and biomedical text processing*, pages 1–8. Citeseer.

Harpaz, R., Vilar, S., DuMouchel, W., Salmasian, H., Haerian, K., Shah, N. H., Chase, H. S., and Friedman, C. (2013). Combing signals from spontaneous reports and electronic health records for detection of adverse drug reactions. *Journal of the American Medical Informatics Association*, 20(3):413–419.

He, R., Ravula, A., Kanagal, B., and Ainslie, J. (2020). Realformer: Transformer likes residual attention. *arXiv preprint arXiv:2012.11747*.

Henry, S., Buchan, K., Filannino, M., Stubbs, A., and Uzuner, O. (2020). 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *Journal of the American Medical Informatics Association*, 27(1):3–12.

Hochreiter, S., Bengio, Y., Frasconi, P., Schmidhuber, J., Kolen, J., and Kremer, S. (2001). A field guide to dynamical recurrent neural networks. *chapter Gradient Flow in Recurrent Nets: The Difficulty of Learning Long-Term Dependencies*, pages 237–243.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Hu, H., Phan, N., Chun, S. A., Geller, J., Vo, H., Ye, X., Jin, R., Ding, K., Kenne, D., and Dou, D. (2019). An insight analysis and detection of drug-abuse risk behavior

on twitter with self-taught deep learning. *Computational Social Networks*, 6(1):1–19.

Jagannatha, A., Liu, F., Liu, W., and Yu, H. (2019). Overview of the first natural language processing challenge for extracting medication, indication, and adverse drug events from electronic health record notes (made 1.0). *Drug safety*, 42:99–111.

Ji, Z., Xia, T., and Han, M. (2021). Paii-nlp at smm4h 2021: Joint extraction and normalization of adverse drug effect mentions in tweets. In *Proceedings of the Sixth Social Media Mining for Health (# SMM4H) Workshop and Shared Task*, pages 126–127.

Kim, Y. (2014). Convolutional neural networks for sentence classification. empirical methods in natural language processing.

Klein, A., Alimova, I., Flores, I., Magge, A., Miftahutdinov, Z., Minard, A.-L., O'connor, K., Sarker, A., Tutubalina, E., Weissenbacher, D., et al. (2020). Overview of the fifth social media mining for health applications (# smm4h) shared tasks at coling 2020. In *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*, pages 27–36.

Kohn, L. T., Corrigan, J. M., Donaldson, M. S., et al. (2000). Institute of medicine. to err is human: building a safer health system.

Kuhn, M., Campillos, M., Letunic, I., Jensen, L. J., and Bork, P. (2010). A side effect resource to capture phenotypic effects of drugs. *Molecular systems biology*, 6(1):343.

Lazarou, J., Pomeranz, B. H., and Corey, P. N. (1998). Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies. *Jama*, 279(15):1200–1205.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Lee, K. (2008). *The World Health Organization (WHO)*. Routledge.

Lian, A. T., Du, J., and Tang, L. (2022). Using a machine learning approach to monitor covid-19 vaccine adverse events (vae) from twitter data. *Vaccines*, 10(1):103.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized BERT pretraining approach. *Computing Research Repository (CoRR) - arXiv*, abs/1907.11692.

Liza, F. F. (2020). Sentence classification with imbalanced data for health applications. In *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*, pages 138–145.

Ma, J., Xie, S., Jin, M., Lianxin, J., Yang, M., and Shen, J. (2020). Xsysigma at semeval-2020 task 7: Method for predicting headlines' humor based on auxiliary sentences with ei-bert. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1077–1084.

Magge, A., Klein, A., Miranda-Escalada, A., Al-Garadi, M. A., Alimova, I., Miftahutdinov, Z., Farre, E., Lima-López, S., Flores, I., O'Connor, K., et al. (2021). Overview of the sixth social media mining for health applications (# smm4h) shared tasks at naacl 2021. In *Proceedings of the Sixth Social Media Mining for Health (# SMM4H) Workshop and Shared Task*, pages 21–32.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Miranda, D. S. (2018). Automated detection of adverse drug reactions in the biomedical literature using convolutional neural networks and biomedical word embeddings. *arXiv preprint arXiv:1804.09148*.

Nikfarjam, A., Sarker, A., O'connor, K., Ginn, R., and Gonzalez, G. (2015). Pharmacovigilance from social media: mining adverse drug reaction mentions using

sequence labeling with word embedding cluster features. *Journal of the American Medical Informatics Association*, 22(3):671–681.

Otter, D. W., Medina, J. R., and Kalita, J. K. (2020). A survey of the usages of deep learning for natural language processing. *IEEE transactions on neural networks and learning systems*, 32(2):604–624.

O'Connor, K., Pimpalkhute, P., Nikfarjam, A., Ginn, R., Smith, K. L., and Gonzalez, G. (2014). Pharmacovigilance on twitter? mining tweets for adverse drug reactions. In *AMIA annual symposium proceedings*, volume 2014, page 924. American Medical Informatics Association.

Pavao, A., Guyon, I., Letournel, A.-C., Baró, X., Escalante, H., Escalera, S., Thomas, T., and Xu, Z. (2022). Codalab competitions: An open source platform to organize scientific challenges. *Technical report*.

Pedro Borges (2018). Deep learning: Recurrent neural networks.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Perambai, Abhishek. (2019). A deep dive into the world of gated recurrent neural networks: Lstm and gru.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Pimpalkhute, V., Nakhate, P., and Diwan, T. (2021). IIITN NLP at SMM4H 2021 tasks: Transformer models for classification on health-related imbalanced Twitter

datasets. In *Proceedings of the Sixth Social Media Mining for Health (#SMM4H) Workshop and Shared Task*, pages 118–122, Mexico City, Mexico. Association for Computational Linguistics.

Pirmohamed, M., James, S., Meakin, S., Green, C., Scott, A. K., Walley, T. J., Farrar, K., Park, B. K., and Breckenridge, A. M. (2004). Adverse drug reactions as cause of admission to hospital: prospective analysis of 18 820 patients. *Bmj*, 329(7456):15–19.

Raghupathi, W. and Raghupathi, V. (2014). Big data analytics in healthcare: promise and potential. *Health information science and systems*, 2(1):1–10.

Ramesh, S., Tiwari, A., Choubey, P., Kashyap, S., Khose, S., Lakara, K., Singh, N., and Verma, U. (2021). Bert based transformers lead the way in extraction of health information from social media. In *Proceedings of the Sixth Social Media Mining for Health (# SMM4H) Workshop and Shared Task*, pages 33–38.

Ruchay, A. and Kober, V. (2017). Impulsive noise removal from color images with morphological filtering. In *International Conference on Analysis of Images, Social Networks and Texts*, pages 280–291. Springer.

Sakhovskiy, A., Miftahutdinov, Z., and Tutubalina, E. (2021). Kfu nlp team at smm4h 2021 tasks: Cross-lingual and cross-modal bert-based models for adverse drug effects. In *Proceedings of the Sixth Social Media Mining for Health (# SMM4H) Workshop and Shared Task*, pages 39–43.

Sánchez-Vega, F. and López-Monroy, A. P. (2021). Bert's auxiliary sentence focused on word's information for offensiveness detection. *IberLEF*, 2943:259–269.

Sarker, A., Ginn, R., Nikfarjam, A., O'Connor, K., Smith, K., Jayaraman, S., Upadhaya, T., and Gonzalez, G. (2015). Utilizing social media data for pharmacovigilance: a review. *Journal of biomedical informatics*, 54:202–212.

Sarker, A. and Gonzalez-Hernandez, G. (2017). Overview of the second social media mining for health (smm4h) shared tasks at amia 2017. *Training*, 1(10,822):1239.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47.

Sebastiani, F. (2005). Text categorization. In *Encyclopedia of database technologies and applications*, pages 683–687. IGI Global.

Social Media Mining for Health (27 de Noviembre de 2022). `https://healthlanguageprocessing.org`.

Sultana, J., Cutroneo, P., and Trifirò, G. (2013). Clinical and economic burden of adverse drug reactions. *Journal of pharmacology & pharmacotherapeutics*, 4(Suppl1):S73.

Sun, C., Huang, L., and Qiu, X. (2019). Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. *Computing Research Repository (CoRR) - arXiv*, abs/1903.09588.

Thompson, M. A., Majhail, N. S., Wood, W. A., Perales, M.-A., and Chaboissier, M. (2015). Social media and the practicing hematologist: Twitter 101 for the busy healthcare provider. *Current hematologic malignancy reports*, 10(4):405–412.

Wang, B., Wang, A., Chen, F., Wang, Y., and Kuo, C.-C. J. (2019). Evaluating word embedding models: methods and experimental results. *APSIPA transactions on signal and information processing*, 8.

Weissenbacher, D., Sarker, A., Magge, A., Daughton, A., O'Connor, K., Paul, M., and Gonzalez, G. (2019). Overview of the fourth social media mining for health (smm4h) shared tasks at acl 2019. In *Proceedings of the fourth social media mining for health applications (# SMM4H) workshop & shared task*, pages 21–30.

Weissenbacher, D., Sarker, A., Paul, M., and Gonzalez, G. (2018). Overview of the third social media mining for health (smm4h) shared tasks at emnlp 2018. In *Proceedings of the 2018 EMNLP workshop SMM4H: the 3rd social media mining for health applications workshop & shared task*, pages 13–16.

Williams, R. J. and Zipser, D. (1989). A learning algorithm for continually running fully recurrent neural networks. *Neural computation*, 1(2):270–280.

Wohlgenannt, G., Chernyak, E., and Ilvovsky, D. (2016). Extracting social networks from literary text with word embedding tools. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 18–25, Osaka, Japan. The COLING 2016 Organizing Committee.

Wu, C., Wu, F., Yuan, Z., Liu, J., Huang, Y., and Xie, X. (2019). Msa: Jointly detecting drug name and adverse drug reaction mentioning tweets with multi-head self-attention. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, WSDM '19, page 33–41, New York, NY, USA. Association for Computing Machinery.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., and Hovy, E. (2016). Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489.

Yaseen, U. and Langer, S. (2021). Neural text classification and stacked heterogeneous embeddings for named entity recognition in smm4h 2021. *arXiv*.

Yu, S., Su, J., and Luo, D. (2019). Improving bert-based text classification with auxiliary sentence and domain knowledge. *IEEE Access*, 7:176600–176612.

Zhou, T., Li, Z., Gan, Z., Zhang, B., Chen, Y., Niu, K., Wan, J., Liu, K., Zhao, J., Shi, Y., Chong, W., and Liu, S. (2021a). Classification, extraction, and normalization : CASIA_Unisound team at the social media mining for health 2021 shared tasks. In *Proceedings of the Sixth Social Media Mining for Health (#SMM4H) Workshop and Shared Task*, pages 77–82, Mexico City, Mexico. Association for Computational Linguistics.

Zhou, T., Li, Z., Gan, Z., Zhang, B., Chen, Y., Niu, K., Wan, J., Liu, K., Zhao, J., Shi, Y., et al. (2021b). Classification, extraction, and normalization: Casia_unisound team at the social media mining for health 2021 shared tasks. In *Proceedings of the Sixth Social Media Mining for Health (# SMM4H) Workshop and Shared Task*, pages 77–82.