Polo del Conocimiento

Pol. Con. (Edición núm. 70) Vol. 7, No 12
Diciembre 2022, pp. 241-273
ISSN: 2550 - 682X
DOI: 10.23857/pc.v7i12

Clasificación de artículos académicos sobre la pandemia de COVID-19 a través de la técnica de minería de texto Word Embeddings

Classification of academic articles on the COVID-19 pandemic through the Word Embeddings text mining technique

Classificação de artigos acadêmicos sobre a pandemia de COVID-19 por meio da técnica de mineração de texto Word Embeddings

Bayron Fernando Vásquez Vanegas [I]
bayron.vasquezv@hotmail.com
https://orcid.org/0000-0003-3537-2296

Marcos Patricio Orellana Cordero [II]
marore@uazuay.edu.ec
https://orcid.org/0000-0002-3671-9362

Correspondencia: bayron.vasquezv@hotmail.com

Ciencias de la Educación
Artículo de Investigación

*** Recibido:** 23 de octubre de 2022 ***Aceptado:** 12 de noviembre de 2022 *** Publicado:** 5 de diciembre de 2022

I. Ingeniero en Sistemas; Analista de Tecnologías de la Información y Comunicaciones; Regional del Instituto Nacional de Patrimonio Cultural, Investigador independiente, Ecuador.
II. Ingeniero de Sistemas; Docente-Investigador en las líneas de Ciencia de los Datos e Inteligencia Artificial; Master en Gestión de Sistemas de Información e Inteligencia de Negocios; Master en Docencia Universitaria; Profesional con amplia experiencia en bases de datos y dirección de departamentos de TI; Desarrollador de Sistemas de Información Transaccionales y de Inteligencia de Negocios; Docente y Coordinador de la Escuela de Ingeniería de Sistemas y Telemática de la Universidad del Azuay; Responsable del programa de Informática y Director del Laboratorio de Investigación y Desarrollo en Informática (LIDI), Universidad del Azuay, Ecuador.

Resumen

La enfermedad de COVID-19, se introdujo y extendió rápidamente como una pandemia global, que necesitaba ser tratada con respuestas inmediatas, oportunas e integradas a los sistemas. Con la presencia de este virus SARS-CoV-2, la comunidad científica, las organizaciones, los individuos y la sociedad en general, han visto la necesidad de obtener información que pueda aportar conocimiento sobre la evolución de la enfermedad, posibles causas, consecuencias, tratamientos, prevención, entre otros aspectos. El presente artículo propone realizar la clasificación de artículos científicos publicados sobre la pandemia de COVID-19, con la aplicación de técnicas de Machine Learning, a través de mecanismos de representación semántica de palabras como el Word Embeddings y tecnologías basadas en redes neuronales; utilizando el análisis y procesamiento de los abstracts de artículos científicos disponibles en las fuentes de información como LitCovid. Los resultados describen los distintos mecanismos y metodologías de clasificación de texto y las maneras de representar el mismo, con el objeto de construir un modelo de clasificación fundamentado en la técnica de minería de texto Word Embeddings y en redes neuronales basadas en la arquitectura LSTM; obteniéndose la metodología a seguir para clasificar artículos científicos, así como, los resultados de desempeño de los modelos propuestos. Se concluye que, no se logró una predicción con resultados favorables en todas las clases, debido a que los datos están desbalanceados y existen clases muy mayoritarias en comparación a otras, por lo que las predicciones se vieron afectadas.

Palabras Clave: Procesamiento de Lenguaje Natural; Word Embedding; Machine Learning; Redes Neuronales; Clasificación de artículos; COVID-19.

Abstract

The COVID-19 disease was rapidly introduced and spread as a global pandemic, which needed to be treated with immediate, timely and systems-integrated responses. With the presence of this SARS-CoV-2 virus, the scientific community, organizations, individuals and society in general have seen the need to obtain information that can provide knowledge about the evolution of the disease, possible causes, consequences, treatments, prevention, among other aspects. This article proposes to carry out the classification of scientific articles published on the COVID-19 pandemic, with the application of Machine Learning techniques, through mechanisms of semantic

representation of words such as Word Embeddings and technologies based on neural networks; using the analysis and processing of the abstracts of scientific articles available in information sources such as LitCovid. The results describe the different text classification mechanisms and methodologies and the ways of representing it, in order to build a classification model based on the Word Embeddings text mining technique and on neural networks based on the LSTM architecture; obtaining the methodology to follow to classify scientific articles, as well as the performance results of the proposed models. It is concluded that a prediction with favorable results was not achieved in all classes, because the data is unbalanced and there are very majority classes compared to others, so the predictions were affected.

Keywords: Natural Language Processing; Word Embedding; machine learning; Neural Networks; Article classification; COVID-19.

Resumo

A doença COVID-19 foi rapidamente introduzida e disseminada como uma pandemia global, que precisava ser tratada com respostas imediatas, oportunas e integradas aos sistemas. Com a presença deste vírus SARS-CoV-2, a comunidade científica, organizações, indivíduos e a sociedade em geral têm visto a necessidade de obter informações que possam fornecer conhecimento sobre a evolução da doença, possíveis causas, consequências, tratamentos, prevenção, entre outros aspectos. Este artigo propõe realizar a classificação de artigos científicos publicados sobre a pandemia de COVID-19, com aplicação de técnicas de Machine Learning, por meio de mecanismos de representação semântica de palavras como Word Embeddings e tecnologias baseadas em redes neurais; utilizando a análise e tratamento dos resumos de artigos científicos disponíveis em fontes de informação como o LitCovid. Os resultados descrevem os diferentes mecanismos e metodologias de classificação de texto e as formas de representá-lo, de forma a construir um modelo de classificação baseado na técnica de mineração de texto Word Embeddings e em redes neurais baseadas na arquitetura LSTM; obter a metodologia a seguir para classificar os artigos científicos, bem como os resultados de desempenho dos modelos propostos. Conclui-se que não foi alcançada uma previsão com resultados favoráveis em todas as classes, pois os dados estão desbalanceados e há classes muito majoritárias em relação às outras, então as previsões foram afetadas.

Palavras-chave: Processamento de linguagem natural; Incorporação de palavras; aprendizado de máquina; Redes neurais; classificação do artigo; COVID-19.

Introducción

La pandemia de COVID-19 pertenece a la familia de los anteriores virus coronavirus, cuyas cepas producían la gripe común. Sin embargo, en el año 2003 surge la primera mutación, dando origen al SARS que tuvo sus inicios en China, con más de ocho mil cuatrocientos pacientes en veintisiete países diferentes, con una letalidad del diez por ciento. Más adelante, en el año 2012 aparece otra nueva cepa mutante de coronavirus en Arabia Saudita, conocida como MERS-CoV, con un registro de más de dos mil cuatrocientos enfermos y una letalidad del treinta y siete por ciento (Thompson, 2003; BMJ Best Practice, 2020).

El actual coronavirus, conocido como COVID-19, surgió en Wuhan, China y se extendió por todo el mundo; tiene predilección por el árbol respiratorio, de modo que, al penetrar causa una respuesta inmune anormal con características inflamatorias e incremento de las citoquinas, agravando al paciente y causando múltiples daños orgánicos (Maguiña et al, 2020). Análisis genómicos han revelado que el SARS-CoV-2 está asociado con los virus de murciélagos, que son similares al síndrome respiratorio agudo severo. La fuente intermedia de origen y transmisión a humanos se desconoce, no obstante, lo que sí está confirmada es la rápida transferencia de humano a humano (Muhammad, et al, 2020).

Sea cual fuere su origen, lo cierto es que el mundo en general se ha visto grandemente afectado por los efectos producidos por la COVID-19. A la fecha, casi todos los países registran miles de infectados, decesos, secuelas físicas y mentales, así como, graves problemas en la economía. El surgimiento de la reciente enfermedad, ha llevado al mundo a una de las mayores crisis de la historia, en los ámbitos económico, social y de salud, nunca antes visto, afectando múltiples aspectos de la vida cotidiana (Ministerio de Sanidad, 2020).

Ante esta nueva realidad, la comunidad científica ha puesto su mayor esfuerzo en hacerle frente a la pandemia, estudiando y entendiendo el origen del nuevo virus, su comportamiento, y los efectos en la salud y la vida del ser humano, para, de esta manera, poder establecer medidas de prevención eficaces, administrar tratamientos adecuados, desarrollar vacunas, e implementar políticas públicas para la gestión y control de la pandemia. Como resultado de ello, la producción de conocimiento científico acerca de la COVID-19 y el nuevo coronavirus ha crecido a un ritmo sin precedentes.

Según estudios realizados por Wang & Lo (2021), "se han publicado más de 50000 artículos sobre COVID-19 desde principios de 2020 y se siguen publicando varios cientos de artículos nuevos todos los días" (p. 781). Esta enorme tasa de productividad científica sobre COVID-19 lleva a una sobrecarga de información, dificultando que los médicos, enfermeros, bioanalistas, funcionarios de salud pública, gobiernos e investigadores, estén al día con los últimos hallazgos sobre la temática; siendo imprescindible que se mantengan actualizados en lo que respecta a dicha literatura (Wang y Lo, 2021).

Desde el mismo momento que se notificó el primer caso de COVID-19 se inició la publicación de una gran cantidad de estudios, intentado aclarar ciertas incógnitas sobre síntomas, pruebas de detección, medidas preventivas y tratamiento. De hecho, la manera rápida en la que se fue propagando el virus y la repercusión inmediata que tuvo en el individuo y la sociedad, creó la necesidad de tomar medidas serias a nivel mundial, desde el punto de vista de la salud, con base en las evidencias disponibles hasta el momento (Greenhalgh et al, 2020).

De manera que, la pandemia de COVID-19 empezó a dar lugar a un estallido de información recogida en publicaciones científicas, donde cada quince días se van duplicando las referencias; tal y como lo menciona Torres-Salinas (2020), cuando señala que con la llegada de la pandemia, el primer problema que ha debido afrontar al universo de la publicación es la avalancha de preprints y artículos científicos, así como, la necesidad de que estos lleguen a ser accesibles . Cabe acotar que, una de las respuestas por parte de editoriales fue la creación de centros de recursos para unificar en una única web y que sea de acceso abierto, todo lo que se vaya publicando acerca de la COVID-19.

En este mismo orden de ideas, según Torres-Salinas (2020), la tasa de crecimiento bibliométrico según el análisis realizado en la base de datos Dimensions se calcula en $R^2 = 0{,}92$, el mismo que determina que la cantidad de publicaciones realizadas es de alrededor de quinientos artículos diarios. Sin duda, toda esta cantidad de información es el reflejo de los esfuerzos de la comunidad científica para hacer frente a esta crisis sanitaria que ha afectado a múltiples aspectos de la vida cotidiana alrededor del mundo.

Toda esta cantidad de publicaciones son de naturaleza multidisciplinar, siendo así que cualquier entidad o persona interesada en realizar investigación sobre COVID-19 con base a un criterio de interés particular debe realizar la búsqueda e ir clasificando los resultados obtenidos de manera

manual. Esto supone un alto costo en términos de tiempo, siendo ahora más que nunca el recurso tiempo un factor primordial para hacer frente a la pandemia.

El principal interés de los investigadores es extraer información a partir de artículos científicos, con base en un criterio u objetivo determinado, según el área de interés; por lo que herramientas que permitan realizar una clasificación automática de tal información son cada vez más importantes y requeridas por parte de la comunidad científica (Chandrasekaran y Fernandes, 2020). Por este motivo, la utilización de técnicas informáticas de procesamiento y clasificación de datos, permitirían obtener información específica en las distintas bases científicas, y lo más importante, que estén clasificadas.

Por lo antes dicho, se requiere el uso de técnicas informáticas que faciliten la búsqueda, lectura y clasificación de un determinado documento; de forma rápida y precisa ante el exceso de información existente (Wang y Lo, 2021). Esto es importante, pues extraer información de interés particular puede llevar mucho esfuerzo y tiempo, ya que cada uno de los artículos de investigación sobre cierta temática en particular, pertenecen a distintas fuentes y dominios, como la medicina y atención médica, el reconocimiento de patrones, la minería de datos, el aprendizaje automático, entre otros (Sonbhadra et al, 2020).

De hecho, organizaciones, editoriales, bibliotecas virtuales, redes académicas, catálogos, directorios académicos, revistas científicas, entre otros, se han esforzado por organizar la información sobre COVID-19 de tal manera que sea encontrada y por ende de utilidad a la comunidad científica. Por ejemplo, la Organización Panamericana de la Salud ha compilado publicaciones científicas, guías técnicas, recomendaciones y protocolos de investigación en curso de América y el resto del mundo, relacionados con la pandemia actual; siendo esto de utilidad para autoridades, profesionales de la salud, investigadores, y la sociedad en general (Organización Panamericana de la Salud, 2022).

Se habla entonces de una técnica denominada minería de textos, la cual, una de sus principales áreas de aplicación biomédica es la gestión de la sobrecarga de información (Ananiadou et al, 2006; Kilicoglu, 2018; Zweigenbaum et al, 2007). La minería de textos se centra en resolver problemas específicos como recuperar documentos relevantes o extraer parte de la información de dichos documentos. Puede utilizar técnicas para la recuperación, extracción y clasificación de la información; además de aprovechar métodos de campos relacionados, como el lenguaje de procesamiento y la construcción de bases de conocimientos (Cohen & Hersh, 2005).

Hoy más que nunca, es imprescindible tener una visión completa del estado del arte de la literatura relacionada con la COVID-19, debido a razones tales como: organizar y categorizar la literatura; explorar temas de investigación; identificar prioridades y necesidades para generar oportunidades de investigación; entender la evolución de la pandemia; reconocer a los líderes de la investigación en esta área, como investigadores, institutos y centro de investigaciones, países líderes, entre otros; y explorar conexiones entre temas y áreas de investigación.

En este sentido, la clasificación de documentos representa un área admirada de investigación en reconocimiento de patrones y minería de datos. Hoy día, la presencia de repositorios de investigación en línea masivos, llevan a que la búsqueda de artículos de investigación de temas específicos o de interés para el usuario, se convierta en un proceso que demanda mucho tiempo. Los motores de búsqueda disponibles para encontrar documentos mediante palabras clave, son útiles, no obstante, a veces representan una tarea limitante y desafiante (Sonbhadra et al, 2020).

Por lo tanto, este artículo propone una metodología que se enfoca en realizar categorizaciones de artículos científicos publicados sobre COVID-19, mediante la aplicación de técnicas de PNL como el Word Embedding. Este procesamiento de lenguaje natural se ha venido aplicando a documentos médicos que se redactan en textos libres a fin de construir bases de datos que programas computarizados puedan no solo entender, sino también analizar (Friedman & Johnson, 2006).

Con base en lo anterior, cabe acotar que una clasificación automática de documentación mediante la aplicación de técnicas de Procesamiento de Lenguaje Natural (PLN), puede tener un gran impacto al momento de organizar y clasificar artículos de interés por campos y temas; facilitando la tarea de búsqueda de información y brindando soporte a las tareas de investigación para ésta nueva temática sobre COVID-19.

Si bien varios estudios e investigaciones realizadas como Jimenez et al (2020), Jelodar et al (2020) y Dynomant et al (2019), han abordado el tema de la problemática de clasificar artículos o documentos de texto acerca del COVID-19 y problemas de salud en general, es importante conocer si la técnica de PLN conocida como Word Embedding puede brindar una clasificación de artículos que permitan extraer conocimientos relevantes, y brindar soporte a la investigación científica. El word embedding ha demostrado ser una técnica útil en diversas tareas del PLN aparte de la similitud de textos; por lo que en la actualidad tienen gran popularidad (Collobert, et al, 2011; Zou, et al, 2013).

Con la llegada de la pandemia, surgieron proyectos para abordar la problemática antes descrita, como el COVIDScholar; un proyecto que nace del esfuerzo por afrontar los problemas aplicando técnicas de PLN, para agregar, analizar y buscar literatura de investigación acerca del COVID-19, mediante la implementación de una infraestructura automatizada y escalable para buscar e integrar investigaciones recientes tal como éstas aparecen, logrando así, levantar un corpus de más de 81,000 artículos científicos y demás documentos relacionados al COVID-19 (Trewartha et al., 2020).

Por otra parte, para afrontar el desafío que ha provocado la pandemia de COVID-19 en múltiples aspectos, se están empleando mecanismos de PLN y aprendizaje automático sobre los artículos de investigación de la Organización Mundial de la Salud (OMS), con el fin de generar conocimiento que pueda guiar tanto las políticas del COVID-19, investigaciones y desarrollo (Awasthi, et al, 2020). Se aplican enfoques de resumen de texto y los modelos entrenados de Word Embeddings para resumir la información publicada, dando como resultado la herramienta CovidNLP.

Un abordaje teórico sobre Procesamiento de Lenguaje Natural (PNL), Machine Learning (ML) y Word Embeddings (WE)

Actualmente es de interés realizar tareas que procesan el lenguaje natural, es decir, la lengua o idioma hablado o escrito por humanos para propósitos generales de comunicación, mediante el empleo de técnicas o métodos de aprendizaje automático. El objetivo del Procesamiento de Lenguaje Natural (PLN), es estudiar, analizar y emplear algoritmos y metodologías para desarrollar modelos computacionales que puedan ser capaces de procesar idiomas en lenguaje natural, que permitan o faciliten la comunicación entre humanos y máquinas o realicen el procesamiento del habla o texto (Jurafsky & Martin, 2020).

Los enfoques de PLN actualmente incorporan algoritmos de Machine Learning (ML) o aprendizaje automático, este enfoque desarrolla técnicas y algoritmos los mismos que aprenden a realizar ciertas tareas en particular mediante el uso de datos o información que no han sido programados para dicho propósito, esto quiere decir que son capaces de desarrollar un modelo generalizado con un grupo de datos y hacer predicciones sobre datos nuevos (Daud et al, 2017).

Machine Learning (ML) o aprendizaje automático, es una rama de la Inteligencia Artificial, que permite lidiar con el problema de grandes cantidades de información que resultan difíciles de analizar, facilitando la entrega de información confiable y rápida, y la toma de decisiones, en

especial de organizaciones de salud (Pedrero et al, 2021). Por tanto, el Machine Learning tiene como objetivo desarrollar mecanismos y algoritmos que partiendo de un conjunto de datos puedan realizar tareas específicas, sin que hayan sido programados específicamente para ello.

En otro orden de ideas, una de las principales tareas de la clasificación de texto dentro de las tareas de PLN, es la representación del mismo, teniendo como objetivo representar de manera numérica los documentos de texto para que luego puedan ser procesados computacionalmente, para ello, es necesario representar los elementos textuales de los documentos como son palabras, caracteres, n-gramas de palabras o incluso información morfológica como categorías gramaticales etc. Usualmente existen dos tipos de representación que son One-Hot y Representación distribuida o Embeddings.

El mayor avance métodos de representación de palabras llega con el trabajo realizado en 2013 por Mikolov, et al., llamados modelos predictivos. Estos modelos tratan de predecir palabras a partir de las palabras que están cercanas a éstas en términos de vectores más pequeños y densos. Estos métodos basan su concepto en que si se puede predecir el contexto en el cual aparece una palabra, entonces se entiende el significado de ésta en su contexto. Por lo que palabras semánticamente similares estarán cerca entre sí en sus representaciones de espacios vectoriales. A estos métodos se los denomina Word Embeddings. (Mikolov, et al, 2013)

Las técnicas de Word Embedding se han convertido en las principales herramientas dentro de los modelos de PLN, capturando el significado de las palabras y convirtiéndolas a una codificación que puede ser utilizada para todo tipo de redes neuronales. Entre las principales aplicaciones de ésta técnica son: sistemas de traducción; análisis de opinión de textos; generación de textos; chatbox; entre otros.

Algunas técnicas para el PLN son el Word2Vec, el FastText y el Glove. El primero, Word2vec, es un grupo de varios modelos relacionados utilizados para producir word embeddings; que generan representaciones de palabras en vectores, los cuales almacenan la relación semántica entre las mismas; estos vectores resultantes son empleados en distintas tareas de PLN, por lo general tienen cientos de dimensiones para cada una de las palabras en el corpus. Una vez que el modelo se ha entrenado, éste puede detectar sinónimos de palabras o sugerencias de las mismas para una oración. (Mikolov, et al, 2013)

El segundo, FastText, representa una palabra mediante la suma de sus composiciones de caracteres llamados n-grams. Por ejemplo, el vector de la palabra "apple" consiste en la suma de los vectores

n-gram “<ap, app, appl, apple, apple>, ppl, pple, pple>, ple, ple>, le>”. En consecuencia, aplicando ésta técnica, se obtiene una mejor representación de las palabras "raras" que pocas veces aparecen en el cuerpo del texto, y así generar vectores para palabras que no existen en el vocabulario de los Word Embeddings. (Bojanowski, et al, 2017)

El tercero, Glove, es un modelo basado en conteo, en el cual se genera una matriz de gran tamaño que almacena la información de la concurrencia entre palabras y contextos. Es decir, para cada palabra se realiza un conteo de las veces que ésta aparece en algún contexto. El objetivo de entrenamiento de dicha matriz es aprender vectores de forma que el producto escalar entre las palabras sea igual al logaritmo de la probabilidad de co-ocurrencia entre las palabras. El número de contextos es muy alto, por lo tanto, se realiza una factorización de dicha matriz para obtener una de menores dimensiones, dando como resultado mejores representaciones de palabras o Word Embeddings (Pennington, et al, 2014)

Materiales y Métodos

Para el presente estudio en la clasificación de texto se propone un modelo basado en redes neuronales, mediante el empleo de arquitectura LSTM, se emplea este enfoque debido a la ventaja que tienen este tipo de redes de almacenar información para la siguiente iteración y controlar la información que llega de entrada y de salida, de la misma manera como se mencionó previamente en el análisis de las redes neuronales, las redes neuronales de tipo RNN, son empleadas mayormente para tareas de PLN por brindar mejores resultados en dichas tareas.

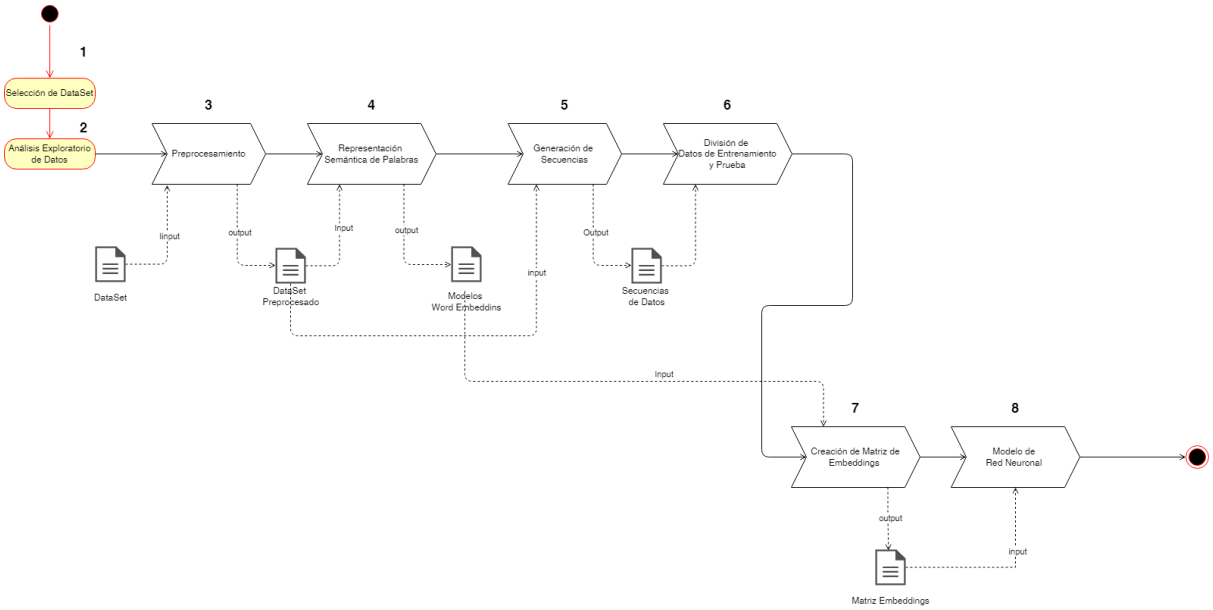
Además, se utiliza un entorno Google Colab, que es una herramienta de Google Research, la cual permite codificar y ejecutar código en lenguaje Python desde el navegador web, esto con el objetivo de obtener las librerías que se requieren para la realizar tareas de PLN. Asimismo, el estudio propone tres modelos de clasificación que emplean la combinación de un modelo de Word Embedding junto con un modelo de red neuronal para la clasificación, a saber,

- Modelo I - Word2Vec + LSTM Bidireccional
- Modelo II - FastText + LSTM Bidireccional
- Modelo III - Glove + LSTM Bidireccional

Para que el texto pueda ser procesado computacionalmente, éste debe ser transformado a una representación que el computador pueda entenderlo, es decir, de forma numérica; para ello, el texto de análisis debe pasar por un proceso de manera que se logre conseguir una representación lo más

aceptable y óptima posible, para que el modelo de aprendizaje profundo pueda realizar de mejor manera las predicciones deseadas. La figura a continuación muestra una representación del proceso a seguir para la metodología propuesta de clasificación de artículos científicos.

Figura 1. Metodología de Clasificación



Metodología propuesta de clasificación de artículos

1. Selección de Conjunto de Datos o DataSet:

El Dataset o conjunto de datos de LitCovid es una recopilación de artículos recientemente publicados, cuyas temáticas están relacionadas con la literatura actual del Coronavirus. Éste conjunto de datos contiene más de 23.000 artículos y en promedio se agregan 2.000 nuevos artículos semanalmente, siendo así un recurso integral para que la comunidad científica pueda actualizarse con información acerca de la crisis que ha provocado la pandemia de la COVID-19. (Jiménez, et al, 2020)

Cada uno de los artículos contenidos en el conjunto de datos de LitCovid, son etiquetados en una de las siguientes temáticas: Prevención, Tratamiento, Diagnóstico, Mecanismo, Reporte de casos, Transmisión, Pronóstico, General. La mayoría de estos artículos pueden ser etiquetados con varias de éstas etiquetas, sin embargo, alrededor del 76% ha sido etiquetado solo con una.

LitCovid se actualiza diariamente con nuevos artículos relacionados con COVID-19 identificados en PubMed y categorizados en Tratamiento, Diagnóstico, Prevención e Infecciones. Inicialmente, toda la recopilación de datos y el almacenamiento de documentos se realizaban de manera manual con poca ayuda de las máquinas. Sin embargo, a medida que avanzaba la pandemia, se implementaron enfoques automatizados para dar soporte al refinado manual y maximizar la productividad de la refinación para mantener al día con la literatura en rápido crecimiento.

Los artículos se afinan o depuran a diario, permitiendo que los usuarios puedan navegar de manera rápida por el entorno de la investigación de temas acerca del COVID-19 con un alto nivel, geolocalización y organizaciones relacionadas. La información afinada integra la búsqueda entre datos y conocimiento, lo que permite el descubrimiento de conocimientos en aplicaciones posteriores, como la síntesis de pruebas y la reutilización de fármacos. Así también, permite descubrir información a través de funciones de búsqueda avanzadas como clasificación de relevancia, búsqueda de frases, entre otras.

Cabe señalar que LitCovid es una fuente de datos abierta por lo que se puede descargar libremente para la investigación, así como para tareas de procesamiento automático. La tarea de afinación o depuración de los artículos de LitCovid se realiza de la siguiente manera:

- Los artículos candidatos son seleccionados utilizando consultas de palabras clave de PubMed por medio de la herramienta E-Utils de NCBI.
- Los artículos seleccionados se examinan y clasifican como relevantes o irrelevantes.

- Los artículos relevantes de COVID-19 se afinan a profundidad.
    - Se les asigna uno o más de los ocho temas generales que correspondan.
    - Se extrae la geolocalización y las menciones de drogas o sustancias químicas en el título y el abstract.
    - Los artículos relevantes son indexados mediante Solr, una plataforma de búsqueda empresarial independiente de código abierto.

2. Análisis Exploratorio de los Datos:

El Análisis Exploratorio de Datos o EDA por sus siglas en inglés (*Exploratory Data Analysis*), permite revisar cómo están los datos antes de crear el modelo, este paso es importante ya que al realizar la inspección del conjunto de datos permite revisar qué distribución tienen sobre ciertas características, si existen datos que aporten a la construcción del modelo o que deban ser descartados, normalizados, entre otros.

Para realizar la experimentación se toma como base el conjunto de datos descritos previamente LitCovid de entrenamiento actualizada hasta el 12-09-2021 (Qingyu, et al, 2021), la misma que consta de un total de 24,960 artículos de LitCovid. Si el análisis exploratorio de los datos o EDA por sus siglas en inglés (Exploratory Data Analysis), no se realiza adecuadamente, pueden darse problemas o dificultades en las etapas o fases siguientes durante la construcción del modelo de ML. Entre los pasos que se emplean para realizar éste análisis se encuentran:

- Revisión de la cantidad de datos, lo que permite determinar si existen los suficientes recursos para el procesamiento de los mismos.
- Identificar si existen filas o columnas en blanco, ya que si estos datos son parte de la construcción del modelo podrían introducir ruido y afectar el cálculo del modelo.
- Identificar el tipo de datos, es decir, si la información a analizar comprende únicamente texto o también se componen de otro tipo de datos números como enteros, decimales, alfanuméricos, etc.
- Tener siempre claro qué tipo de tarea es la que se va a realizar, es decir, si la tarea consiste en abordar un problema supervisado, si es de salida binaria o multiclase, ya que esto permitirá seleccionar la arquitectura adecuada para la construcción del modelo.

- Visualización del corpus en una nube de palabras (representación gráfica de la frecuencia de las palabras en un texto), esta representación gráfica puede proveer una descripción general del corpus de texto, permitiendo visualizar si el texto a ser analizado contiene los temas de interés.
- Revisión de la distribución de los datos, esto permite revisar cómo se distribuyen los mismos en relación a cierta característica a lo largo del dataset.

A continuación, se muestran algunas imágenes del proceso EDA realizado sobre el conjunto de datos, compuesto de 24,960 registros.

Tabla 1. Descripción del Conjunto de Datos

#	**Columna**	**Cant. No-Null**
0	Pmid	24,960 non-null
1	Journal	24,960 non-null
2	Title	24,960 non-null
3	Abstract	24,960 non-null
4	Keywords	18,968 non-null
5	pub_type	24,960 non-null
6	Authors	24,859 non-null
7	Doi	24,406 non-null
8	Label	24,960 non-null

Como se puede observar, algunos de estos datos no están completos como el campo de “keywords”, “authors” y “doi”, sin embargo, estos datos no son representativos, ya que el presente estudio se enfoca en el análisis del abstract. Con ésta información inicial se puede empezar a trabajar sobre la tarea a desarrollarse, ya que por la inspección realizada el conjunto de datos seleccionado no

contienen vacíos o datos de tipo null en el *abstract*, sobre el cual se va a realizar la construcción del modelo de contexto o modelo de Word embeddings y sobre el que se va a realizar la clasificación.

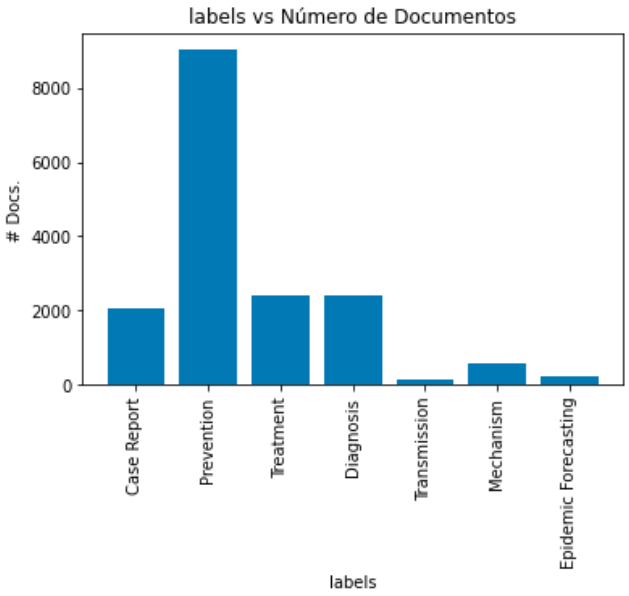
3. Pre-procesamiento:

El pre-procesamiento del texto antes de ejecutar cualquier tarea de PLN es un proceso importante, para el presente estudio se abordan las siguientes subtareas de pre-procesamiento: eliminación de stopwords o palabras vacías, eliminación de signos o caracteres especiales, normalización y Stemming. Debido a que el conjunto de datos se compone de información de varias fuentes poseen distintas características, lo que hace necesario estandarizar todas estas características de manera que el modelo que va a realizar la predicción para la clasificación de texto contenga únicamente información que sea relevante.

Es importante destacar que no existe un método estandarizado para realizar el pre-procesamiento, ya que muchos de estos procedimientos pueden utilizarse dependiendo del tipo de tarea a realizar y del texto que vaya a ser analizado, ya que podría ser el caso que, para ciertas tareas de PLN, puede requerir realizar ciertos procedimientos de pre-procesamiento y para otras tareas no.

- *Filtrado de Datos:* Para lograr el objetivo de realizar una clasificación de artículos se toma únicamente los artículos que contengan éste atributo en el conjunto de datos, ya que pueden existir diversos documentos que no contengan éste atributo, pudiendo introducir ruido al momento de realizar el análisis de texto y podría provocar una mala precisión del modelo de clasificación. El conjunto de datos está compuesto de 24,960 artículos, una vez aplicado el filtro los artículos resultantes para el análisis son 16,814.

Figura 2. Filtro de datos



- *Eliminación de caracteres especiales y puntuación:* La tarea de clasificación que se propone en el presente proyecto se basa en la aplicación de enfoque de word embedding, y en vista que estas representaciones vectoriales de texto no proporcionan representaciones para signos de puntuación y caracteres especiales, estos deben eliminarse.
- *Eliminación de enlaces o URLs:* De la misma manera que los caracteres especiales las direcciones web o urls no aportan información semántica o sintáctica para establecer relación del texto, por lo que debe eliminarse éste tipo de contenido.
- *Eliminación de stopword o palabras vacías:* El lenguaje natural está conformado de dos clases de palabras las que contienen significado asociado entre ellas y palabras funcionales que no contienen ningún significado. Las stopwords o palabras vacías, son términos utilizados para identificar palabras funcionales y no necesitan ser parte del procesamiento de tareas de PLN por su bajo aporte al análisis. Las stopwords o palabras vacías son palabras funcionales que carecen de sentido en el contexto de tareas de clasificación de texto. Estas deben ser eliminadas con el propósito de reducir el tamaño del texto y analizar palabras que únicamente aportan al contexto dentro del corpus.

- *Identificación de n-grams:* El proceso de identificación de n-grams permite identificar características dentro del documento como, por ejemplo, determinar conjuntos de palabras que ocurren con frecuencia, para el caso del presente estudio se ha definido la identificación de *unigrams*, *bigrams* y *trigrams*; por ejemplo: *Unigram*: *"coronavirus"*, *Bigrams*: *"coronavirus pandemic" y Trigrams*: *"test positive coronavirus".*

- *Tokenización:* La identificación de ngram tokenización no es más que el proceso de dividir el texto en unidades textuales más pequeñas, se puede interpretar como dividir un conjunto de información en símbolos, es decir los token o símbolos de una palabra son cada una de sus letras; de un párrafo un símbolo o token podría ser toda una oración.

4. Representación semántica de palabras:

La representación del texto debe ser capaz de mantener la similitud semántica entre las palabras que componen el texto, la representación por Word Embedding es generar vectores de manera que las palabras que sean similares entre sí semánticamente, están cerca una de las otras en el espacio vectorial. Con esto se logra que los vectores resultantes de éste modelo puedan ser utilizados como entrada para el modelo de clasificación y tengan un mejor rendimiento, al momento de realizar las predicciones de clasificación.

A pesar de que existen ya modelos pre-establecidos de vectores por Word Embeddings generalizados para tareas de PLN, para el presente estudio se realiza la construcción de un modelo de contexto propio a partir del corpus del conjunto de datos seleccionado, por lo que se obtienen tres modelos de contexto con las distintas arquitecturas antes mencionadas: Word2Vec, FastText y Glove.

Para la construcción de los mencionados modelos de contexto se tienen que establecer ciertos hiper parámetros, los cuales afectan la calidad de entrenamiento, así como la velocidad del mismo. Para los modelos de Word2Vec y FastText se han determinado los siguientes hiper parámetros para el presente estudio:

- MIN_COUNT: Este parámetro se utiliza para delimitar el número de veces que la palabra se repite dentro del corpus, éste valor por defecto es 5, sin embargo, depende mucho del tamaño del conjunto de datos para entrenar.

- SIZE: Este parámetro determina el tamaño del vector resultante que va representar cada palabra; para el presente estudio se configura con un tamaño de 300, ya que son los tamaños por defecto que maneja ésta arquitectura.
- WINDOW: La ventana o tamaño de ventana significa que la palabra del centro es la palabra objetivo y las demás son las palabras de contexto, para el presente estudio se ha considerado un valor de 5.
- SG: este parámetro indica que arquitectura de Word2Vec se utiliza, para el caso de Skip-Gram es 1 y para CBOW es 0.

En el caso del modelo Glove, se define únicamente el hiper parámetro NO_COMPONENT, el mismo que indica la dimensión que van a tener los vectores para cada palabra, lo que equivale al hiper parámetro *SIZE,* del modelo anterior. La tabla a continuación muestra la información del modelo de contexto obtenido mediante el empleo de la arquitectura Word2Vec con la arquitectura Skip-Gram.

Tabla 2. Descripción del Modelo Word Embedding - Word2Vec

Descripción de elemento	**Valor**
Número de Documentos	24,960
Tamaño del Corpus	3191,187 total words
Tamaño del Vocabulario	83,439
Tiempo de entrenamiento	564.44 segundos
Pérdidas en el entrenamiento	0.0
Épocas	30
Tamaño del vector	300
Arquitectura	Skip-Gram

Como se puede observar, se han generado vectores de palabras de trecientas dimensiones, esto quiere decir que, para cada palabra dentro del corpus, existe un vector donde cada una de sus dimensiones representa una relación que tiene ésta palabra con el resto de palabras del texto, como se muestra en la siguiente ilustración:

	1	2	3	4	5	6							299	300
pandemic	0.68	-1.5	0.8	0,765	0.0034	-0.1	0.412	...	...	...	...	...	...	0,0056	0.8329
desease	-0.3467	0.8121	-0.0012	0.0038	0.665	-0.44567	-0.0867	...	...	...	...	...	...	0.4007	-0.0451

Figura 3. Representación de Vectores de Palabras por Word Embedding

5. Generación de Secuencias:

Se transforma el corpus de texto en secuencias rellenas de identificadores de palabras para obtener una matriz de características, cabe resaltar que el relleno de las secuencias se determina con base al tamaño de la secuencia de mayor tamaño, por lo que, secuencias de menor tamaño son rellenadas con cero, hasta lograr un tamaño igual a la secuencia mayor, para el presente estudio, el artículo con mayor número de palabras dentro del abstract es de 847 palabras, por lo que las secuencias de los demás abstracts de menor tamaño serán rellenados con cero, hasta completar el tamaño mencionado.

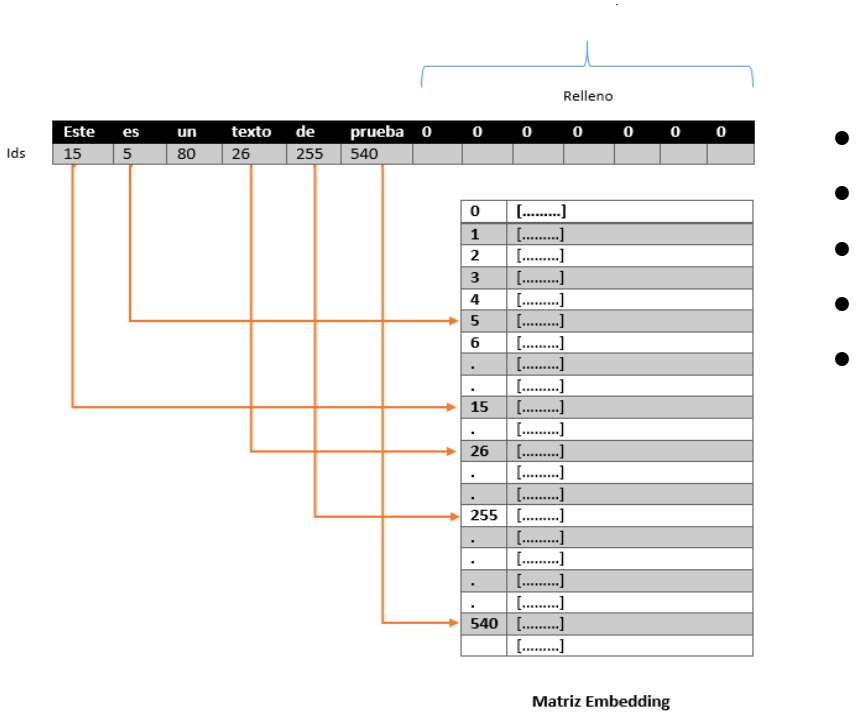
6. División de Datos de Entrenamiento y Prueba:

Una vez obtenidas las secuencias para cada uno de los abstracts del corpus, se procede a dividir el conjunto de datos de las secuencias obtenidas en datos de entrenamiento y prueba. Los artículos o datos de entrenamiento son los que aportan a la identificación de patrones en los datos, también en ésta etapa se reducen las tasas de error para la etapa de prueba y evaluación del rendimiento del modelo. Algunos estudios como Khan, et al (2010), indican que, para realizar el entrenamiento de modelos de ML, es necesario contar con un subconjunto representativo lo suficiente para evitar el sobreentrenamiento. Del conjunto de datos seleccionado, el 70% de ellos se consideran como datos de entrenamiento y el 30% como datos de prueba del modelo.

7. Creación de Matriz de Embeddings

Esta matriz actúa como una matriz de pesos, donde el vector de la palabra n se ubica en la fila n, esto se logra mediante la tokenización, que se realizó en la etapa de Pre-procesamiento, donde se obtuvo un diccionario para cada palabra dentro del corpus. Este diccionario contiene un índice de cada palabra, el mismo que sirve para ubicar cada vector de palabra en la fila correspondiente a dicho índice. La figura a continuación muestra el proceso de generación de la matriz de embeddings.

Figura 4. Matriz de Embeddings



8. Modelo de Red Neuronal

Por último, el modelo de red neuronal es creado con la matriz de pesos de cada una de las palabras en el vector de secuencias correspondiente. El objetivo es desarrollar un modelo de clasificación basado en redes neuronales basado en la arquitectura LSTM, un tipo especial de redes recurrentes; las cuales tienen un mejor desempeño al momento de procesar secuencias de datos, y predecir la salida.

La arquitectura de la red neuronal utilizada en el presente estudio consiste en una red neuronal recurrente LSTM bidireccional, la misma que consta de capas hacia adelante y hacia atrás que están conectadas juntas a la capa de salida, de esta manera, tales redes neuronales mantienen la información contextual en ambas direcciones, lo que es precisamente útil para el caso de tareas de clasificación de texto.

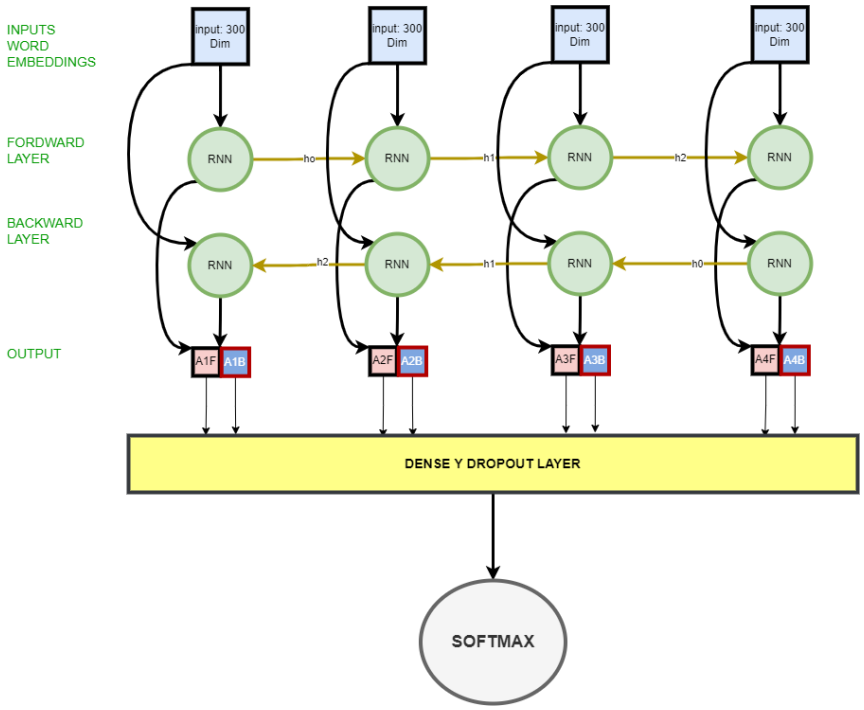
Para entenderlo de mejor manera, la celda RNN toma como valor de entrada un estado oculto o vector, y un vector de palabra, luego esta celda produce como salida el siguiente estado oculto, esta celda RNN tiene algunos pesos que se autoajustan mediante backpropagation de las pérdidas. Además, a todas las palabras se aplica la misma celda para que los pesos se compartan.

Una red neuronal RNN tradicional, para una secuencia longitud determinada proporciona el mismo número de salidas que se pueden vincular y luego esta pasarse a la capa de densidad hacia adelante. Por otra parte, la diferencia con las redes LSTM Bidireccionales es que toma la secuencia de entrada tanto en su forma inicial, así como inversa (forward y backward); se aplican dos RNN en paralelo y se obtiene una salida del doble de tamaño de la entrada. Una vez obtenida esta salida se envía a la capa de densidad para luego aplicar una función softmax y obtener el clasificador de texto. (Abduljabbar, et al, 2021).

Teniendo en cuenta este tipo de red neuronal, se ha construido el modelo de clasificación para el presente estudio de la siguiente manera:

- La capa de embedding toma las secuencias como entrada y los vectores de palabras como pesos.
- Dos capas de red neuronal LSTM Bidireccional, que tienen como objetivo modelar el orden de palabras en una secuencia en ambas direcciones.
- Dos capas finales de densidad que lo que hacen es predecir la probabilidad de cada una de las distintas categorías.
- Debido a que es un problema multiclase, se emplea una función softmax, ésta función devuelve valores entre 0 y 1, los cuales representan las probabilidades para cada categoría.

Figura 5. Modelo de Clasificación basado en una Red Neuronal LSTM Bidireccional



De la misma manera que para la construcción de los modelos de contexto o word embeddings se establecieron ciertos hiper parámetros, así también, se debe realizar para el modelo de clasificación basado en redes neuronales. La siguiente tabla muestra los hiper parámetros utilizados en el modelo neuronal del presente estudio junto con una descripción de cada uno de ellos.

Tabla 3. Hiper Parámetros del Modelo de Clasificación

Hyper Parámetro	Valor	Descripción
Neuronas en capas BiDirectional LSTM	32	Número de neuronas en cada una de las capas de la red neuronal
Número de capas	2	Número de capas ocultas de la red neuronal
Tamaño de vocabulario	83,439	Tamaño del vocabulario del corpus de texto, palabras únicas.

Tamaño de vectores	300	Tamaño del vector de cada palabra obtenido en el modelo
Dropout	0.2	Técnica para regularizar el sobreajuste en modelos de redes neuronales
Optimizador	adam	
Activación	Softmax	Función de activación brinda la probabilidad de cada clase en la salida

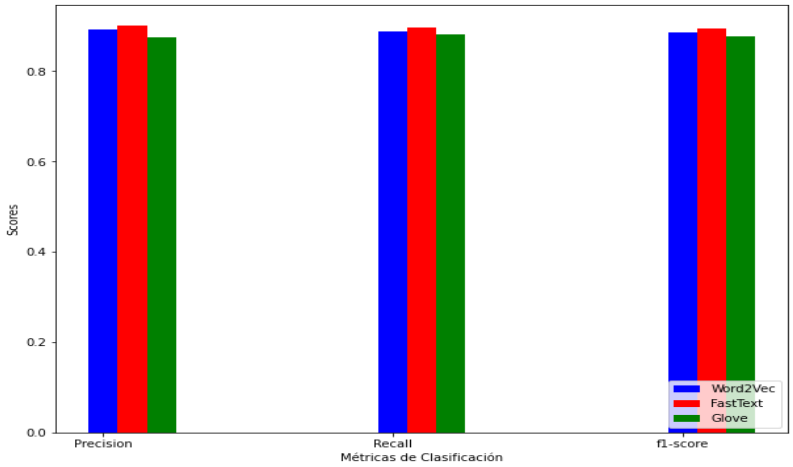
Análisis y discusión de resultados

El desarrollo de la experimentación se realiza en dos partes. La primera consiste en el análisis de los datos directamente, empleando el modelo propuesto sin considerar la distribución de los datos y aplicando la metodología propuesta. Para la segunda parte de la experimentación se considera la distribución del conjunto de datos, y en vista que los mismos tienen una distribución desbalanceada, donde existen clases muy minoritarias, las predicciones del modelo pueden verse afectadas al tener éste tipo de distribución.

Con el objetivo de afrontar éste fenómeno se aplica la técnica de muestreo estratificado, la cual consiste en dividir los datos de forma aleatoria en grupos o muestras del mismo tamaño, estos grupos o muestras son utilizados para entrenar el modelo. Se han obtenido los resultados de clasificación con base a los tres modelos propuestos (Mikolov, et al, 2013; Bojanowski, et al, 2017; Pennington, et al, 2014); para la clasificación de artículos científicos mediante el análisis de los abstracts de cada uno de éstos artículos.

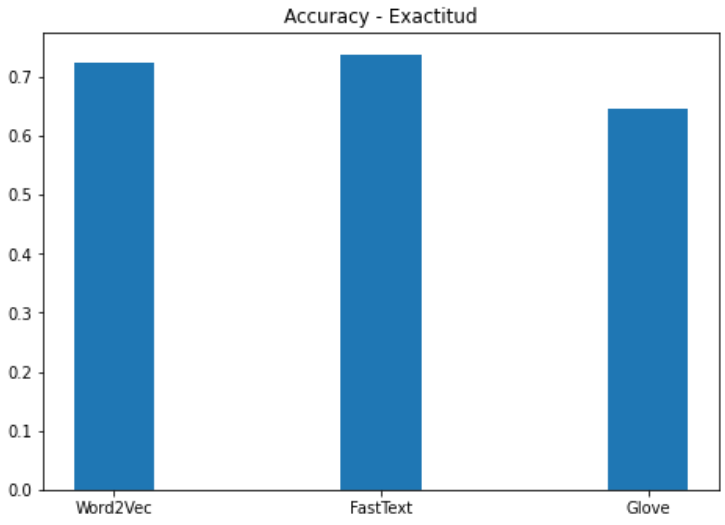
La Figura a continuación visualiza el desempeño de los tres modelos propuestos, aplicando las técnicas de evaluación de ML, las cuales, de acuerdo con Pedrero et al (2021), permiten lidiar con el problema de grandes cantidades de información que resultan difíciles de analizar, facilitando la entrega de información confiable y rápida, y la toma de decisiones. Los tres modelos presentan un desempeño de las métricas de *precision*, *recall* y *f1-score* superiores al 80%, lo que se puede entender como un rendimiento aceptable de las predicciones realizadas.

Figura 6. Métricas de Evaluación



Para determinar la exactitud o *accuracy* de cada uno de los modelos, se evalúa mediante la precisión equilibrada, que no es más que la media aritmética sobre la métrica *recall*, obtenida para cada clase, ésta medida se aplica en vista de que se tiene el conjunto de datos desbalanceado, entonces la métrica *recall*, brinda el porcentaje de clasificaciones acertadas que el modelo es capaz de realizar. La siguiente figura muestra la exactitud de cada uno de los modelos.

Figura 7. Exactitud de los modelos



Tal como se observa en la imagen anterior, la exactitud o *accuracy* de los modelos que emplean Word2Vec y FastText está entre el 72% y 74% respectivamente, mientras que en el caso del modelo empleando Glove se encuentra en el 65%. Esta evaluación de cada modelo muestra qué tan

eficientes son cada uno de ellos, sin embargo, al ser una tarea de clasificación de tipo multiclase es importante conocer cuál es el desempeño que los modelos tienen al predecir cada una de las clases. A continuación, se presentan las distintas métricas para cada uno de los modelos desarrollados.

Tabla 4. Métricas de Clase de Modelo Word2Vec

	Métricas		
Clase	precision	recall	f1-score
Case Report	0,72	0,93	0,81
Diagnosis	0,88	0,82	0,85
Epidemic Forecasting	0,45	0,83	0,58
Mechanism	0,90	0,70	0,79
Prevention	0,97	0,93	0,95
Transmission	0,00	0,00	0,00
Treatment	0,83	0,86	0,85

Tabla 5. Métricas de Clase de Modelo FastText

	Métricas		
Clase	precision	recall	f1-score
Case Report	0,81	0,86	0,83
Diagnosis	0,87	0,89	0,88
Epidemic Forecasting	0,63	0,75	0,68
Mechanism	0,73	0,84	0,78
Prevention	0,96	0,94	0,95
Transmission	1,00	0,02	0,04
Treatment	0,84	0,87	0,86

Tabla 6. Métricas de Clase de Modelo Glove

	Métricas

Clase	precision	recall	f1-score
Case Report	0,77	0,83	0,80
Diagnosis	0,83	0,85	0,84
Epidemic Forecasting	0,61	0,34	0,43
Mechanism	0,78	0,66	0,71
Prevention	0,95	0,94	0,94
Transmission	0,00	0,00	0,00
Treatment	0,81	0,90	0,85

Como se puede observar en las tablas anteriormente presentadas, los resultados de las métricas de evaluación para cada clase son similares, sin embargo, la métrica de *precision*, para el caso del modelo que emplea FastText es del 100% en el caso de la clase *Transmission*, mientras que los modelos de Word2Vec y Glove son de 0% para dicha clase. Esto se debe a que el conjunto de datos está desbalanceado y existen clases muy mayoritarias en comparación con otras, por lo que los resultados de la clasificación realizada se ven afectados por éste fenómeno. Tal como se observó en la etapa de preprocesamiento y análisis exploratorio de datos, la clase *Transmission* representa apenas el 0.79% de artículos etiquetados con esta clase, por lo que las predicciones al entrenar los modelos afectan.

Esto también puede visualizarse de mejor manera mediante las matrices de confusión de cada modelo, donde se analizan los valores reales de cada clase versus los valores predichos.

Figura 8. Matriz de confusión Modelo: Word2Vec

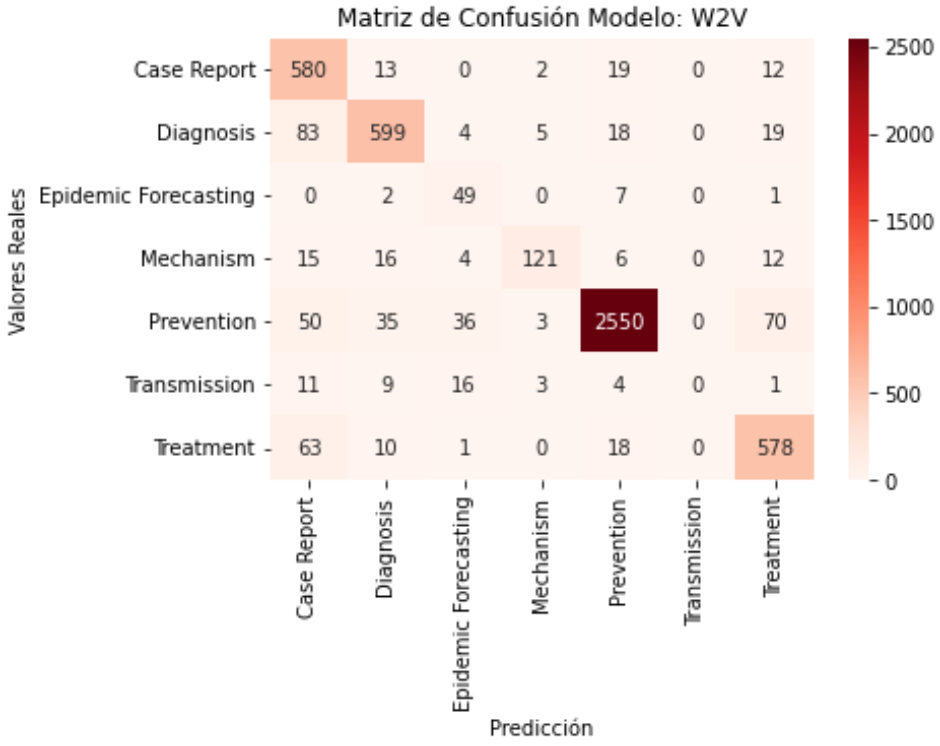


Figura 9. Matriz de confusión Modelo: FastText

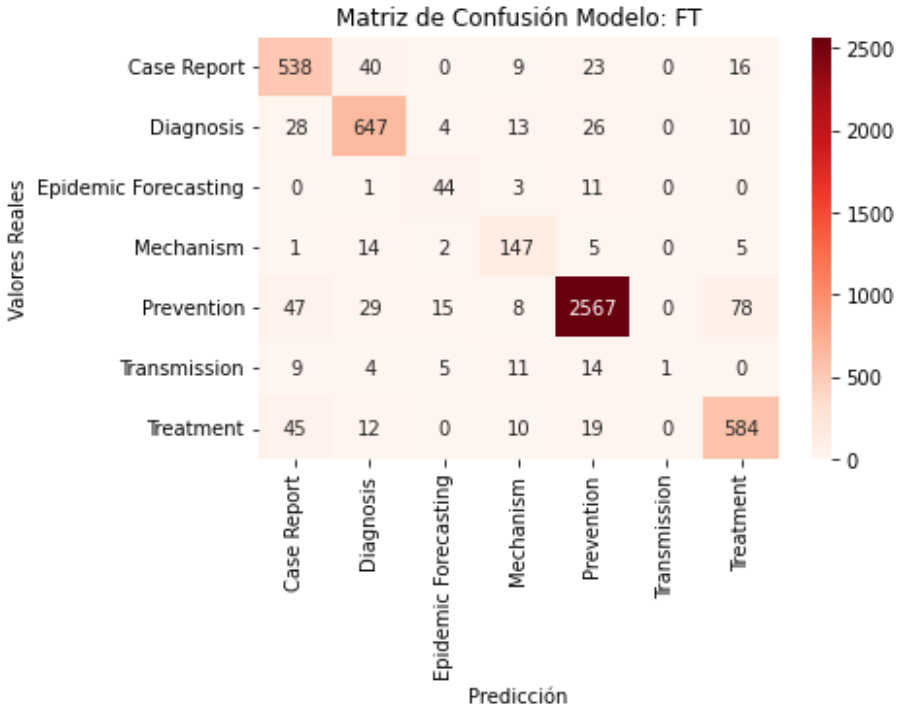
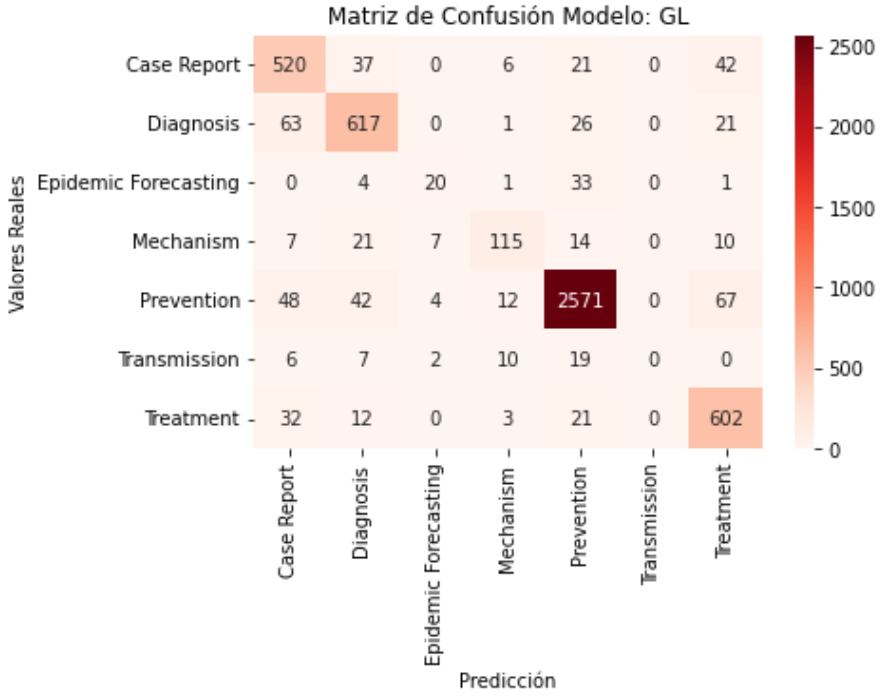


Figura 10. Matriz de confusión Modelo: Glove



Como muestran las figuras anteriores sobre las matrices de confusión de cada modelo, el mayor número de artículos que se han logrado clasificar efectivamente corresponde a la clase de *Prevention*, siendo las siguientes clases con mejor cantidad de artículos clasificados las de

Diagnosis, *Treatment* y *Case Report*, por otra parte las clases minoritarias la cantidad de artículos clasificados se ven muy afectados, tal es el caso para la clase *Transmission*, en donde el modelo que emplea FastText realiza una sola predicción correcta mientras que los otros dos modelos no realizan ninguna.

Con base en el análisis anteriormente planteado, puede observarse que el presente estudio se ha centrado en la clasificación de artículos académicos sobre la pandemia de COVID-19 a través de la técnica de minería de texto Word Embedding; algo muy útil hoy día debido a la avalancha de información publicada que ha traído consigo la pandemia. Así lo afirman Wang & Lo (2021) al mencionar que se han publicado más de cincuenta mil artículos sobre COVID-19 desde principios de 2020 y se siguen publicando varios cientos de artículos nuevos todos los días. Esta enorme tasa de productividad científica, lleva a una sobrecarga de información, dificultando que los médicos, enfermeros, bioanalistas, funcionarios de salud pública, gobiernos e investigadores, estén al día con los últimos hallazgos sobre la temática.

La metodología analizada parte del procedimiento para la clasificación multiclase, del conjunto de entrenamiento de LitCovid, que, según Jiménez, et al (2020) representa una recopilación de artículos recientemente publicados, cuyas temáticas están relacionadas con la literatura actual del Coronavirus. Este conjunto de datos contiene más de 23.000 artículos y en promedio se agregan 2.000 nuevos artículos semanalmente, siendo así un recurso integral para que la comunidad científica pueda actualizarse con información acerca de la crisis que ha provocado la pandemia de la COVID-19.

Se desarrolla la metodología a seguir para la clasificación de artículos científicos, mediante la aplicación de técnicas de Deep Learning como lo es Word Embedding, un modelo que, de acuerdo con Mikolov, et al (2013), predice palabras a partir de términos que están cercanos a éstas en función a vectores más pequeños y densos. Este tipo de métodos basan su concepto en que si se puede predecir el contexto en el cual aparece una palabra, entonces se entiende el significado de ésta en su contexto. Por lo que palabras semánticamente similares estarán cerca entre sí en sus representaciones de espacios vectoriales.

Se evaluaron tres modelos propuestos que se basan en tres arquitecturas diferentes de Word Embeddings, a saber, Word2Vec, el FastText y el Glove (Mikolov, et al, 2013; Bojanowski, et al, 2017; Pennington, et al, 2014), con la arquitectura LSTM Bidireccional. La comparación de los resultados de rendimiento obtenidos para cada modelo mostró que la exactitud o accuracy de cada

modelo se encuentran en un rango del 65% al 74%, siendo el modelo que emplea FastText el que alcanzó el mayor porcentaje de exactitud mientras que el modelo que emplea Glove alcanzó la menor exactitud de los tres.

Sin embargo, al analizar los resultados obtenidos por cada uno de los modelos se observa que no se logró una predicción con resultados favorables en todas las clases, esto debido a que los datos están desbalanceados y existen clases muy mayoritarias en comparación a otras, por lo que las predicciones se ven afectadas por estas clases.

Conclusiones

A modo de conclusión, debido a la avalancha de información visible en la web sobre COVID-19, es imprescindible la clasificación de artículos científicos publicados sobre la mencionada temática, para lo cual pueden aplicarse técnicas de Machine Learning, a través de mecanismos de representación semántica de palabras como el Word Embeddings y tecnologías basadas en redes neuronales; utilizando el análisis y procesamiento de los abstracts de artículos científicos disponibles en las fuentes de información como LitCovid.

Con la aplicación de la propuesta del modelo de clasificación, se puede concluir que al analizar los resultados obtenidos por cada uno de los modelos se observa que no se logró una predicción con resultados favorables en todas las clases, esto debido a que los datos están desbalanceados y existen clases muy mayoritarias en comparación a otras, por lo que las predicciones se ven afectadas por estas clases.

Por lo tanto, se concluye que, si bien los resultados obtenidos han demostrado que la clasificación de los artículos académicos de tipo multiclase es posible realizarla aplicando la metodología propuesta, es necesario señalar que se puede mejorar el rendimiento de los modelos aplicando otras técnicas de selección de datos para aminorar el problema que se presenta con el desbalance en la distribución de los mismos. Además, a fin de obtener una mayor calidad la representación semántica de las palabras, pudiera emplearse no solo el análisis del abstract, sino también de partes de segmentos de mayor tamaño, como, por ejemplo, la introducción u otros apartados del documento.

Referencias

1. Abduljabbar, R., Dia, H., & Tsai, P. (2021). Modelos LSTM unidireccionales y bidireccionales para la predicción del tráfico a corto plazo . Journal of Advanced Transportation , 2021(5589075). doi: https://doi.org/10.1155/2021/5589075
2. Ananiadou, S., Kell, D., & Tsujii, J. (2006). Text mining and its potential applications in systems biology. Trends Biotechnol (24), 571–579.
3. Awasthi, R., Pal, R., Singh, P., Nagori, A., Reddy, S., Gulati, A., . . . Sethi, T. (2020). CovidNLP: A Web Application for Distilling Systemic Implications of COVID-19 Pandemic with Natural Language Processing. MedRxiv.
4. BMJ Best Practice. (17 de agosto de 2020). Visión general de los coronavirus. (B. P. Group, Ed.) Obtenido de https://bestpractice.bmj.com: https://bestpractice.bmj.com/topics/eses/3000165/
5. Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. Transactions of the Association for Computational Linguistics, 5, 135-146. doi:https://doi.org/10.1162/tacl_a_00051
6. Chandrasekaran, B., & Fernandes, S. (january de 2020). Target specific mining of COVID-19 scholarly articles using one-class approach. Diabetes Metab Syndr, 14(4), 337–339.
7. Cohen, A., & Hersh, W. (2005). A survey of current work in biomedical text mining. Brief Bioinform(6), 57-71.
8. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. J. Mach. Learn. Res.(12), 2493–2537.
9. Daud, A., Khan, W., & Che, D. (2017). Urdu language processing: a survey. . Artificial Intelligence Review, 47(3), 279–311. doi:https://doi.org/10.1007/s10462-016-9482-x
10. Friedman, C., & Johnson, S. (2006). Natural language and text processing in biomedicine. . En E. Shortliffe, & J. Cimino, Biomedical informatics: computer applications in health care and biomedicine (Third ed., págs. 312 - 343). New York: Springer.
11. Greenhalgh, T., Choon, G., & Koh, H. (2020). Covid-19: una evaluación remota en atención primaria. Practice(368:m1182), 1-5. doi:doi: 10.1136/bmj.m1182
12. Jiménez, B., Zeng, J., Zhang, D., Zhang, P., & Su, Y. (2020). Clasificación de documentos para la literatura COVID-19. En A. d. Computacional (Ed.), En Hallazgos de la Asociación de Lingüística Computacional: : EMNLP 2020 (págs. 3715-3722). doi:10.18653/v1/2020.hallazgos-emnlp.332

13. Jurafsky, D., & Martin, J. (2020). Speech and Language Processing: An Introduction to Natural Language Processing Computational Linguistics and Speech Recognition (Third Edition ed.).
14. Khan, A., Baharudin, B., Hong, L., & khan, K. (2010). A Review of Machine Learning Algorithms for Text-Documents Classification. ournal of Advances in Information Technology, 1(1). doi:https://doi.org/10.4304/jait.1.1.4-20
15. Kilicoglu, H. (2018). Biomedical text mining for research rigor and integrity: tasks, challenges, directions. Brief Bioinform (19), 1400-1414.
16. Maguiña, C., Gastelo, R., & Tequen, A. (2020). El nuevo Coronavirus y la pandemia del Covid-19. Revista Medica Herediana, 31(2), 125-131. Obtenido de https://doi.org/10.20453/rmh.v31i2.3776
17. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (16 de January de 2013). Efficient estimation of word representations in vector space. En C. University (Ed.), 1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings. Obtenido de https://arxiv.org/abs/1301.3
18. Ministerio de Sanidad. (2020). Neumonía por nuevo coronavirus (2019-nCoV) en Wuhan, provincia de Hubei, (China). Informe Actualización nº 13, Ministerio de Sanidad, Madrid. Obtenido de https://www.mscbs.gob.es/profesionales/saludPublica/ccayes/alertasActual/nCov-China/documentos/Actualizacion_13_2019-nCoV_China.pdf
19. Muhammad, A., Suliman, K., Abeer, K., Nadia, B., & Rabeea, S. (2020). COVID-19 infection: Emergence, transmission, and characteristics of human coronaviruses. Journal ofAdvanced Research(24), 91-98.
20. Organización Panamericana de la Salud. (2022). Artículos científicos y recursos sobre la COVID-19. Revista Panamericana de Salud Pública(Números Especiales).
21. Pedrero, V., Reynaldos-Grandón, K., Ureta-Achurra, J., & Cortez-Pinto, E. (2021). Generalidades del Machine Learning y su aplicación en la gestión sanitaria en Servicios de Urgencia. Revista médica de Chile, 149(2), 248-254. doi:https://dx.doi.org/10.4067/s0034-98872021000200248
22. Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global Vectors for Word Representation. Actas de la Conferencia de 2014 sobre métodos empíricos en el

procesamiento del lenguaje natural (EMNLP) (págs. 1532-1543). Doha, Qatar: Asociación de Lingüística Computacional.

23. Qingyu, C., Alexis, A., & Zhiyong, L. (2021). LitCovid: una base de datos abierta de literatura sobre COVID-19. Nucleic Acids Research, 49(D1), D1534-D1540. doi:https://doi.org/10.1093/nar /gkaa952
24. Sonbhadra, S., Agarwal, S., & Nagabhushan, P. (2020). Apunte a la extracción específica de artículos académicos sobre el COVID-19 utilizando un enfoque de clase única. Caos, solitones y fractales(140 , 110155). Obtenido de https://doi.org/10.1016/j.chaos.2020.110155
25. Thompson, L. (2003). Inicio de una nueva epidemia, SARS. Rev Med Hered, 14(2), 49.
26. Torres-Salinas, D. (2020). Ritmo de crecimiento diario de la producción científica sobre Covid-19. Análisis en bases de datos y repositorios en acceso abierto. El profesional de la informacion(29:e290215). doi:10.3145/epi.2020.mar.15
27. Wang, L., & Lo, K. (2021). Text mining approaches for dealing with the rapidly expanding literature on COVID-19 . Briefings in Bioinformatics, 22(2), 781–799. Obtenido de https://doi.org/10.1093/bib/bbaa296
28. Zou, W., Socher, R., Cer, D., & Manning, C. (2013). Bilingual word embeddings for phrase-based machine translation. EMNLP, 1393 - 1398.
29. Zweigenbaum, P., Demner-Fushman, D., Yu, H., & al, e. (2007). Frontiers of biomedical text mining: current progress. Brief Bioinform (8), 358-375.