Artículo original

Caracterización de un *corpus* extraído de historias clínicas electrónicas de maternas a través de técnicas de procesamiento de lenguaje natural

Characterization of a *corpus* extracted from maternal electronic health records through natural language processing techniques

María Camila Durango Barrera[1*] https://orcid.org/0000-0002-2779-2461
Ever Augusto Torres Silva[1] https://orcid.org/0000-0002-6302-6131
José Fernando Florez-Arango[2] https://orcid.org/0000-0001-9083-0195
Andrés Orozgo-Duque[1] https://orcid.org/0000-0001-8582-8015

[1]Instituto Tecnológico Metropolitano. Medellín, Colombia.
[2] Universidad A&M. Texas, Estados Unidos de América.

*Autor para la correspondencia: camiladurangob@gmail.com

RESUMEN

Este artículo tuvo como propósito caracterizar el texto libre disponible en una historia clínica electrónica de una institución orientada a la atención de pacientes en embarazo. La historia clínica electrónica, más que ser un repositorio de datos, se ha convertido en un sistema de soporte a la toma de decisiones clínicas. Sin embargo, debido al alto volumen de información y a que parte de la información clave de las historias clínicas electrónicas está en forma de texto libre, utilizar todo el potencial que ofrece la información de la historia clínica electrónica para mejorar la toma de decisiones clínicas requiere el apoyo de métodos de minería de texto y procesamiento de lenguaje natural. Particularmente, en el área de Ginecología y Obstetricia, la implementación de métodos del procesamiento de lenguaje natural podría ayudar a agilizar la identificación de factores asociados al

riesgo materno. A pesar de esto, en la literatura no se registran trabajos que integren técnicas de procesamiento de lenguaje natural en las historias clínicas electrónicas asociadas al seguimiento materno en idioma español. En este trabajo se obtuvieron 659 789 tokens mediante los métodos de minería de texto, un diccionario con palabras únicas dado por 7 334 tokens y se estudiaron los n-grams más frecuentes. Se generó una caracterización con una arquitectura de red neuronal CBOW (*continuos bag of words*) para la incrustación de palabras. Utilizando algoritmos de *clustering* se obtuvo evidencia que indica que palabras cercanas en el espacio de incrustación de 300 dimensiones pueden llegar a representar asociaciones referentes a tipos de pacientes, o agrupar palabras similares, incluyendo palabras escritas con errores ortográficos. El *corpus* generado y los resultados encontrados sientan las bases para trabajos futuros en la detección de entidades (síntomas, signos, diagnósticos, tratamientos), la corrección de errores ortográficos y las relaciones semánticas entre palabras para generar resúmenes de historias clínicas o asistir el seguimiento de las maternas mediante la revisión automatizada de la historia clínica electrónica.

Palabras clave: Procesamiento de lenguaje natural; historia clínica electrónica; aprendizaje de máquina; *word embedding*; redes neuronales artificiales.

ABSTRACT

The purpose of this article was to characterize the free text available in an electronic health record of an institution, directed at the care of patients in pregnancy. More than being a data repository, the electronic health record (HCE) has become a clinical decision support system (CDSS). However, due to the high volume of information, as some of the key information in EHR is in free text form, using the full potential that EHR information offers to improve clinical decision-making requires the support of methods of text mining and natural language processing (PLN). Particularly in the area of gynecology and obstetrics, the implementation of PLN methods could help speed up the identification of factors associated with maternal risk. Despite this, in the literature there are no papers that integrate PLN techniques in EHR associated with maternal follow-up in

Spanish. Taking into account this knowledge gap, in this work a corpus was generated and characterized from the EHRs of a gynecology and obstetrics service characterized by treating high-risk maternal patients. PLN and text mining methods were implemented on the data, obtaining 659 789 tokens and a dictionary with unique words given by 7 334 tokens. The characterization of the data was developed from the identification of the most frequent words and n-grams and a vector representation of embedding words in a 300-dimensional space was performed using a CBOW (Continuous Bag Of Words) neural network architecture. The embedding of words allowed to verify by means of Clustering algorithms, that the words associated to the same group can come to represent associations referring to types of patients, or group similar words, including words written with spelling errors. The corpus generated and the results found lay the foundations for future work in the detection of entities (symptoms, signs, diagnoses, treatments), correction of spelling errors and semantic relationships between words to generate summaries of medical records or assist the follow-up of mothers through the automated review of the electronic health record.

Key words: Natural language processing; electronic health record; machine learning; word embedding; artificial neural networks.

Recibido: 16/11/2020
Aceptado: 03/02/2021

Introducción

La historia clínica electrónica (HCE) es un elemento fundamental en las instituciones de salud alrededor del mundo, en la cual se diligencia la información relacionada de todos los encuentros de salud con el paciente. En cada visita se registra un conjunto de entidades y conceptos médicos, tales como los signos vitales, las notas médicas, los síntomas, las enfermedades, entre otros. En los

últimos años ha habido un progreso significativo en el desarrollo de sistemas de información en salud, como la historia clínica electrónica. Los avances en los recursos computacionales han permitido que estos sistemas pasen de ser un repositorio de datos a tener capacidades de sistemas de soporte a la toma de decisiones clínicas (CDSS) aprovechando la gran cantidad de datos disponibles. El área del procesamiento de lenguaje natural y aprendizaje de máquina ha sacado provecho para generar soluciones de diagnóstico y tratamiento y ha contribuido a la calidad de la atención de los pacientes.[1]

Un alto porcentaje de la información contenida sobre estos registros electrónicos se encuentra de forma no estructurada o en texto libre.[1,2] El lenguaje natural facilita la comunicación entre el personal de salud y contiene información que frecuentemente no es bien representada en formas estructuradas[3] y que no es una tarea fácil para la extracción y análisis por el personal asistencial.[4] No obstante, aún siguen siendo preferidos los CDSS basados en datos estructurados dada su interpretabilidad para los expertos del dominio.[5] Es por eso que en la literatura se ha reportado el análisis automático de datos en el entorno clínico por medio de la minería de texto, con la aplicación de técnicas de procesamiento de lenguaje natural (PLN) para la extracción de información en documentos de texto de forma automática,[6] lo que permite disponer de insumos para el análisis clínico predictivo en los CDSS.[7] Entre las tareas del PLN se ha destacado el reconocimiento de entidades nombradas (NER),[8] basado en un método supervisado que contribuye al análisis semántico y a la extracción de las relaciones entre las entidades médicas. Esta tarea depende totalmente de las etiquetas, las cuales conllevan una revisión manual de las HCE que hace el proceso costoso y de mucho tiempo. Para esta tarea se requiere representar el texto a partir de modelos de *word embedding*, lo que permite extraer la relación semántica y sintáctica entre las palabras que componen el *corpus*.[9]

Dado que son limitados los estudios de PLN en HCE, y más aún en el área de Ginecología y Obstetricia, este artículo tuvo como propósito caracterizar el texto

libre disponible en una HCE de una institución orientada a la atención de pacientes en embarazo. Para esto se empleó una metodología a través del conteo de palabras y n-grams, además de la incrustación de palabras y la caracterización vectorial mediante el método conocido como *word embedding*, a partir del modelo de la bolsa continua de palabras (*continuous bag of words*- CBOW), el cual permite representar la semántica de los datos contenidos en el *corpus* de notas clínicas y obtener una caracterización de los datos a partir del aprendizaje no supervisado y sin la necesidad de etiquetar previamente los datos.[9] Posteriormente se aplicarán diferentes métodos de evaluación intrínseca, como el subconjunto de palabras sobre el espacio de las incrustaciones, la visualización en un espacio dimensional reducido en 2D y 3D y los algoritmos de agrupamiento *clustering* jerárquico aglomerativo y *clustering K-means*. Por último, se aplicarán métodos de validación interna en la elección del número de clústeres.[10]

La metodología propuesta apunta a crear las bases para el posterior desarrollo de CDSS que faciliten el reconocimiento temprano de factores que puedan desencadenar patologías de gravedad, y que contribuyan a la disminución de la mortalidad materna.[11] Esta aproximación constituye un aporte en esta área del conocimiento, dado que es limitada la evidencia de estudios relacionados con la caracterización vectorial de un *corpus* de notas clínicas sin recurrir a etiquetas asociadas al alto riesgo obstétrico.

Métodos

En este artículo se realizó una metodología que permitió el procesamiento y la caracterización del texto libre sobre un *corpus* de notas clínicas asociadas al alto riesgo obstétrico. El esquema metodológico general que se presenta en la figura 1 incluye la extracción de los datos de la HCE para la generación del *corpus*; el preprocesamiento de los datos con técnicas de minería de texto y PLN; la generación de la representación para la incrustación de palabras, que permitió capturar la relación semántica entre las palabras según el contexto; y finalmente

una comprobación sobre la caracterización a través de los métodos de evaluación intrínseca para determinar la relación entre la similitud de las palabras a partir de métodos como: subconjunto de palabras con visualización en 2D y 3D y algoritmos de agrupamiento *clustering* tales como *clustering* jerárquico aglomerativo y el *clustering K-means*, junto con un método de validación interna. El comité de ética de investigación de la Institución Bolivariana aprobó en el Acta 02 del 10 de febrero de 2020 el uso retrospectivo de los datos de la Historia Clínica Electrónica.

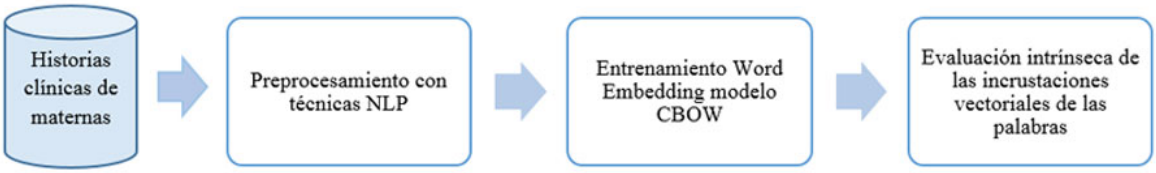

Fig. 1 - Cómo entrenar las incrustaciones de palabras en *corpus* con notas clínicas y usarlas para la automatización de tareas en el procesamiento de lenguaje natural.

Generación del *corpus*

Como primera etapa se hizo una extracción de la información de la HCE de la Clínica Universitaria Bolivariana, una institución de alto nivel de complejidad, orientado a la población materno-infantil, con alrededor de 7 000 partos anuales con base en la ciudad de Medellín, Colombia. Seguidamente, se generó el *corpus*, donde se extrajo notas del cuadro de evoluciones médica a maternas que tuvieron la terminación del embarazo en el mes de enero del año 2015. Todos los datos fueron anonimizados y los datos personales removidos del banco de datos.

Se hizo uso del software Oracale ® SQL Developer a través de sentencias SQL, donde se identificaron las maternas y las tablas que contenían el texto libre, y se exportaron en archivos planos.

Las HCE de consulta permite al personal de salud tomar notas en texto libre o de forma semiestructurada, en plantillas predefinidas. Dado que cada una de las

plantillas corresponden a diferentes eventos clínicos, solo se tomaron los datos provenientes de las tablas de evoluciones médicas para la construcción del *corpus*.

Preprocesamiento del texto

Como segunda etapa, se realizó un preprocesamiento sobre los datos para generar un modelo de representación de PLN, incluyendo tokenización, *stopwords* para el idioma español, eliminación de los signos de puntuación como caracteres alfanuméricos, espacios vacíos, tabulación, filtrado sobre las cadenas de caracteres mayores a tres en longitud y normalización sobre cada uno de los tokens. En esta etapa también se incluyó el análisis y el conteo de las palabras más usuales junto con los n-grams, los cuales consisten en secuencias de n-tokens; se obtuvo Bigrams y Trigrams con su respectiva representación gráfica. Luego, a cada token único se le asignó un índice entero irrepetible que permitió mapear y tener una representación de cada uno de estos tokens en un diccionario con una longitud de V de acuerdo con la cantidad de estos tokens en un espacio vectorial dado.[7]

Finalmente, se realizó una representación tipo *one-hot vector,* donde cada palabra se representa con un vector y todos los valores son cero excepto la posición de la determinada palabra según su ubicación de acuerdo con el valor del índice entero. Todas las etapas en el desarrollo de este trabajo se realizaron en Spyder un IDE en el lenguaje Python.

Entrenamiento del modelo CBOW (Continuous Bag of Words)

Como tercera etapa se desarrolló el entrenamiento a través del modelo CBOW de *Word Embedding*, el cual permitió la extracción de la incrustación de texto dado por el *corpus*. *Word Embedding* es una técnica de incrustación de palabras usada en el aprendizaje de máquina en la construcción de la representación vectorial de las palabras, debido a que captura las relaciones semánticas y sintácticas entre las palabras que componen a un *corpus*.[12] Entre las arquitecturas existentes de modelos que han sido implementadas para esta representación en PLN están el Skip-gram (SG) y el CBOW. Tanto el modelo SG como el CBOW utilizan redes

neuronales para aprender el mapeo de palabras a un punto en un espacio vectorial. El primero se basa en la predicción de las palabras de contexto a partir de la palabra central y el segundo hace predicciones de la palabra central a partir de las palabras de contexto.[13] Estos modelos poseen dos hiperparámetros claves importantes para el entrenamiento. Uno de ellos es el número de dimensiones de incrustación N, en el cual cada dimensión captura algún componente de la relación semántica y sintáctica de las palabras. Se ha establecido que este parámetro puede variar entre 50 a 500; y un tamaño en la ventana de embebimiento que generalmente se encuentra entre 2 a 10 representado por C.

La figura 2 muestra el esquema del modelo utilizado en esta propuesta, donde V está dado por el diccionario, el cual está representado por las palabras únicas en el *corpus* previamente preprocesadas, m representa al set de datos del conjunto de entrenamiento. N está dado por el número de dimensiones o incrustaciones a representar.

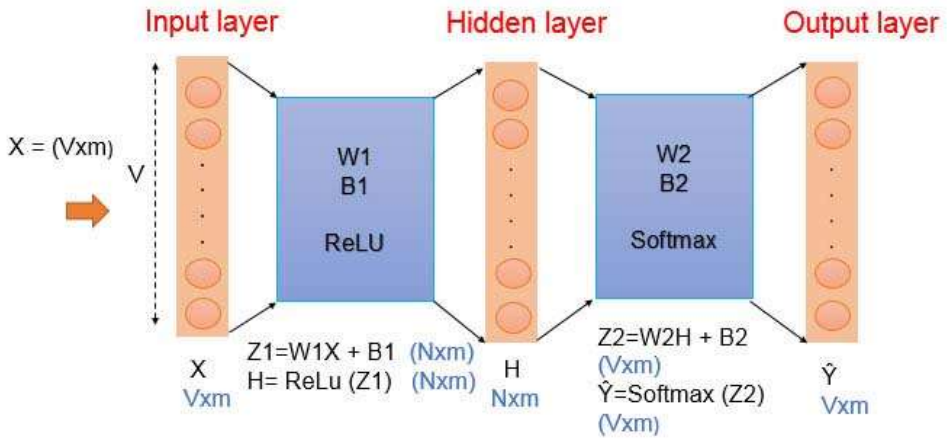

Fig. 2 - Esquema del modelo CBOW.

Para el desarrollo del contexto y las palabras activas que representaron el set del conjunto de entrenamiento se creó un vector que representó a forma del contexto todas las palabras que componen el diccionario con los tokens únicos que fueron previamente procesados a partir de la codificación y representación *one-hot-vector* de acuerdo con el tamaño de la ventana de embebimiento, el cual indicó

el número de palabras antes de la palabra central y el número de palabras después de la palabra central con una C igual a 2. Posteriormente, se hizo el promedio de la ventana de acuerdo con la frecuencia de las palabras a lo largo de esta. Estas representaciones son importantes para el entrenamiento del modelo debido a que las palabras que aparecen en los mismos contextos deben tener un significado similar o relacionado.[13]

Para realizar el proceso de entrenamiento del modelo CBOW a través de la red neuronal, primeramente se establecieron los parámetros con valores iniciales de los pesos representados comúnmente con la letra W y los bias B. La capa de entrada está compuesta por la longitud del diccionario, el cual se basó en V tokens únicos, mientras que m se basó en el set de conjunto de entrenamiento dado por las representaciones de las palabras de contexto previamente con la elección en el tamaño en la ventana de embebimiento C a través la codificación *one-hot-vector* y su respectivo promedio. Seguido de la elección del número de neuronas N en la capa oculta correspondiente a la captura de la semántica y sintáctica de las palabras contenidas en V.

Finalmente, en la capa de salida se obtuvo un arreglo que contuvo a las palabras centrales de acuerdo con la longitud de V y el número de m; la función de coste a la salida del modelo estuvo relacionada directamente sobre la elección de los parámetros e indicó qué tan bien hizo la predicción la red neuronal con base en la observación que se le dio con las palabras de contexto y el resultado esperado sobre la palabra central, determinando el valor del error. Se recurrió a la función de entropía cruzada binaria por su fácil interpretación y manejo.

Se implementó el gradiente descendente como un optimizador de la función de coste para hallar el mínimo global de dicha función y optimizar los parámetros iniciales del modelo con base en una tasa de aprendizaje (Alpha) entre valores de 0,01 y 0,04. Mientras que en los valores de los hiperparámetros también se optimizaron, ya que la función de coste también depende de su elección. Para

esto, se realizaron iteraciones con el algoritmo con valores de N de 100,150, 200, 250, 350 y 400, mientras que para C se realizó con valores de 3,4,5, 7 y 8.

Una vez que se ha entrenado el modelo CBOW, se hace extracción de la compactación de los datos por la capa oculta sobre el conjunto de los datos de entrenamiento que da lugar a la incrustación de palabras. Esta extracción de la incrustación de palabras permitió obtener la representación vectorial de las palabras con N dimensiones de acuerdo con la relación semántica de acuerdo con el *corpus* utilizado. Esta relación semántica dependerá totalmente de los datos del *corpus* con el que se esté trabajando, ya que de este van a derivar las incrustaciones.

Evaluación intrínseca de las incrustaciones vectoriales de las palabras

En la última etapa se realizó la evaluación intrínseca sobre la incrustación de palabras de forma cualitativa y cuantitativa, que permitió el análisis visual de la incrustación de palabras a través de gráficas en relación con un subconjunto de palabras sobre el espacio de la incrustación de palabras extraído. Para la evaluación cuantitativa, se realizaron métodos a partir de algoritmos *clustering* jerárquico aglomerativo y *K-means*,[14,15] junto con medidas de similitud entre la agrupación de palabras a partir de la distancia Euclidiana.

La evaluación intrínseca permitió evaluar los resultados obtenidos de acuerdo con las relaciones semánticas o sintácticas entre las palabras de contexto. Esta etapa tuvo como finalidad estudiar el significado atribuible a expresiones de palabras, dada una dimensión vectorial reducida que se hizo a partir de la implementación de vecinos estocásticos con distribución (t-SNE).[16]

El *testing* se hizo mediante el uso de las de algoritmos de *clustering*, analogías y visualización. Las técnicas sobre *clustering* que se efectuaron en este estudio se basaron en técnicas de *cluster* aglomerativo jerárquico y *K-means* junto con

técnicas de validación interna como el índice de Elbow (SSE).[10] Los algoritmos de agrupamiento permitieron agrupar el conjunto de datos según la similitud entre sí. Cada instancia de los datos estuvo representada vectorialmente por medio de las características; en este caso por la dimensionalidad de la incrustación de palabras, es decir, las 300 dimensiones. Se utilizó la distancia euclidiana como la medida de similitud según los agrupamientos por parte del algoritmo de agrupamiento *K-means*.

Resultados y discusión

A partir de las HCE asociadas a las tablas de evoluciones médicas de 597 maternas de interés que tuvieron la terminación del embarazo en la Institución en el año 2015, se obtuvo un *corpus* con un total de 2 542 655 de caracteres, 659 789 tokens sin procesar, y después de una etapa de preprocesamiento se obtuvieron 189 292 tokens que dieron lugar a un diccionario único con una longitud V con 7 334 tokens. Algunos ejemplos de frases obtenidas en el *corpus* se presentan a continuación:

> "Lactancia +, sangrado vaginal escaso, muy difícil la extracción del feto por gran contracción del útero que requirió de nitroglicerina, niega cefalea, niega prurito, niega náuseas o vómito, adecuada diuresis espontánea".

Se realizó un análisis de los tokens más frecuentes junto con los n-grams, los cuales constan de una secuencia de n tokens, se obtuvo con n igual a 2 y 3, es decir, Bigrams y Trigrams, respectivamente, como se muestra en la figura 3. Entre las palabras de mayor ocurrencia, se encuentra la palabra "niega", lo que indica que el análisis de negaciones es importante para dar claridad si determinado síntoma está asociado a la paciente o es una negación.

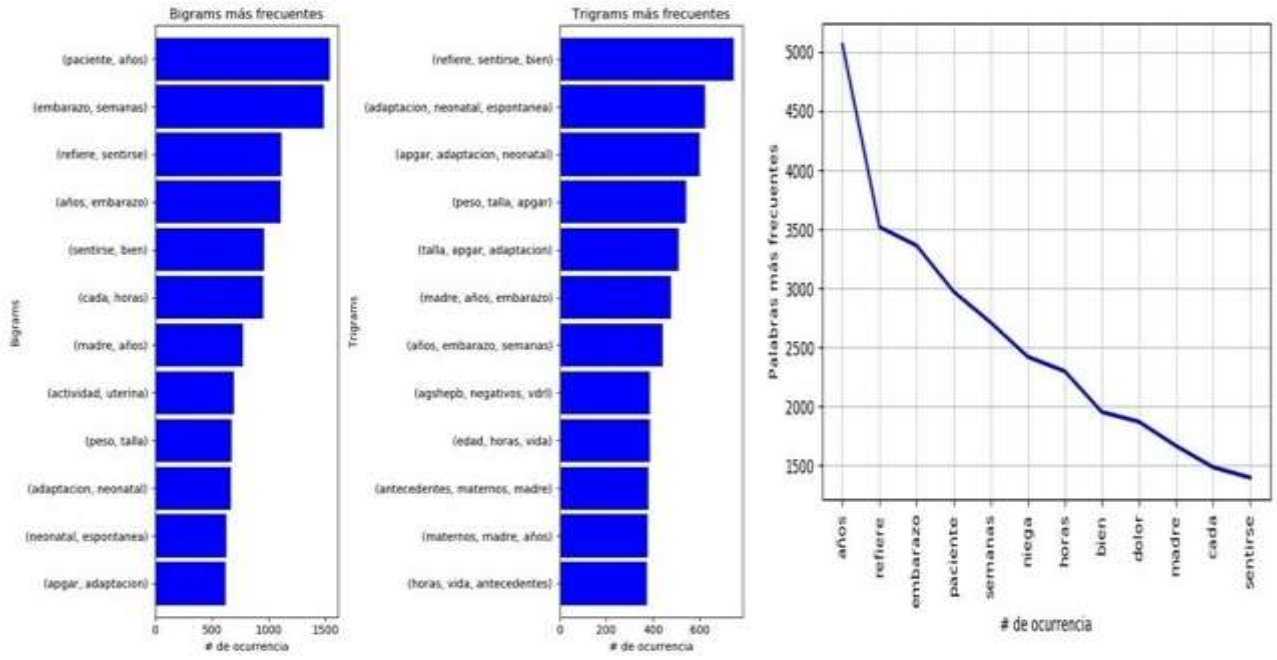

Fig. 3 - Conteo de los Bigrams y Trigrams más frecuentes con su respectiva frecuencia a lo largo del *corpus* de notas clínicas.

El Bigram más frecuente [paciente, años] se explica por la tradición de identificación de las notas de evolución como: "paciente de x años …"; es útil para la identificación de edades de riesgo, como el embarazo adolescente, las primigestantes mayores de 35 años o las mujeres en edad avanzada (mayor de 40 años) para una gestación. El siguiente Bigram [embarazo, semana] representa también elementos de la identificación de las notas de evolución que establece la duración actual del embarazo. [Refiere, sentirse] es una estructura básica del motivo de consulta y la enfermedad actual, donde se pueden capturar síntomas y signos de importancia en el análisis de riesgo.

Analizando los Trigrams [refiere, sentirse, bien] habla del estado de evolución de las pacientes, que podría utilizarse en análisis inverso, enfocando el procesamiento en esas maternas sin este trigrama. Los siguientes trigramas (que incluyen apgar) hacen referencia a la evaluación del neonato, que indica que esa nota es una nota posterminación del embarazo, probablemente de seguimiento, que aún puede ser

importante para la identificación de causas de muerte materna como la sepsis o la hemorragia posparto. Igualmente, el análisis de Bigrams y Trigrams puede ser una primera aproximación para el análisis de las palabras en un contexto. Es de anotar que es importante diferenciar cuándo el contexto se refiere a la madre y cuándo al feto o al bebé.

Tras la finalización del entrenamiento del modelo CBOW se realizó la extracción de la incrustación de palabras. El modelo estuvo compuesto por un arreglo de 7 334 tokens únicos del *corpus* y 300 dimensiones dadas por el tamaño de la capa oculta del modelo de la red neuronal. Dentro de los métodos de evaluación intrínseca se desarrolló el algoritmo de *clustering* sobre las incrustaciones de palabras. En primera instancia se hizo a partir del *clustering* jerárquico aglomerativo sin ninguna reducción de la dimensionalidad sobre las incrustaciones.

Tabla 1 - Algunos ejemplos de las palabras agrupadas por el *clustering* aglomerativo con un número de *clusters* de 7 en 300 dimensiones

Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
Palabra	**Palabra**	**Palabra**	**Palabra**	**Palabra**	**Palabra**	**Palabra**
ecocardiografía	analgesicos	alteración	paciente	ambulatorias	biamniótico	hipertensión
ecografía	anestesiólogo	alteraciones	-	ambulatorio	biliares	hipertensiva
ectopico	anestésicas	alterada	-	ambulatorios	bilioso	hipocromica
ectópico	anti	alterado	-	prenatal	cirugia	hipocrómica
espontáneamente	antibiograma	antibiotico	-	preparto	cirugias	hipotenia
esporadicamente	anticoagulación	antibioticos	-	prerrenal	cirujana	hipotiroidea
esporadicos	anticoagulante	antibiótico	-	neural	cirujano	hipotensión
esporádica	anticoagulada	fétido	-	neurológico	saturacion	hipotonico
fetal	antiembólicas	fétida	-	neuropatía	saturaciones	hipertensión

En la tabla 1 se presentan algunos ejemplos de palabras agrupadas por el algoritmo de *clustering* aglomerativo jerárquico. En la elección para el número de *cluster* se realizó un previo análisis a partir de la representación gráfica del dendograma,[17] donde se determinó un número de 7 clusters. Dentro de los clusters se encuentran agrupaciones de ciertas palabras en relación con su escritura, a pesar de la presencia de los errores ortográficos y con la derivación de la palabra base o palabra raíz , por ejemplo, con la palabra base del cluster 6 “hipo”. El *cluster* contiene agrupadas palabras como “hipotenia”, “hipotiroidea”, “hipotensión” e “hipotónico”. A partir de esto, se infiere que un correcto proceso de incrustación de palabras puede ser útil como paso previo al desarrollo de sistemas corrección de ortografía específicamente adaptado al contexto en que fue entrenado el modelo; en este caso, en un servicio de Ginecología y Obstetricia.

Por otro lado, los *clusters* encontrados tienen cierta correlación clínica que amerita un análisis más detallado, pero es claro cómo la ecografía en las palabras agrupadas dentro del *cluster* 0 tiene alta relación con los movimientos espontáneos y esporádicos del feto, y su utilidad como herramienta diagnóstica en el embarazo ectópico. También un hallazgo importante es cómo las faltas de ortografía se pueden agrupar. Aquí existe un potencial para la conversión sintáctica en la semántica de los contenidos. El *cluster* 1 parece referirse a pacientes en este posoperatorio, probablemente poscesárea, que requiere reposo y control del trombo embolismo pulmonar. El *cluster* 2 está relacionado con las infecciones, especialmente por la característica fétida de las secreciones y la presencia de antibióticos.

Posteriormente, con el fin de determinar la relación entre palabras cercanas entre sí en el espacio de embebimiento, se obtuvo la distancia euclidiana como medida de similitud entre una palabra utilizada como referencia respecto a las demás palabras del mismo cluster. Se escogieron tres clústeres con una palabra de referencia que dio como resultado la selección de las primeras cinco palabras más cercanas a ella, como se muestra en la Tabla 2.

Tabla 2 - Distancia euclidiana a partir del *clustering* aglomerativo en 300 dimensiones

Cluster 0 (palabra= fetal)		Cluster 2 (palabra=antibiótica)		Cluster 6 (palabra=hipertensión)	
Palabra	**Distancia**	**Palabra**	**Distancia**	**Palabra**	**Distancia**
cama	0,0436447	coco	0,0436144	premonitorios	0,0426295
madre	0,043973	glucómetros	0,0443347	bilaterales	0,0430633
enzimático	0,0443119	mareada	0,0448449	intestinal	0,0438262
disminuida	0,0445143	extremo	0,0451252	hipotonía	0,0442704
orales	0,0449239	colocarle	0,0455021	obstrucción	0,0443133

Se esperaba obtener como palabras cercanas, dado la palabra de referencia, aquellas palabras derivadas en su escritura, como se mostraba en la agrupación de palabras dentro de los *clusters* en la Tabla 1. Sin embargo, el contexto asociado a estas palabras podría estar dado por el contexto clínico en torno a las complicaciones y síntomas durante el embarazo (Tabla 2). Para el *cluster* 2, la palabra más cercana a antibiótica fue coco, la cual puede estar asociada a un tipo de bacteria. Lo anterior muestra que cuando se utiliza un esquema de incrustación de palabras, se debería también tener en cuenta el objetivo de dicho proceso para su entrenamiento. De esta forma, se podrían buscar métodos de validación extrínseca que orienten el espacio de embebimiento para hacer más cercanas palabras de relación sintáctica o semántica según el caso, corrección ortográfica, o relación de entidades, respectivamente.

La proximidad de las palabras nos puede dar una percepción de la condición de las maternas, como en los casos anteriores, para la palabra fetal en el *cluster* 1. estas son condiciones antes del parto, por lo que estar en cama y referirse a la mamá tiene alta correlación. Llama la atención la palabra disminuida, que se espera para la actividad uterina o para la frecuencia cardiaca fetal, la cual sería una señal de

alarma y la necesidad de intervención clínica. El *cluster* 2 tiene un tema relacionado con la infección y la posible sepsis. Finalmente, el *cluster* 6 parece referirse a condiciones asociadas a la hipertensión inducida por el embarazo y su complicación, la preeclampsia, causa importante de mortalidad materna.

Desde otro punto de vista, la figura 4 muestra la representación del subconjunto de palabras como otro método de evaluación intrínseco sobre el espacio en baja dimensión 3D de la incrustación de palabras, realizado por la función t-SNE. Esta representación visual facilitó la interpretación entre la relación semántica y sintáctica de las palabras. Permitió analizar visualmente la validez de las incorporaciones aprendidas por el modelo y detallar en su contexto.

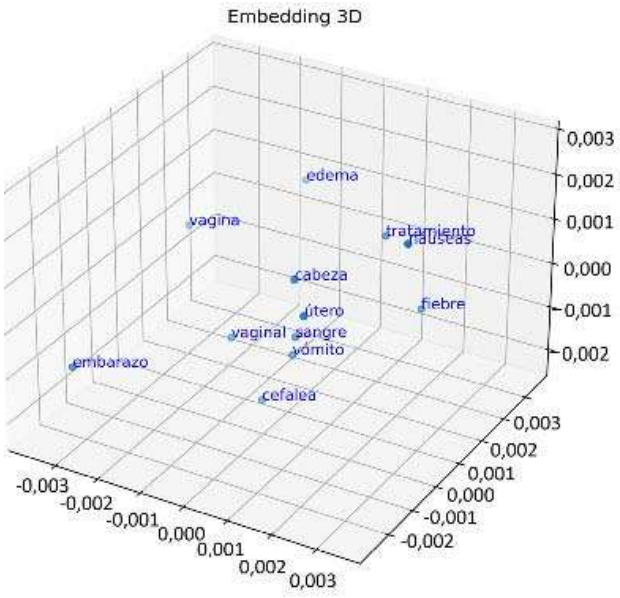

Fig. 4 - Representación de un subconjunto de palabras sobre la incrustación extraída en un espacio reducido 3D.

Se muestra también la relación semántica entre algunas palabras asociadas al contexto de las materna en el embarazo. Se obtiene una fuerte relación entre las palabras “vomito” y “embarazo”, esta relación entre estas palabras está más cerca respecto a otra palabras debido a que uno de los síntomas más comunes y normales durante el embarazo son las náuseas y el vómito, ya que en la gestación se produce

una hormona que activa las sensaciones de náuseas y vómito, llamada la hormona del embarazo: la gonadotropina coriónica humana (hCG). Por otro parte, se puede ver que el subconjunto de palabras "cefalea" y "cabeza" se encuentran alejadas del espacio vectorial debido a que, en el contexto del embarazo para referirse al dolor de cabeza, es común utilizar el término de "cefalea", mientras que el término "cabeza" puede ser usado para referirse a esta parte anatómica del bebé.

En la figura 5 se visualiza la incrustación de palabras en un espacio dimensional reducido previamente por la función t-SNE con los parámetros (perplexity=20, n_iters=1000) a 2D después de haber extraído las incrustaciones de palabras tras el entrenamiento por el modelo CBOW. Para un mejor enfoque y un análisis en la gráfica, se filtraron los tokens que tuvieran una frecuencia superior e igual a 290. También se realizó una comparativa a través del *clustering K-means* con reducción en la dimensionalidad en un espacio 2D a partir de t-SNE.

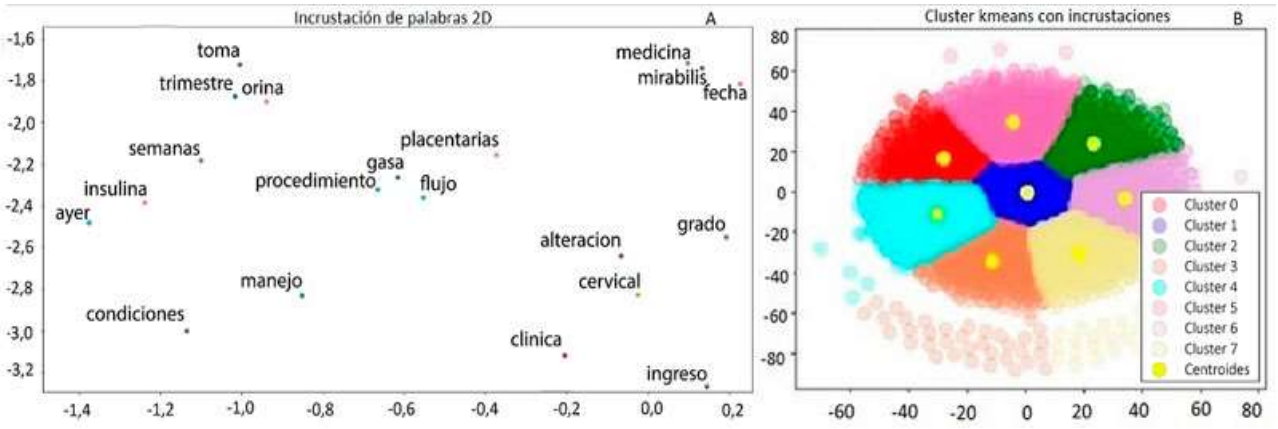

Fig. 5 - A) Extracción del embebimiento de palabras en una dimensionalidad reducida a 2D. B)

Clusters con los datos y los centroides sobre espacio 2D (B).

Se obtuvo la representación de cada *cluster* con la agrupación de las palabras y su respectivo centroide. Para la elección del número de clústeres para el *K-Means* se efectuó el método de validación interna (SSE),[11,16] en donde se sugirió un SSE con k=8. Se realizó una medida de similitud a partir de la distancia euclidiana de las

palabras más cercanas al respectivo centroide de cada cluster, y se obtuvieron las primeras nueve palabras cercanas, como se muestra en la tabla 3. Es de anotar que en esta tabla se están presentando las palabras más cercanas al centroide de cada cluster. Como se mencionó anteriormente, no aparecen palabras relacionadas sintácticamente, sino más bien palabras que podrían ser referidas en un contexto similar. En este trabajo se buscó realizar una incrustación de palabras genéricas de forma no supervisada; sin embargo, trabajos futuros podrían tener algún tipo de evaluación supervisada para orientar la incrustación hacia el objetivo que se esté persiguiendo.

Tabla 3 - Agrupación de palabras por el *cluster* según la mínima distancia Euclidiana entre la palabra y el centroide a través del algoritmo *K-means* con k=8 en 2 dimensiones

Cluster 0	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 5**	**Cluster 6**	**Cluster 7**
Palabra	**Palabra**	**Palabra**	**Palabra**	**Palabra**	**Palabra**	**Palabra**	**Palabra**
cistoflo	catéteres	sepsis	remitido	Ansiosa	perforación	quirúgica	intraepitelial
antihipertensiva	rasgo	diaria	morfina	Ahogo	tamaño	inicialmente	aguda
hidroxicina	triada	malformación	física	Pparaórticas	dextrometer	ingesta	uronalisis
acúfenos	hipertrofia	estabiliza	descarta	ginecobstetricia	esfínteres	estabilizan	intoxicación
infecciones	alerta	trasladada	nitroprusiato	Carbetosina	nocturna	respiratorios	escotomas
hemiabdomen	revisada	laparotomía	malignidad	Parto	hipertensivo	laparatomía	miomectomía
placenta	obstétricos	tranexámico	fetocardias	Asintomático	tratadas	anteparto	cetoacedosis
pancreatitis	artroscopia	adenoides	inestabilidad	Punzada	sérica	neonatos	misoporstol
cavidad	contracciones	intrauterina	cabeza	Sonofobia	glucometría	frialdad	hinchados

En contraste con los *clusters* que se mostró en la tabla 2, es difícil establecer un tema de significancia clínica en estos hallazgos. Se require mejor refinamiento de este escenario.

Conclusiones

En este trabajo se generó un *corpus* a partir de HCE de un Servicio de Ginecoobstetricia de la ciudad de Medellín, Colombia, el cual se caracteriza por atender pacientes de alto riesgo materno. Utilizando métodos de minería de texto se realiza una caracterización del *corpus*, el cual estuvo limitado a las maternas que terminaron el embarazo durante un mes, y se encontraron 659 789 tokens y un diccionario compuesto por 7 334 tokens únicos. Se construyó un *corpus* junto con la arquitectura del modelo CBOW para obtener una representación vectorial de las palabras con 300 dimensiones, con una función de coste a partir de la función de entropía cruzada binaria de 0,001813. Se proporcionó una caracterización vectorial sin la necesidad de recurrir a etiquetas sobre las características o entidades asociadas a alto riesgo obstétrico a partir de técnicas de PLN, lo que permitió implementar un proceso de extracción de la información de manera automática para su posterior análisis.

El *corpus* generado y los resultados encontrados sientan las bases para trabajos futuros en detección de entidades (síntomas, signos, diagnósticos, tratamientos), corrección de errores ortográficos y relaciones semánticas entre palabras para generar resúmenes de historias clínicas o asistir el seguimiento de las maternas mediante la revisión automatizada de la historia clínica electrónica.

Agradecimientos

Este proyecto de investigación se hizo posible gracias a la convocatoria de jóvenes investigadores e innovadores para grupos de investigación ITM – Resolución 000163 de 2020 y a la financiación otorgada por el ITM al proyecto P20242 mediante convocatoria interna. Por último, un agradecimiento a la Clínica Universitaria Bolivariana por facilitar la información de las HCE y hacer posible el desarrollo de este trabajo.

Referencias bibliográficas

1. Chen J, Wei W, Guo C, Tang L, Sun L. Textual analysis and visualization of research trends in data mining for electronic health records. Heal Policy Technol. 2017;6(4):389–400. DOI: http://dx.doi.org/10.1016/j.hlpt.2017.10.003

2. González Bernaldo de Quirós F, Otero C, Luna D. Terminology Services: Standard Terminologies to Control Health Vocabulary. Yearb Med Inform. 2018;27(1):227–33. DOI: http://dx.doi.org/10.1055/s-0038-1641200

3. Resnik P, Niv M, Nossal M, Kapit A, Toren R. Communication of Clinically Relevant Information in Electronic Health Records: A Comparison between Structured Data and Unrestricted Physician Language. Perspect Health Inf Manag. 2008 [acceso: 12/11/2020]. Disponible en: https://perspectives.ahima.org/communication

4. Peng X, Long G, Pan S, Jiang J, Niu Z. Attentive dual embedding for understanding medical concepts in electronic health records. Proc Int Jt Conf Neural Networks. 2019;2019. DOI: http://dx.doi.org/10.1109/IJCNN.2019.8852429

5. Giamouzi M. Discover research from City, University of London. City. 2008;34(2019):51–79. Available from: http://openaccess.city.ac.uk/1189/

6. Neuraz A, Looten V, Rance B, Garcelon N, Llanos LC, et al. Do you need embeddings trained on a massive specialized corpus for your clinical natural language processing task? Stud Health Technol Inform. 2019;264:1558–9. DOI: http://dx.doi.org/10.3233/SHTI190533

7. Khattak FK, Jeblee S, Pou-Prom C, Abdalla M, Meaney C, Rudzicz F. A survey of word embeddings for clinical text. J Biomed Informatics X. 2019;4. DOI: https://doi.org/10.1016/j.yjbinx.2019.100057

8. Khan W, Daud A, Alotaibi F, Aljohani N, Arafat S. Deep recurrent neural networks with word embeddings for Urdu named entity recognition. ETRI J. 2020;42(1):90–100. DOI: https://doi.org/10.4218/etrij.2018-0553

9. Ruas T, Ferreira CHP, Grosky W, de França FO, de Medeiros DMR. Enhanced word embeddings using multi-semantic representation through lexical chains. Inf Sci. 2020;532:16–32. DOI: https://doi.org/10.1016/j.ins.2020.04.048

10. Liu Y, Li Z, Xiong H, Gao X, Wu J. Understanding of internal clustering validation measures. IEEE International Conference on Data Mining (ICDM); 2010.
11. Arrieta Rodríguez EL, Martínez Santos JC. Predicción temprana de morbilidad materna extrema usando aprendizaje automático. Cartagena de Indias: Universidad Tecnológica de Bolívar; 2017.
12. Mohamed EH, Shokry EM. QSST: A quranic semantic search tool based on word embedding. J King Saud Univ - Comput Inf Sci. 2020;(40). DOI: https://doi.org/10.1016/j.jksuci.2020.01.004
13. McDonald S, Ramscar M. Testing the distributional hypothesis: The influence of context on judgements of semantic similarity. University of Edinburgh, Institute for Communicating and Collaborative Systems; 2021.
14. Zakrzewska D. Cluster analysis in personalized e-learning systems. Stud Comput Intell. 2009 [acceso: 12/11/2020];252:29–50. Disponible en: https://link.springer.com/chapter/10.1007/978-3-642-04170-9_10
15. Berzal F. Clustering jerárquico: métodos de agrupamiento. Universidad de Granada; 2020 [acceso: 12/11/2020]. Disponible en: https://elvex.ugr.es/idbis/dm/slides/42Clustering-Hierarchical.pdf
16. García-Alonso CR, Pérez-Naranjo LM, Fernández-Caballero JC. Multiobjective evolutionary algorithms to identify highly autocorrelated areas: The case of spatial distribution in financially compromised farms. Ann Oper Res. 2014 [acceso: 12/11/2020];219(1):187–202. Disponible en: https://link.springer.com/article/10.1007/s10479-011-0841-3
17. Vilà R, Rubio MJ, Berlanga V, Torrado M. Cómo aplicar un cluster jerárquico en SPSS. REIRE Rev d'Innovació i Recer en Educ. 2014 [acceso: 12/11/2020];7(2):113–27. Disponible en: http://revistes.ub.edu/index.php/REIRE

Conflicto de intereses

Los autores declaran que no tienen conflicto de intereses.

Contribución de los autores

María Camila Durango Barrera: Curación de datos, análisis formal, investigación, metodología, recursos software, validación, visualización, redacción – borrador original.

Ever Augusto Torres Silva: Curación de datos, investigación, metodología, recursos, redacción – borrador original.

José Fernando Florez-Arango: Análisis formal, investigación, recursos, validación, redacción – revisión y edición.

Andrés Orozgo-Duque: Conceptualización, adquisición de fondos, investigación, metodología, administración del proyecto, recursos, supervisión, validación, redacción – revisión y edición.