risti
Revista Ibérica de Sistemas e Tecnologias de Informação
Revista Ibérica de Sistemas y Tecnologías de Información

Recebido/Submission: 13/06/2019
Aceitação/Acceptance: 29/08/2019

Descubrimiento de Conocimiento en Historias Clínicas mediante Minería de Texto

Ana Isabel Oviedo Carrascal[1], David Sanguino Cotte[2], Natalia Andrea Restrepo Arango[1], Andrés Felipe Patiño Vélez[1]

ana.oviedo@upb.edu.co, david.sanguino@sanvicentefundacion.com, nataliaandrear@gmail.com, patino.andres@gmail.com

[1] Universidad Pontificia Bolivariana, Circular 1 No. 70-01, 050031, Medellín, Colombia

[2] Hospital San Vicente Fundación, Calle 64 No. 51D-154, 050010, Medellín, Colombia

DOI: 10.17013/risti.34.29–43

Resumen: Las instituciones clínicas presentan una alta generación de datos no estructurados tanto en el registro de procedimientos en texto libre por parte del personal médico, como por las imágenes y videos generados por las ayudas diagnósticas. En este trabajo se plantea un proceso de descubrimiento de conocimiento en el texto no estructurado de las historias clínicas del área de traumatología del Hospital San Vicente Fundación mediante minería de texto. Para el estudio se aplicaron técnicas de preparación de texto como eliminación de palabras no relevantes, sustitución de términos, eliminación de acentos y derivación de palabras. Respecto a los procesos de minería se aplicaron técnicas de aprendizaje supervisado y no supervisado como árboles de decisión, regresión logística, k-vecinos más cercanos, clustering jerárquico y reglas de asociación. El resultado obtenido es la conformación de un modelo de las palabras más relevantes en los registros clínicos del Hospital en el área de traumatología.

Palabras-clave: Minería de texto; minería de datos de la salud; procesamiento del lenguaje natural.

Knowledge Discovery in Medical Records through Text Mining

Abstract: The clinical institutions generate a large amount of unstructured data both in the registration of procedures in free text by medical staff, and by the images and videos generated by diagnostic aids. This paper proposes a process of knowledge discovery in the unstructured text of the medical records of the trauma area of the San Vicente Foundation Hospital through text mining. Text preparation techniques were applied such as elimination of non-relevant words, substitution of terms, elimination of accents and derivation of words. Regarding mining processes, supervised and unsupervised learning techniques were applied such as decision trees, logistic regression, nearest k-neighbors, hierarchical clustering and association rules. The result obtained is the conformation of a model of the most relevant words in the clinical records of the Hospital in the area of traumatology.

Keywords: Text mining; health data mining; natural language processing.

1. Introducción

Análisis mundiales muestran que cerca del 80% de la información en las organizaciones se encuentra almacenada como datos no estructurados, los cuales requieren procesos de organización interna para la generación automática de conocimiento (Kharrazi et al., 2018). Entre los datos no estructurados se encuentra el texto libre, el cual requiere procesos avanzados de analítica y de procesamiento de lenguaje natural para descubrir conocimiento relevante.

La minería de texto puede entenderse como un área interdisciplinaria utilizada para encontrar conocimiento útil a partir de documentos de texto no estructurado, mediante la aplicación de métodos que incluyen técnicas de procesamiento de lenguaje natural, recuperación de información, aprendizaje automático, estadística, lingüística computacional, además de aquellos métodos aplicados a la minería de datos convencional. La minería de texto es similar a la minería de datos en el sentido de que se hacen exploraciones en grandes cantidades de información para extraer nuevo conocimiento. La diferencia radica en que, mientras la minería de datos se enfoca en datos estructurados, la minería de texto se enfoca en datos no estructurados, por lo que se hace necesario el uso de técnicas adicionales que permitan un adecuado procesamiento de lenguaje natural (Sukanya & Biruntha, 2012) (Otsuka & Matsushita, 2012).

La minería de texto es aplicable a todos los procesos en los que se involucre el lenguaje escrito, lo cual abre las posibilidades a prácticamente todos los procesos humanos, como medicina, mercadeo, educación, idiomática, internet, etc. El caso de las instituciones clínicas en particular es de alta generación de datos no estructurados, debido a la naturaleza de los procesos que allí se tienen, puesto que el personal médico registra procedimientos mediante redacción de texto libre, además de las imágenes generadas por las ayudas diagnósticas. El conocimiento oculto en estos datos tiene un alto potencial para las instituciones clínicas respecto al mejoramiento de la calidad de sus servicios, que hoy en día se ven medidos en indicadores de gestión como índice de mortandad, días promedio de estancia de los pacientes y egresos por cama, entre otros (Hospital San Vicente Fundación, 2017).

La minería de texto ha sido ampliamente utilizada en la salud para el estudio de diferentes enfermedades. En (Kushima & Nikama, 2012) se realizó un trabajo de minería de texto sobre los registros médicos realizados por enfermeras y doctores para los pacientes con hepatitis crónica en el Hospital Universitario de Miyazaki en Japón, con el propósito de explotar síntomas similares. En (Pereira & Agostinho, 2013) se propone un proceso automático de clasificación para diagnósticos de epilepsia, mediante la aplicación de la técnica de k-vecinos más cercanos permitiendo mapear los códigos de las enfermedades de acuerdo al estándar preestablecido por ICD-9 (Clasificación Estadística Internacional de Enfermedades y Problemas Relacionados con la Salud). En (Vijayakrishnan & Stewart, 2014) se encuentra una investigación en la que se aplica minería de texto para la detección temprana de las señales y síntomas de fallas cardíacas, dentro de pacientes participantes en programas de salud preventiva. En (Karystianis et al., 2015) se usaron 56 mil prescripciones de medicamentos para la atención primaria escritas en texto libre, las cuales fueron analizadas con reglas de diccionario y composición léxica que permitieron la estructuración de datos. En (Lucini et al., 2017) se aplican métodos de

minería de texto para predecir futuras hospitalizaciones y altas utilizando registros médicos tempranos del departamento de emergencias. En (Judd, 2018) se analiza el texto de registros médicos electrónicos para definir las opciones de tratamiento para pacientes con dolor lumbar, usando la herramienta clínica cTAKES (*NLP Apache Clinical Text Analysis and Knowledge Extraction System*), el algoritmo de aprendizaje automático utiliza siete años de notas clínicas extraídas del médico de atención primaria para clasificar el patrón de dolor de espalda de 20 pacientes.

De la revisión de la literatura se puede inferir que la minería de datos tiene un gran potencial para el diagnóstico temprano de enfermedades, perfilamiento de pacientes, generación de historia clínica, búsqueda de información y variables de interés en investigación médica, entre otros (Oviedo & Sánchez, 2017). Entre los retos para el sector está el hecho que los datos médicos son complejos y difíciles de analizar, por lo que se debe trabajar en estandarizar la terminología médica para aplicar procesos de minería de textos que generen datos de alta calidad para obtener resultados óptimos desde los procesos de minería (Sun et al., 2018).

Con el objetivo de aportar en el descubrimiento de conocimiento en historias clínicas, en este artículo se realiza un estudio de los registros clínicos de pacientes de traumatología del Hospital Universitario San Vicente Fundación mediante minería de texto, como apoyo al análisis de las palabras utilizadas en los diagnósticos del área.

El Hospital Universitario San Vicente Fundación es una de las principales instituciones hospitalarias de Colombia, siendo una institución privada sin ánimo de lucro que presta servicios de salud con énfasis en la atención del paciente de alta complejidad. La unidad de Urgencias Adultos es uno de los principales servicios que ofrece el Hospital a la comunidad, en el cual se brindan servicios de primeros auxilios, cirugía y unidad de cuidado intensivo. Esta unidad, es considerada como el primer centro de atención de trauma del país, siendo la principal puerta de entrada de los pacientes, con un 96% del total de los 85.000 pacientes que en promedio atiende el Hospital durante cada año. El 26% de los casos atendidos en urgencias se debe a trauma, área en la cual se centra el presente proyecto.

En el Hospital, el proceso de documentación de la historia clínica se realiza de dos maneras: una estructurada, en donde se almacenan los datos básicos del paciente, los tratamientos y procedimientos realizados, además de los medicamentos que recibe; otra no estructurada, donde se redacta de manera libre los acontecimientos médicos del paciente por cada ingreso al hospital, técnicamente cada ingreso recibe el nombre de "episodios". La evolución médica de un episodio es consignada de forma transaccional, generando una serie de documentos por medio de texto libre donde se diligencian los campos: SUBJETIVO, OBJETIVO, ANÁLISIS Y PLAN. El campo SUBJETIVO almacena lo que dice el paciente, el campo OBJETIVO almacena lo que el médico encuentra en el paciente al examinarlo, el campo ANALISIS almacena lo que el médico piensa según su criterio y el campo PLAN almacena el tratamiento que debe seguir el paciente según el criterio del médico.

En el presente trabajo se desarrolla un modelo de palabras para el área de traumatología del Hospital San Vicente Fundación, analizando los campos SUBJETIVO, OBJETIVO, ANÁLISIS y PLAN de las historias clínicas mediante técnicas de minería de texto.

2. Materiales y Métodos

En esta sección se describe la metodología CRISP-DM usada en el proceso de minería, se detallan los datos analizados y se presenta el diseño de los modelos analíticos desarrollados en este trabajo para el descubrimiento de conocimiento en historias clínicas del Hospital San Vicente Fundación.

2.1. Metodología CRISP-DM para el proceso de minería de datos

CRISP-DM (*CRoss-Industry Standard Process for Data Mining*) es una metodología de circulación libre, destacada por su rigurosidad y completitud a través de 6 fases: entendimiento del negocio, entendimiento de los datos, preparación de los datos, modelado, evaluación y despliegue (Wirth & Hipp, 2000).

El entendimiento del negocio se centra en la comprensión de los objetivos y requisitos del proyecto desde una perspectiva empresarial, formalizando el objetivo de desarrollar un modelo de palabras para el área de traumatología del Hospital San Vicente Fundación, analizando las historias clínicas mediante técnicas de minería de texto. El entendimiento de los datos realiza una aproximación estadística de la información, realizando un perfilamiento de los campos analizados de las historias clínicas correspondientes al diagnóstico según el código CIE-10 y los campos de texto libre SUBJETIVO, OBJETIVO, ANÁLISIS y PLAN. La preparación de datos es una de las fases más importantes y con frecuencia es la que más tiempo requiere para realizar actividades como limpieza y transformación de los datos, que en este trabajo requiere la aplicación de técnicas de procesamiento de lenguaje natural. El modelado involucra la aplicación de técnicas de aprendizaje de máquinas que den solución al problema planteado. La evaluación utiliza métricas para determinar si los resultados del modelado cumplen los criterios de calidad esperados, requiriendo una comprensión de los objetivos del negocio. Finalmente, el despliegue consiste en utilizar los nuevos conocimientos generados para implementar acciones en la organización.

2.2. Datos analizados

Los datos analizados corresponden a 19.078 pacientes, 30.958 episodios, 174.474 documentos y 351.627 registros de traumatología en un año seleccionado por el Hospital. En la tabla 1 se especifican los atributos seleccionados para el estudio de las historias clínicas de los pacientes de traumatología.

Variable	Tipo de Variable	Descripción de la variable
Episodio	Numérica	Número consecutivo asociado a cada ingreso del paciente al hospital
Documento	Numérica	Número consecutivo asociado a cada ingreso de información en los campos de texto. La base contiene tantos registros para un documento, como diagnósticos se haya asociados al documento.
Código del diagnóstico	Categórica	Código alfanumérico del diagnóstico. Campo asociado a la clasificación internacional CIE-10.
Descripción del diagnóstico	Categórica	Descripción del diagnóstico

Variable	Tipo de Variable	Descripción de la variable
Análisis	Texto	Campos "Análisis": lo que el médico piensa según su especialidad y criterio
Objetivo	Texto	Campos "Objetivo": lo que el médico le encuentra al paciente al examinarlo
Plan	Texto	Campo "Plan": lo que el médico considera según su especialidad debe seguir en el tratamiento del paciente
Subjetivo	Texto	Campo "Subjetivo": lo que dice el paciente o el acompañante

Tabla 1 – Variables de la historia clínica seleccionadas para el estudio.

El diagnóstico es la variable objetivo considerada en este estudio, ya que se plantea la creación de un modelo de palabras para cada código CIE-10. Los diagnósticos con mayor cantidad de registros asociados se presentan en la tabla 2.

Código	Diagnóstico	Cantidad	%
S069	Traumatismo intracraneal, no especificado	18.232	4,8%
T07X	Traumatismos múltiples, no especificados	11.115	2,9%
S099	Traumatismo de la cabeza, no especificado	6.158	1,6%
S822	Fractura de la diafisis de la tibia	6.124	1,6%
S065	Hemorragia subdural traumatica	5.262	1,4%
T814	Infección consecutiva a procedimiento, no clasificada en otra parte	5.046	1,3%
T212	Quemadura del tronco, de segundo grado	4.974	1,3%
T202	Quemadura de la cabeza y del cuello de segundo grado	4.846	1,3%
S211	Herida de la pared anterior del torax	4.576	1,2%
T810	Hemorragia y hematoma que complican un procedimiento, no clasificados en otra parte	4.442	1,2%

Tabla 2 – Lista de los 10 diagnósticos con mayor cantidad de registros.

Las variables de tipo texto fueron sometidas a un proceso de preparación conocido como "Bolsa de Palabras", en el cual se crea un diccionario con todas las palabras disponibles en el conjunto de documentos y se representan por medio de una matriz de términos. Esta representación ignora el orden en el cual aparecen las palabras y sólo tienen en cuenta la frecuencia de aparición (Cummins et al., 2018).

En el enfoque de "Bolsa de Palabras" se aplica un proceso de limpieza a los campos tipo texto conformado por 7 pasos:

- Eliminación de registros con campos nulos.

- Eliminación de caracteres especiales (tabulaciones, caracteres no imprimibles), signos de puntuación, números y espacios múltiples.
- Eliminación de artículos y preposiciones (stop words).
- Sustitución de sinónimos y acrónimos, los cuales fueron extraídos con el apoyo del Manual de Normas y Procedimientos en Trauma (Martiniano, Restrepo, & Múnera, 2016), el cual es un libro de referencia tanto en la facultad de medicina de la Universidad de Antioquia como en el Hospital para la formación en el área de trauma. Se realizaron un total de 45.886 sustituciones de términos.
- Eliminación de acentos para evitar problemas ortográficos mediante la sustitución de los acentos del idioma español correspondientes a los siguientes caracteres {á, é, í, ó, ú, ñ, ü}, siendo sustituidos respectivamente por su respectiva letra sin acento {a, e, i, o, u, n, u}.
- Eliminación de palabras no relevantes para el área ya que se presentan constantemente en todos los registros como {“paciente”, “anos”, “dias”, “ahora”, “manejo”, “buena”, “manana”, “medico”, “mas”, “adecuada”, “dia”, “ayer”, “debe”, “aun”, “ademas”, “hoy”, “pte”}.
- Reducción de las palabras a sus raíces (stemming): este proceso permite reemplazar las palabras “niña” y “niño” simplemente por la raíz “niñ” y así eliminar la variabilidad de las palabras.

Finalmente, después de ser limpiado el texto es representado de forma numérica por medio de la frecuencia de aparición de las palabras en cada documento. Algunos pasos adicionales para limpiar el texto implican eliminar las palabras que tiene una frecuencia muy baja (se encuentran en muy pocos documentos) y las palabras que tienen una frecuencia muy alta (están en casi todos los documentos).

2.3. Modelos analíticos

Una vez preparado el texto y representado de forma numérica por la frecuencia de las palabras, se aplican técnicas de minería de datos para el descubrimiento de información relevante. En la minería de datos se desarrollan principalmente dos tipos de análisis: predictivos y descriptivos. El análisis predictivo permite analizar datos futuros, prediciendo tanto valores categóricos (clasificación) como numéricos (regresión) por medio de técnicas de aprendizaje supervisado de máquinas. Algunos algoritmos supervisados son árboles de decisión, máquinas de soporte vectorial, naive bayes, redes neuronales, k vecinos más cercadnos y regresiones. Por su parte, el análisis descriptivo permite descubrir conocimiento en los datos actuales por medio de agrupaciones, reglas de asociación y selección de factores aplicando técnicas de aprendizaje no supervisado de máquinas. Algunos algoritmos no supervisados son k-means, clustering jerárquico, a priori, entre otros (Oviedo, Vélez, & Oviedo, 2015).

Para encontrar el modelo de palabras usadas en traumatología, se aplicaron tanto análisis predictivos como descriptivos, creando 5 modelos analíticos. Las técnicas empleadas fueron seleccionadas según la revisión bibliográfica sobre minería de texto aplicada a datos de la salud.

- Modelo 1 - Palabras más frecuentes en el área de traumatología: tiene por objetivo encontrar un ranking de las palabras más comúnmente utilizadas en el área de traumatología mediante la frecuencia de aparición de cada término.

- Modelo 2 - Segmentación de los episodios para encontrar similitudes en los traumas según las palabras: tiene por objetivo buscar similitudes entre los diferentes tipos de traumas según las palabras empleadas por los médicos mediante el algoritmo de clúster jerárquico aglomerativo.
- Modelo 3 - Encontrar las palabras más utilizadas en cada tipo de trauma: tiene como objetivo encontrar las palabras más utilizadas por los médicos en cada tipo de trauma, mediante el algoritmo predictivo de árbol de decisión.
- Modelo 4 - Analizar la ocurrencia conjunta de las palabras en los diferentes tipos de trauma: tiene por objetivo buscar las co-ocurrencias de las palabras teniendo en cuenta los diferentes tipos de trauma, mediante el algoritmo descriptivo apriori.
- Modelo 5 - Predecir el tipo de trauma según las palabras ingresadas por el médico: tiene por objetivo realizar una predicción del tipo de trauma de un episodio, teniendo en cuenta las palabras utilizadas por el médico y las variables estructuradas de la base de datos.

3. Resultados

Los experimentos se realizaron con la herramienta R, debido a su enfoque estadístico con una base analítica flexible y confiable para la informática. A continuación, se reportan los resultados de los modelos analíticos.

3.1. Modelo 1 - Palabras más frecuentes en el área de traumatología

Después de preparar el texto, se calcula la frecuencia de aparición de cada palabra. La Figura 1 presenta la nube de palabras después del proceso de preparación de los campos tipo texto, donde el tamaño de la palabra corresponde a la frecuencia de aparición en los registros. Las palabras se aprecian incompletas debido al proceso de reducción de raíces (*stemming*) aplicado en la preparación de datos.

Figura 1 – Nube de palabras después de aplicar la preparación del texto. El tamaño de la palabra corresponde a la frecuencia de aparición.

Mediante un histograma de frecuencias se verifican las palabras más comunes en el conjunto de datos: evolución, fractura, derecho, trauma, estable, quemadura, continúa, izquierdo, clínico, momento.

3.2. Modelo 2 - Segmentación de los episodios para encontrar similitudes en los traumas

A la matriz de frecuencias de las palabras se aplicó un clúster jerárquico aglomerativo, el cual realiza una matriz de distancias que es posible visualizar mediante un dendograma. Para este método no es necesario indicar la cantidad de grupos, sino que se traza una línea horizontal que corta la gráfica y agrupa las palabras por las jerarquías que quedan en la parte inferior. A partir de las relaciones vistas en el dendograma se eligieron cuatro clústeres, como se aprecia en la Figura 2 correspondientes a: (1) lesiones en los miembros superiores e inferiores ocasionados por accidentes de tránsito, (2) lesiones por quemaduras, (3) seguimientos clínicos y (4) seguimientos donde se encuentran involucradas observaciones referentes a la circulación sanguínea.

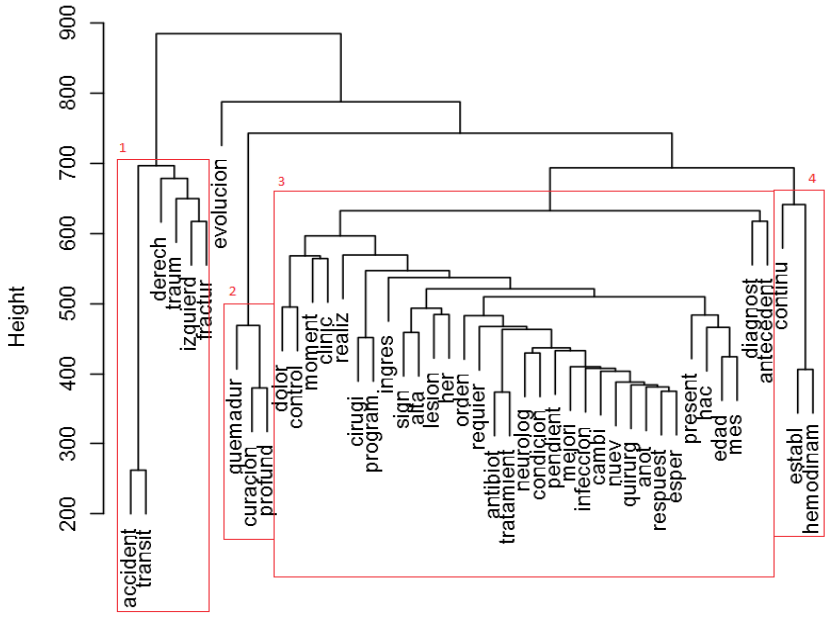

Figura 2 – Selección de clústers del dendograma para buscar similitudes entre las palabras.

3.3. Modelo 3 - Encontrar las palabras más utilizadas en cada tipo de trauma

A la matriz de frecuencia de las palabras se aplica el algoritmo "*Recursive Partitioning And Regression Trees*" de árboles de decisión, para hacer una predicción del tipo de trauma según las palabras de la historia clínica. Para la evaluación de este método, se separó el conjunto de entrenamiento del conjunto de prueba con una división aleatoria

70%-30% respectivamente. En el conjunto de prueba, se calculan las medidas de matriz de confusión, precisión, cobertura y área ROC.

La variable objetivo a predecir corresponde a 23 grupos de diagnósticos de la norma CIE-10, así que se realizaron 23 clasificaciones binomiales, obteniendo un árbol de predicción para cada código CIE-10. En la Figura 3 se presentan los árboles para los dos primeros códigos CIE-10, donde un color azul fuerte en las hojas de los árboles indica que sí pertenece al tipo de trauma evaluado, mientras que un color muy claro significa que no pertenece al tipo de trauma. En cada nodo, adicionalmente se indica el porcentaje de los datos que recibe la predicción.

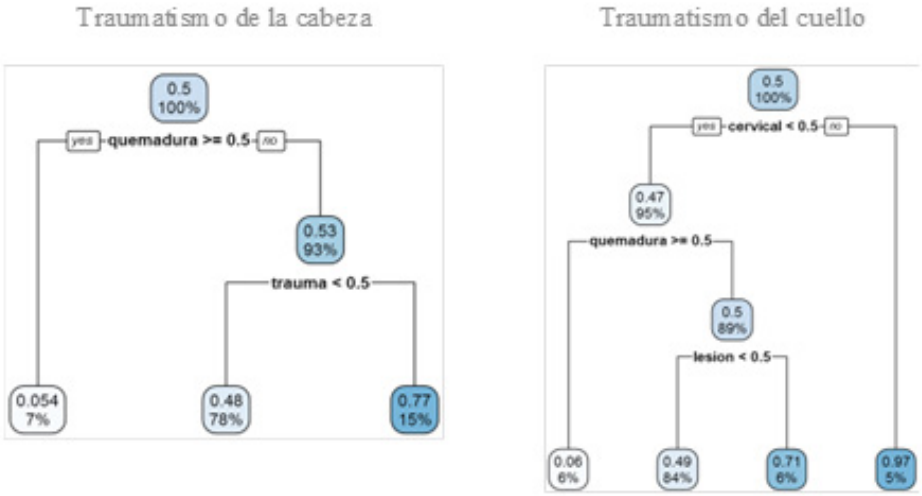

Figura 3 – Árboles de decisión para los dos primeros tipos de trauma según el código CIE-10. Los otros 21 árboles no se presentan en el documento.

Según los árboles de decisión, las palabras más relevantes de los 10 primeros traumas del código CIE-10 se presentan en la tabla 3.

Tipo de Trauma	Palabras relevantes
Traumatismo de la cabeza	• Quemadura • Trauma
Traumatismos del cuello	• Cervical • Quemadura • Lesión
Traumatismos del tórax	• Tórax • Quemadura
Traumatismos del abdomen, de la región lumbosacra de la columna lumbar y de la pelvis	• Quemadura
Traumatismos del hombro y del brazo	• Humero • Hombro • Quemadura

Tipo de Trauma	Palabras relevantes
Traumatismos del antebrazo y del codo	• Radio • Quemadura • Fractura
Traumatismos de la muñeca y de la mano	• Mano • Quemadura • Dedo
Traumatismos de la cadera y del muslo	• Cadera • Fémur • Quemadura
Traumatismos de la rodilla y de la pierna	• Tibia • Rodilla • Quemadura • Tránsito
Traumatismos del tobillo y del pie	• Pie • Tobillo • Fractura • Tránsito • Quemadura

Tabla 3 – Palabras relevantes de los 10 primeros códigos CIE-10 de traumatología. Los otros 13 códigos analizados no son presentados en este documento.

3.4. Modelo 4 - Analizar la ocurrencia de las palabras en los diferentes tipos de trauma

Después de realizar la preparación de datos, los episodios son representados en vectores binarios que indican la presencia de las palabras en los documentos. Posterior a la representación se aplica el algoritmo apriori, el cual busca las co-ocurrencias de las palabras, es decir las palabras que aparecen de forma conjunta en los diagnósticos de trauma. El método se evalúa mediante la confianza de las reglas. En la Figura 4 se presentan algunas reglas de asociación entre las palabras. Se evidencia un bajo nivel de co-ocurrencia para las palabras analizadas, la gran mayoría de ellos con valores inferiores al 20%. La confianza más alta se presenta con las palabras "quemadura" y "profunda", indicando que el 57% de las veces que ocurre la palabra "quemadura", está acompañada de "profunda".

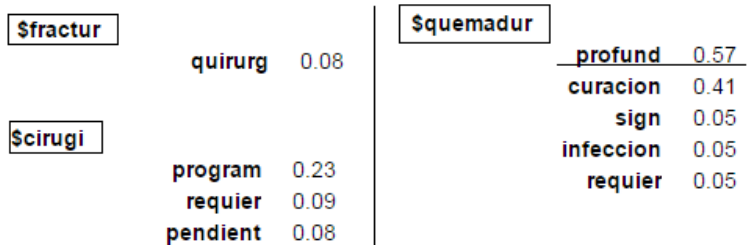

Figura 4 – Reglas de asociación encontradas entre las palabras.

3.5. Modelo 5 - Predecir el tipo de trauma según las palabras ingresadas por el médico

La predicción del tipo de trauma según las palabras empleadas por los médicos fue realizada mediante tres métodos: árbol de decisión, regresión logística y clasificador de vecinos cercanos. Para evaluar los métodos se aplicó una división aleatoria del 70% para los datos de entrenamiento y 30% para los datos de prueba.

El desempeño de los métodos puede apreciarse gráficamente en la Figura 5 que presenta el área ROC para la predicción de los 23 tipos de trauma.

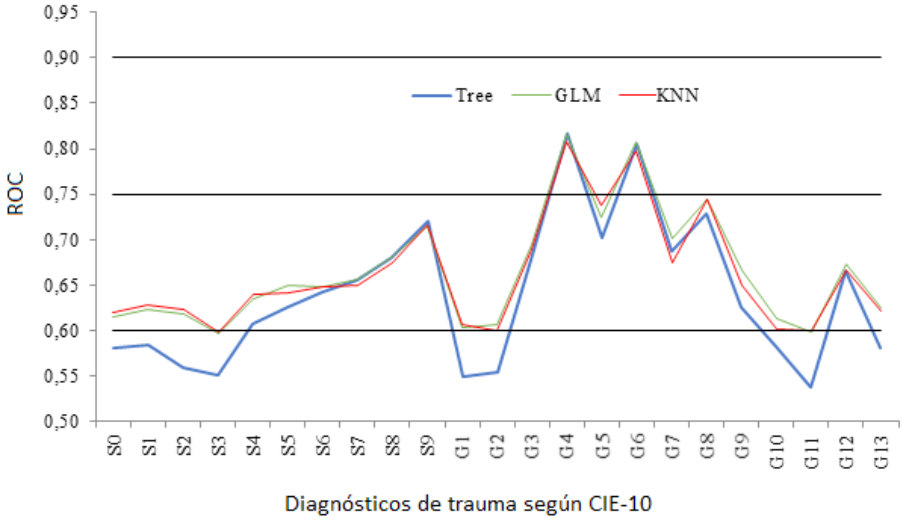

Figura 5 – Gráfica de líneas con el área bajo la curva ROC de los métodos predictivos: árbol de decisión, regresión y k-vecinos más cercanos.

En la Figura 5, los diagnósticos de trauma según CIE-10 son:

S0: Traumatismo de la cabeza
S1: Traumatismos del cuello
S2: Traumatismos del tórax
S3: Traumatismos del abdomen, de la región lumbosacra de la columna lumbar y de la pelvis
S4: Traumatismos del hombro y del brazo
S5: Traumatismos del antebrazo y del codo
S6: Traumatismos de la muñeca y de la mano
S7: Traumatismos de la cadera y del muslo
S8: Traumatismos de la rodilla y de la pierna
S9: Traumatismos del tobillo y del pie
G1: Traumatismos que afectan múltiples regiones del cuerpo
G2: Traumatismos de parte no especificada del tronco, miembro o región del cuerpo
G3: Efectos de cuerpos extraños que penetran por orificios naturales
G4: Quemaduras y corrosiones de la superficie externa del cuerpo, especificadas por sitio
G5: Quemaduras y corrosiones limitadas al ojo y órganos internos
G6: Quemaduras y corrosiones de múltiples regiones del cuerpo y las no especificadas
G7: Congelamiento
G8: Envenenamiento por drogas, medicamentos y sustancias biológicas
G9: Efectos tóxicos de sustancias de procedencia principalmente no medicinal
G10: Otros efectos y los no especificados de causas externas
G11: Algunas complicaciones precoces de traumatismos
G12: Complicaciones de la atención médica y quirúrgica, no clasificadas en otra parte
G13: Secuela de traumatismos, de envenenamientos y de otras consecuencias de causa externa

Según los resultados de la Figura 5, el método de predicción seleccionado es la regresión logística, ya que presenta mayor valor en el área ROC. El procedimiento para realizar la predicción se presenta a continuación:

- Se ingresa el texto del episodio y se realiza el proceso de depuración de texto.
- Posteriormente se construye la matriz de términos del texto ingresado.
- Se aplica el método de predicción.

Para el caso de prueba, se ingresó el siguiente texto: "masculino raza negra edad actualmente hospitalizado intoxicacion causticos consecuencia misma quemaduras tracto gastrointestinal hallazgo incidental tac contr". Obteniendo los resultados para cada uno de los grupos de trauma, tal como se muestra en la tabla 4. El tipo de trauma al que pertenece el texto según la predicción, es el que tiene mayor puntaje en la tabla de probabilidad. También puede ocurrir, que un texto esté relacionado con varios traumas y esto puede ser útil a los médicos en el momento del diagnóstico. Para el ejemplo, en la tabla 4 se presentan ordenados de mayor a menor probabilidad, los resultados del algoritmo predictor donde los grupos con mayor probabilidad están relacionados con intoxicaciones por ingestión de sustancias, seguido por grupos relacionados con quemaduras.

Grupo	Valor	Descripción del grupo
prob_G8	0,98	Envenenamiento por drogas, medicamentos y sustancias biológicas
prob_G9	0,98	Efectos tóxicos de sustancias de procedencia principalmente no medicinal
prob_G4	0,93	Quemaduras y corrosiones de la superficie externa del cuerpo, especificadas por sitio
prob_G6	0,93	Quemaduras y corrosiones de múltiples regiones del cuerpo y las no especificadas
prob_G5	0,85	Quemaduras y corrosiones limitadas al ojo y órganos internos

Tabla 4 – Resultado del algoritmo de predicción.

4. Discusión

A continuación, se presenta la discusión de resultados de los modelos analíticos.

4.1. Palabras más frecuentes en el área de traumatología

Según su frecuencia de aparición en los episodios, las 10 palabras más utilizadas en los diagnósticos son: Evolución, Fractura, Derecho, Trauma, Estable, Quemadura, Continúa, Izquierdo, Clínico, Momento.

De estas palabras puede apreciarse que {evolución, estable, continúa, clínico, momento} son palabras asociadas al seguimiento del paciente, mientras que las palabras {fractura, derecho, trauma, quemadura, izquierdo} son palabras asociadas al diagnóstico.

4.2. Segmentación de los episodios para encontrar similitudes en los traumas según las palabras

Por medio de un clustering jerárquico aglomerativo se encontraron 4 grupos, los cuales se describen a continuación.

- Cluster 1: Lesiones en los miembros superiores e inferiores ocasionados por accidentes de tránsito
- Cluster 2: Lesiones por quemaduras
- Cluster 3: Seguimientos clínicos
- Cluster 4: Seguimientos con observaciones sobre circulación sanguínea

4.3. Encontrar las palabras más utilizadas en cada tipo de trauma

La creación de los 23 modelos predictivos para cada tipo de trauma según el código CIE-10 permite identificar aquellas palabras relevantes para predecir dichos tipos de trauma, a diferencia del primer modelo realizado donde se seleccionaron las palabras por la frecuencia de aparición. A continuación, se listan algunas de las palabras seleccionadas por los 23 experimentos de predicción: Quemadura, Fractura, Curación, Tránsito, Intoxicación, Cervical, Rodilla, Pie, Fistula, Muñón, Mano, Radio, Lesión, Tibia, Fémur, Tobillo, entre otras.

4.4. Analizar la ocurrencia conjunta de las palabras en los diferentes tipos de trauma

Los resultados encontrados con este modelo analítico concuerdan con los experimentos anteriores, donde las reglas de asociación para las palabras {Evolución, Continúa, Estable, Clínico, Cirugía, Dolor} parecen estar relacionadas a seguimientos de pacientes; mientras que las reglas descritas para las palabras {Fractura, Trauma, Derecha, Izquierda, Quemadura, Diagnóstico, Antecedente} son reglas que por sus palabras asociadas parecen estar relacionadas con documentos en donde se diagnóstica al paciente. Estos resultados sugieren como trabajo futuro crear modelos de palabras por separado para las expresiones textuales de los pacientes y las expresiones médicas.

4.5. Predecir el tipo de trauma según las palabras ingresadas por el médico

De los resultados encontrados, se aprecia que la predicción con mejor desempeño corresponde a los grupos de diagnóstico relacionados con quemaduras, luego de estos se tienen los grupos relacionados con traumatismos en las extremidades y en la cadera con un desempeño aceptable. Por último, con menor desempeño los predictores para los diagnósticos relacionados con traumatismos en cabeza, tórax, abdomen y partes del cuerpo no especificadas.

El bajo desempeño de los predictores en los grupos diagnósticos relacionados con traumatismos en cabeza, tórax y abdomen, puede estar relacionado con las múltiples complicaciones que se presentan en los órganos ubicados en estas partes de cuerpo, lo cual genera términos dispersos a lo largo de todos los registros y por ende los algoritmos no logran crear relaciones entre ellos.

El bajo desempeño de los predictores en los grupos diagnósticos de los grupos G1, G2, G10 y G11, puede deberse a que son grupos que abarcan complicaciones muy generales, presentándose el mismo fenómeno descrito en el párrafo anterior. Además, la generalidad de estos grupos de diagnóstico les permite ser usados para clasificar de manera poco específica los pacientes en su ingreso al área de traumatología. Los grupos mencionados corresponden a las siguientes clasificaciones de trauma: (G1) traumatismos que afectan

múltiples regiones del cuerpo; (G2) traumatismos de parte no especificada del tronco, miembro o región del cuerpo; (G10) Otros efectos y los no especificados de causas externas; y (G11) Algunas complicaciones precoces de traumatismos.

5. Conclusiones

Con los modelos analíticos desarrollados con las historias clínicas del Hospital San Vicente Fundación, fue posible identificar el modelo de palabras usado en el área de traumatología, identificando:

- Las palabras más usadas en el conjunto global de documentos.
- Cuatro grupos de traumas que usan palabras muy similares.
- Las palabras asociadas a diferentes tipos de traumas.
- Las palabras que ocurren de forma conjunta.
- La predicción de tipo de trauma según las palabras del médico.

Para mejorar el desempeño de los modelos analíticos se plantea como sugerencia ampliar el diccionario de sinónimos y acrónimos para realizar una mejor limpieza de los datos. Adicionalmente, se sugiere realizar un ejercicio de extracción de los números presentes en los campos de texto, buscando construir gráficas evolutivas de variables médicas importantes que no se registran en los campos estructurados. Por ejemplo, índices de gravedad en trauma, frecuencia cardiaca, presión arterial, saturación de oxígeno, temperatura corporal, etc.

Referencias

Cummins, N., Amiriparian, S., Ottl, S., Gerczuk, M., Schmitt, M., & Schuller, B. (2018). Multimodal Bag-of-Words for cross domains sentiment analysis. *En 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4954–4958. Doi: 10.1109/ICASSP.2018.8462660

Hospital San Vicente Fundación (2017). Hospital Universitario San Vicente Fundación – Indicadores. Recuperado de: http://hospitaluniversitario.sanvicentefundacion.com/documentos/HU_Indicadores_Resultado.pdf

Judd, M. (2018). A Medical Decision Support Tool Using Text-mining Techniques with Electronic Medical Records. *In Inquiry@ Queen's Undergraduate Research Conference Proceedings.*

Karystianis, G., Sheppard, T., Dixon, W.G., & Nenadic, G. (2015). Modelling and extraction of variability in free-text medication prescriptions from an anonymised primary care electronic medical record research database. *BMC Medical Informatics and Decision Making*, *16*(1), 18.

Kharrazi, H., Anzaldi, L.J., Hernandez, L., Davison, A., Boyd, C.M., Leff, B., & Weiner, J.P. (2018). The value of unstructured electronic health record data in geriatric syndrome case identification. *Journal of the American Geriatrics Society*, 66(8), 1499–1507. doi:10.1111/jgs.15411

Kushima, M.A., & Nikama, T. (2012). Text Data Mining of the Electronic Medical Record of the Chronic Hepatitis Patient. *En International multiconference of engineers and computer scientists*, 1, Hong Kong.

Lucini, F.R., Fogliatto, F.S., Da Silveira, G.J., Neyeloff, J.L., Anzanello, M.J., Kuchenbecker, R.D.S., & Schaan, B.D. (2017). Text mining approach to predict hospital admissions using early medical records from the emergency department. *International journal of medical informatics*, 100, 1–8. Doi: 10.1016/j.ijmedinf.2017.01.001

Martiniano, J., Restrepo, J., & Múnera, A. (2016). *Manual de normas y procedimientos en trauma*. Medellín: Editorial Universidad de Antioquia.

Otsuka, N., & Matsushita, M. (2014). Constructing knowledge using exploratory text mining. *En Joint 7th International Conference on and Advanced Intelligent Systems (ISIS)*, 15th International Symposium, Kitakyushu, 1392–1397. Doi: 10.1109/SCIS-ISIS.2014.7044806

Oviedo, A., & Sanchez, S. (2017). Minería de datos de la salud: Sistema de votación de técnicas analíticas para identificar los factores que influyen en la realización de cirugías estéticas. *Revista Politecnica,* 13, 43–52. Doi: 10.33571/rpolitec

Oviedo, A., Velez, G., & Oviedo, E. (2015). Minería de datos: aportes y tendencias en el servicio de salud de ciudades inteligentes. *Revista Politecnica*, 11, 111-120. Doi: 10.33571/rpolitec

Pereira, L.R., & Agostinho, M. (2013). ICD9-based Text Mining Approach to Children Epilepsy Classification. *Procedia Technology*, 9, 1351–1360. doi: 10.1016/j.protcy.2013.12.152

Sukanya, M., & Biruntha, S. (2012). Techniques on text mining. *En 2012 IEEE International Conference on Advanced Communication Control and Computing Technologies (ICACCCT)*, 267–271. Doi: 10.1109/ICACCCT.2012.6320784

Sun, W., Cai, Z., Li, Y., Liu, F., Fang, S., & Wang, G. (2018). Data processing and text mining technologies on electronic medical records: a review. *Journal of healthcare engineering*. Doi: 10.1155/2018/4302425

Vijayakrishnan, R.S., & Stewart, W.F. (2014). Prevalence of Heart Failure Signs and Symptoms in a Large Primary Care Population Identified Through the Use of Text and Data Mining of the Electronic Health Record. *Journal of Cardiac Failure*, 20(7), 459–464. Doi: 10.1016/j.cardfail.2014.03.008

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. *En Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, 29–39.