ISSN 1870-4069

Minería de textos para el análisis del subregistro de lesiones por causa externa en el servicio de urgencias de un hospital de tercer nivel

José Ramón Consuelo Estrada[1,2], Otniel Portillo Rodríguez[2], Laura Soraya Gaona Valle[1]

[1] Instituto de Salud del Estado de México, Centro Médico "Lic. Adolfo López Mateos", México

[2] Universidad Autónoma del Estado de México, Facultad de Ingeniería, Ciudad de México, México

jramon208@gmail.com, oportillor@uaemex.mx, gaonav_81@yahoo.com.mx

Resumen. Las Lesiones por Causa Externa (LCE) representan un serio problema de salud, sin embargo el diagnóstico mediante códigos inespecíficos puede subestimar su gravedad. Esto es común en muchos hospitales, afectando la asignación de recursos y prioridades en salud. Analizando datos de un servicio de urgencias de un hospital de tercer nivel, se implementaron cuatro modelos predictivos: regresión logística con texto (TF-IDF), regresión logística, árboles de decisión y boosting para determinar el porcentaje de diagnósticos de *dolor agudo* que pudieran ser LCE. El método más exacto fue el basado en texto y estima que, 12240 (82.56 % n=14826) de los *dolores agudos* son LCE. El porcentaje de LCE subestimado como resultado del uso códigos inespecíficos es alto y la minería de textos es una opción viable para su estimación.

Palabras clave: Minería de textos, machine learning, código inespecífico, lesiones por causa externa.

Text Mining to Analyze Subrecord of External Causes of Injuries in the Emergency Department of a Third Level Hospital

Abstract. External Cause of Injuries (ECI) represent a serious health problem, however diagnosis using nonspecific codes may underestimate their severity. This situation is common in many hospitals, affecting the resources allocation and health priorities. Analyzing data from the emergency department of a hospital, four predictive models were implemented: Text-Logistic regression (TF-IDF), logistic regression, decision

trees and boosting to determine the percentage of *acute pain* diagnoses that could be classified as ECI. The text-based method shows better accuracy and estimates that 12240 (82.56% n = 14826) of acute pain are ECI. The percentage of LCE underestimated by using nonspecific codes is high and text mining is a viable option for their estimation.

Keywords: Text mining, machine learning, non-specific code, external cause of injuries.

1. Introducción

Las Lesiones por Causa Externa (LCE) contribuyen aproximadamente con el 10 % de la mortalidad mundial y con el 12 % de la morbilidad [1]. Estas lesiones engloban situaciones entre las que podemos encontrar los accidentes de tránsito, las caídas, las agresiones interpersonales dentro y fuera del hogar o golpes de cualquier tipo. Cada una de ellas puede representar problemas serios a nivel mundial para los sistemas de salud y para la ciudadanía en general [2].

El Catálogo Internacional de Enfermedades (CIE-10) se ha desarrollado como herramienta para la clasificación de diagnósticos y de las causas que los generan [3]. Ciertos diagnósticos dentro del catálogo están asociados a lo que se denomina *Causa Externa* la cual se refiere a un evento que generó el padecimiento y que requiere el registro de los datos relacionados (lugar de ocurrencia, atención prehospitalaria, agente de la lesión, datos del agresor, etc). En la Figura 1 se muestra el proceso general de registro en los servicios de urgencias, donde se puede apreciar que, cuando el diagnóstico está relacionado con una LCE, entonces se requiere realizar el registro de los datos de la lesión.

Dentro del propio CIE-10 se ha identificado un grupo de diagnósticos que no proporcionan la información suficiente para su posterior análisis estadístico o epidemiológico. A estos diagnósticos se les ha denominado *inespecíficos* o, incluso *códigos basura* [4]. Esta situación también puede presentarse debido al diseño de los Sistemas de Información (SI) cuando los usuarios tienen la posibilidad de elegir opciones tales como *Otros*, *se desconoce*, *no disponible*, etc. dentro de una lista desplegable y no se solicita la aclaración pertinente; o cuando se permite que se guarde información con datos faltantes.

El problema de códigos inespecíficos se ha estudiado desde distintos puntos de vista, Pérez-Nuñez et al. [4] proponen el uso de modelos de imputación simple [5] para identificar muertes que pudieran ser atribuidas al tránsito. Sus resultados muestran un subregistro del 18.85 % y presentan una lista de diagnósticos CIE-10 que pueden ser considerados como inespecíficos. Bhalla et al. [1] realizan una evaluación de la calidad y disponibilidad de los datos sobre muertes causadas por lesiones, donde un registro menor al 20 % de causas inespecíficas se considera bueno. Se encontró que solo 20 de los 83 países analizados presentan datos de calidad. En el trabajo presentado por Híjar et al. [6] se presentan tres modelos basados en el método proporcional, imputación múltiple y regresiones en un

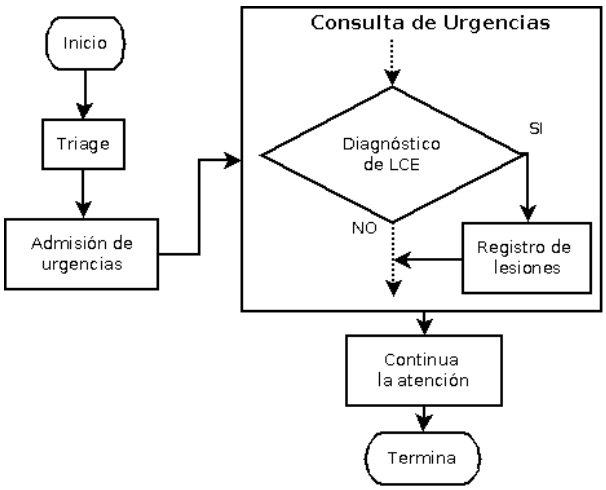

Fig. 1. Registro de lesiones en el servicio de urgencias.

intento por corregir la subestimación de muertes por tránsito debida a los códigos inespecíficos. Los autores reportan un incremento del 18 al 45 % dejando clara la magnitud del problema.

Los enfoques tradicionales para abordar el tema de la inespecificidad de los datos se basan en análisis de un conjunto de datos formado por variables cualitativas y/o cuantitativas (aquí: datos estructurados). Sin embargo, gran parte de la información hospitalaria disponible se encuentra en texto proveniente de hojas escritas por los médicos, normalmente en formato libre, ya sea manual o electrónico.

En este trabajo se propone el análisis de datos en formato de texto para abordar el problema de códigos inespecíficos. Diversas investigaciones han demostrado que el análisis de texto puede ayudar a la solución de problemas de administración hospitalaria, Lucini et al. [7] utilizaron distintas metodologías de minería de textos para entrenar modelos que permitan predecir futuras hospitalizaciones y costos de atención. Los modelos utilizados incluyen: árboles de decisión, regresión logística y máquinas de soporte vectorial, entre otras. Los autores concluyen que la minería de textos puede brindar información valiosa para la toma de decisiones.

Por otro lado, en el campo de la farmacovigilancia Abacha et al. [8] proponen el uso de minería de texto para analizar artículos científicos y registros médicos que ayuden a comprender y dar soporte al tema del suministro de medicamentos.

Kocbek et al. [9] presentan un sistema de minería de textos que utiliza información de radiología, patología y datos del paciente y de su ingreso para detectar distintos tipos de cáncer tales como el de pulmón, pecho y colon, entre otros. Concluyen que la combinación de distintas fuentes de información y el uso de minería de textos pueden mejorar las predicciones de estos padecimientos.

En nuestro caso de estudio, los datos disponibles muestran que las LCE tienen una tendencia a la baja mientras que los registros del diagnóstico inespecífico *dolor agudo* (R52.0) aumentan en proporciones similares (ver Figura 2), si bien los datos del Instituto Nacional de Estadística y Geografía (INEGI) muestran una tendencia a la baja en accidentes de tránsito [10], otros informes reportan aumentos en casos de: lesiones no intencionales, agresiones interpersonales caídas o accidentes laborales [2], [11]. Incluso la Secretaría de Salud reporta incremento en egresos hospitalarios por lesiones causadas por el tránsito, en copilotos y motociclistas, con un aumento del 62.90 % de 2010 a 2014 [12].

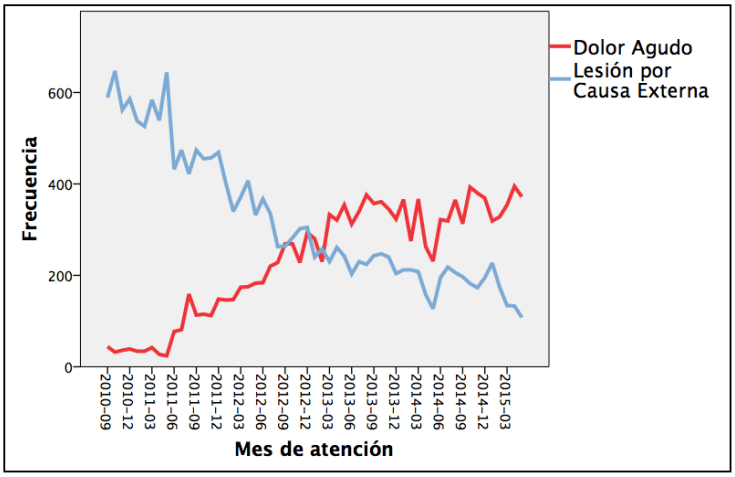

Fig. 2. Tendencia de los registros en el periodo analizado.

Este trabajo pretende estimar el porcentaje de registros con diagnóstico de dolor agudo pudieran ser LCE y hacer una comparación entre los métodos estructurados contra los basados en el análisis de textos.

2. Material y métodos

Previa aprobación ética, se extrajeron los registros electrónicos del servicio de urgencias de un hospital de tercer nivel ubicado en la ciudad de Toluca, Estado de México. La información disponible corresponde al periodo comprendido entre el 1 septiembre de 2010 al 31 de mayo de 2015 y proviene de las áreas de Triage, Admisión y Consulta, todas del servicio de urgencias y del área de Expediente Clínico que es común a todo el nosocomio.

El registro de los diagnósticos asignados a los pacientes se realizó en base al CIE-10. Utilizando el diagnóstico se definió una etiqueta con la cual se determina si el diagnóstico es Dolor Agudo (DA), Lesión por Causa Externa (LCE) o cualquier otro diagnóstico (OTRO), la Tabla 1 muestra la distribución resultante.

Tabla 1. Distribución de los datos según la etiqueta asignada.

Etiqueta	n	%
Lesión por Causa Externa (LCE)	19230	16.13 %
Dolor Agudo (DA)	15859	13.30 %
Otros diagnósticos (OTRO)	84131	70.57 %
Total	119220	100.00 %

El conjunto de datos resultante cuenta con variables estructuradas tales como: edad y sexo del paciente, escolaridad, si presenta discapacidad (ya sea mental o física), fecha de la atención, peso, talla, frecuencia cardiaca, presión arterial (sistólica y diastólica), frecuencia respiratoria, área a la que se envía al paciente después de la atención, diagnóstico (CIE-10), procedimiento realizado, medicamentos prescritos, área de atención (Consulta de urgencias, Consulta de Traumatológia de urgencia, Observación o Choque), antecedentes de Diabetes Mellitus, antecedente de alcoholismo, antecedente de tabaquismo, índice de Glasgow y forma de arribo al hospital, entre otras.

Así mismo, cada registro cuenta con variables en texto libre relativas a campos como: motivo de la consulta, antecedentes, padecimiento actual, exploración física, tratamiento y plan, observaciones, estudios realizados y pronóstico. Para el análisis propuesto, estas variables fueron concatenadas y utilizadas como un solo campo de texto.

Tabla 2. Resumen de modelos generados.

Modelo	Variables	Método	Entrenamiento	Validación	Prueba	Clasificación
M1	EST	RL	28648	3666	3526	13297
M2	EST	DT	28648	3666	3526	13297
M3	EST	Boosting	28648	3666	3526	13297
M4	TFIDF	RL	30498	3880	3738	14826

EST.- Datos estructurados; RL.- Regresión Logística, DT.- árbol de decisión; TFIDF.- Frecuencia inversa de palabras

Realizando una separación de variables se crearon dos conjuntos uno para el análisis estructurado y otro para el análisis de texto. El primer modelo M1 se basa en datos estructurados utilizando Regresión logística con regularización Ridge Regression (RR) conforme a la ecuación 1 [13], así como árboles de decisión y método de boosting, con restricciones de profundidad e iteración (número de clasificadores simples), respectivamente. Para el último modelo se realizó un análisis de texto utilizando el método conocido como Frecuencia de Término - Frecuencia de Documento Inverso (TF-IDF, por sus siglas en Inglés) [14]. Para la

elección de parámetros, cada modelo fue sometido a un proceso de validación. La Tabla 2 resume los modelos generados donde la columna *Clasificación* se refiere a la cantidad de registros de *dolor agudo* que son clasificados en la parte final del trabajo:

$$\beta_0^*, \beta^* = argmin_{\beta_0, \beta b} \left(\frac{1}{N} \sum_{k=1}^{N} (y_i - (\beta_0 + x_i \beta))^2 + \lambda \beta^T \beta \right), \tag{1}$$

donde la expresión *argmin* significa "los valores β_0 y β que minimizan la expresión". β_0 es el término *intersección* y β representan el vector de coeficientes. y_i es el vector de etiquetas conocidas. La expresión $(\beta_0 + x_i \beta)$ es el valor calculado por el modelo y λ es el parámetro de regularización. La diferencia $y_i - (\beta_0 + x_i \beta)$ es el error entre el término esperado y el término calculado. Ajustando el valor de λ, la ecuación 1 busca encontrar los parámetros β_0 y β que minimicen el error de predicción [13].

Ambos conjuntos de datos requirieron preprocesamiento. En el conjunto de datos estructurados se eliminan registros duplicados o sesgados. Los datos nulos fueron tratados con imputación por promedio para las variables numéricas y a las variables categóricas se les agregó la categoría *dato no registrado*. Las variables numéricas se estandarizaron en base a la ecuación 2 mientras que las variables categóricas fueron dicotomizadas:

$$S_i' = \frac{S_i - \mu}{\sigma}, \tag{2}$$

donde S' es el nuevo valor estandarizado para el *i-ésimo* registro, S_i es el valor actual, μ es la media del conjunto de datos y σ su desviación estándar. Con esta ecuación se busca homogeneizar los valores para evitar que ciertas variables tengan más peso que otras como resultado de las diferencias en sus escalas de medición.

Por otro lado en el conjunto con datos de texto, para eliminar palabras mal escritas se formó un listado tomando como base notas médicas, un diccionario de la lengua española y un glosario de términos médicos [15]. También se eliminaron palabras poco relevantes para el análisis (que, por, de, el, etc.). Todas las palabras se consideraron en mayúsculas y sin acentos. Para este análisis se utilizó el método TF-IDF que se basa en la ecuación 3:

$$F_i(p) = \log \left(\frac{N}{1+n} \right), \tag{3}$$

donde la frecuencia inversa de la palabra p es el logaritmo del total de registros N entre *1* mas el número de registros n que contienen la palabra p [16] [17]. Con esto se forma un conjunto estructurado con el que posteriormente se entrenó, validó y probó un modelo de regresión logística.

Los registros con diagnóstico de *dolor agudo* fueron separados para ser clasificados al final; constan de 13297 registros, para datos estructurados y 14826 para texto. El entrenamiento, validación y prueba se realizó con los datos etiquetados como LCE y una parte proporcional de los etiquetados como OTROS,

ambos conjuntos fueron balanceados utilizando el método *submuestreo* [18]. El total de registros fue de 35580 y 38116 para datos estructurados y de texto respectivamente. Finalmente, los conjuntos de datos resultantes se dividieron para entrenamiento (80 %), validación (10 %) y prueba (10 %) de los modelos. (ver Tabla 2).

Para cada modelo se realizó un ajuste de parámetros que varía según el clasificador: a los modelos M1 y M4, basados en regresión logística, se les aplicó regularización RR para evitar el sobre ajuste de los coeficientes. Los modelos M2 y M3 variaron en profundidad de árbol y número de árboles, respectivamente. Para estos ajustes se utilizó el conjunto de datos de entrenamiento y de validación [13]. El método de *boosting*, del modelo M3 esta basado en una colección de clasificadores básicos combinados a través de una técnica conocida como gradient boosting [19].

Utilizando los parámetros encontrados en la fase de validación, se clasifican los datos de prueba para realizar la evaluación de los modelos. Finalmente, se realiza la clasificación de los datos de dolor agudo, separado previamente.

Para la programación, tratamiento de datos y entrenamiento de los modelos se utilizó Python 2.5 y GraphLab Create[19].

3. Resultados

En esta sección se muestran los resultados de los análisis realizados, se describe cada modelo en el orden presentado en la Tabla 2. El rendimiento de todos los modelos se muestra en la Tabla 5 y fue calculado utilizando el conjunto de datos de prueba.

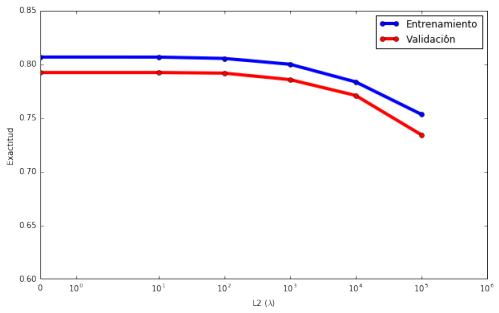

Fig. 3. Exactitud del modelo M1 con diferentes valores de regularización λ.

El modelo M1 se basa en el conjunto de datos estructurados y se realiza una regresión logística que incluye el uso de la ecuación 1 para la validación mediante una regularización RR. La gráfica de la Figura 3 muestra la exactitud para cada valor de λ tanto de los valores de entrenamiento como de validación. En este

proceso se busca que el valor de λ maximice la exactitud del conjunto de datos de validación.

Tabla 3. Variables con coeficientes más representativos del modelo M1, $\lambda = 100$.

Variable	Coeficiente	Descripción
proc.3809	10.0069	Incisión de venas, miembros inferiores
proc.8149	8.3752	Otra reparación de tobillo
med.02608	8.3734	Carbamazepina
med.00569	7.3166	Nitroprusiato de sodio
med.05506	7.2910	Celecoxib
med.03253	-6.2843	Haloperidol
proc.9393	-6.0935	Métodos de resucitación no mecánicos
proc.9908	-5.9974	Transfusión de expansor sanguíneo
med.03631	-5.9859	Solución de glucosa
med.01901	-5.7695	Sulfadiazina

proc.- Procedimiento médico, med.- Medicamento.

En este caso hay poca variabilidad en la exactitud del modelo (ver Figura 3). El valor de λ que minimiza el error y que por lo tanto aumenta la exactitud es 100 (10^2); este valor fue utilizado para entrenar la versión final del modelo M1. Los coeficientes del modelo resultante con pesos mas representativos (positivos y negativos) son los mostrados en la Tabla 3.

En la Tabla 3 los prefijos *proc* y *med*, hacen referencia al procedimiento aplicado al paciente y a los medicamentos prescritos, ambos valores basados en sus respectivos catálogos. Debido a que el nombre de la variable es largo y para mejorar la lectura, se han acotado al prefijo *proc* y *med* más la clave correspondiente. El comportamiento de estos coeficientes a diferentes valores λ se muestra en la Figura 4.

El modelo M2 también utiliza datos estructurados para entrenar un árbol de decisión. En este caso el parámetro a considerar es la profundidad del árbol. En la Figura 5 se puede apreciar que el mejor valor de profundidad es 10. Con este valor se realiza el entrenamiento final del modelo y posteriormente se realiza la prueba.

El modelo M3 utiliza un conjunto de árboles (boosting). En este modelo el parámetro a encontrar es la cantidad de árboles o *iteración* que conformarán el modelo. En la Figura 6 se muestra sus respectivos valores de exactitud a diferente número de árboles. Aunque se puede apreciar que al aumentar indefinidamente el número de árboles la exactitud en el conjunto de entrenamiento se hace casi perfecta, también se aprecia que la exactitud en la validación disminuye a partir del valor 50, por lo que se ocupa este como parámetro del modelo.

Finalmente, para el modelo M4 una vez realizada la representación del texto con el método de TF-IDF, se calcula la inversa mediante la ecuación 3. Posteriormente se entrena un modelo de regresión logística donde el proceso es similar al

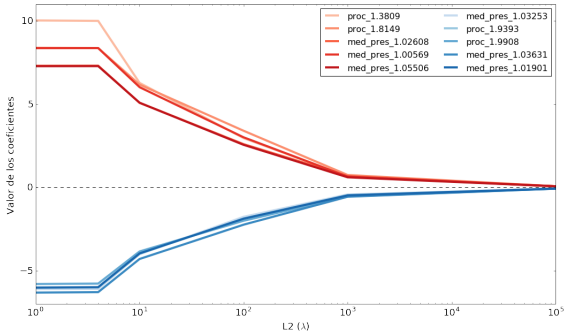

Fig. 4. Comportamiento de los coeficientes del modelo M1 a distintos valores de regularización λ.

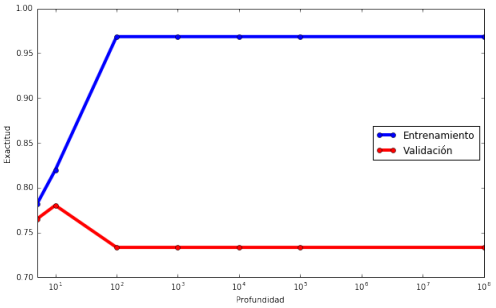

Fig. 5. Exactitud del modelo M2 con diferentes valores de profundidad del árbol de decisión

del modelo M1. En la Figura 7 se muestran los valores de la exactitud según varia el valor del parámetro λ. En este caso el valor que presenta el mejor rendimiento en el conjunto de validación es con $\lambda = 1000(10^3)$.

A diferencia del modelo M1, en este caso el lugar de las variables es tomado por las palabras utilizadas en el análisis. Las que tienen coeficientes más representativos se muestran en la Tabla 4 (Las palabras fueron analizadas en mayúsculas y sin acentos).

Los resultados muestran que los modelos implementados obtienen valores de evaluación superiores al 80.00 %. La Tabla 5 presenta los resultados de las pruebas de rendimiento de los modelos donde se puede apreciar que el modelo M4 de análisis de texto tiene los valores más altos.

La parte final del trabajo consiste en utilizar los modelos implementados como resultado del proceso de validación y prueba para clasificar los registros de atenciones médicas a pacientes que fueron diagnosticados con dolor agudo. En la columna *clasificación* de la Tabla 2, se muestra la cantidad de datos estructurados y de texto a clasificar. Los datos son sometidos a cada uno de los modelos según su tipo de dato para finalmente obtener el porcentaje de registros

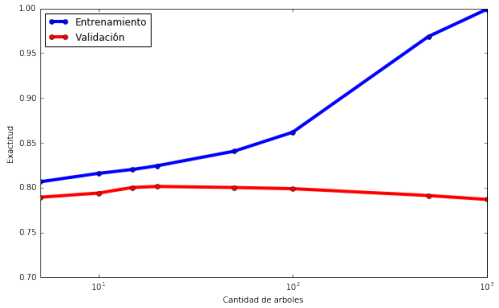

Fig. 6. Exactitud del modelo M3 para distintos valores de regularización

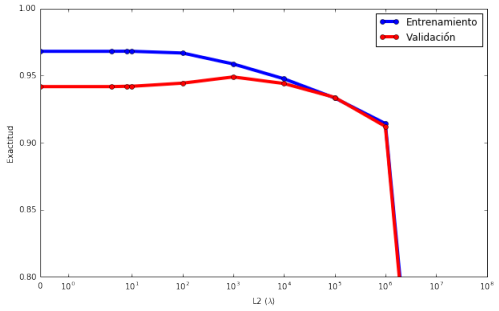

Fig. 7. Exactitud del modelo M4 con diferentes valores de regularización λ.

que pertenecen a LCE. Los resultados de este análisis se muestran en la Tabla 6 y en la gráfica de la Figura 8.

4. Discusión

El problema de los registros con valores inespecíficos es común a distintas áreas. En el sector salud se puede presentar en registros de mortalidad, de natalidad, de LCE o de prevalencia y/o incidencia de enfermedades, solo por mencionar algunas. Este trabajo se enfoca en el subregistro de LCE como consecuencia del código CIE-10 del dolor agudo, sin embargo, otros padecimientos pueden verse afectados por este o por algún otro código inespecífico [4].

Los datos disponibles para este trabajo muestran que en el periodo analizado, el 13.30 % de los registros totales en el servicio de urgencias del hospital son diagnósticos de dolor agudo. Si se compara con el 16.13 % que corresponde a LCE en el mismo periodo, se puede considerar como factor importante en las estadísticas que se reportan estatal y nacionalmente. Los resultados muestran que en promedio 83.10 % de los registros de dolor agudo son LCE lo que aumentaría su prevalencia del 16.13 % al 25.66 % del total de atenciones en el servicio de urgencias del hospital analizado.

Tabla 4. Palabras con coeficientes más representativos del modelo M4 con $\lambda = 1000$.

Palabra	Coeficiente
Privación	0.5067
Minimizar	0.4201
Difteria	0.3803
Refuerza	0.3747
Arteriovenoso	0.3392
Alcali	-0.3454
Escotoma	-0.3140
Preparan	-0.3110
Leucemias	-0.2905
Purgante	-0.2692

Tabla 5. Desempeño de los modelos contra los datos de prueba.

Modelo	Exactitud	F1-score	Precisión	Recall	AUC
M1	0.8117	0.8184	0.7941	0.8442	0.8758
M2	0.8001	0.8006	0.8026	0.7985	0.8702
M3	0.8142	0.8181	0.8054	0.8313	0.8878
M4	0.9393	0.9392	0.9538	0.9250	0.9729

Gran parte de la información hospitalaria se encuentra en formato no estructurado y suele dejarse fuera del análisis de datos tradicional (con la pérdida de información que esto implica) sin embargo, los resultados obtenidos en este trabajo sugieren que la minería de textos puede ser de utilidad en este tipo de situaciones ya sea como complemento a las técnicas tradicionales o como una alternativa confiable cuando no existen datos estructurados suficientes y de calidad.

La dicotomización de las variables categóricas crean conjuntos de datos con altas dimensiones, por tal motivo solo se presentan los coeficientes de las variables más relevantes, y se realiza una regularización RR para evitar sobre ajuste (overfitting), sin embargo, no se realizaron regularizaciones para selección de variables. Así mismo, para mejorar la velocidad de entrenamiento, validación y prueba, en este trabajo se utilizaron conjuntos de validación fijos y seleccionados al azar, no obstante, para mejorar la confiabilidad de los resultados conviene realizar mas pruebas de entrenamiento-validación, por ejemplo utilizando el método de validación cruzada.

Normalmente, este tipo de estudios se realiza considerando los conjuntos de entrenamiento, validación y pruebas. En este trabajo también se muestran los resultados de datos reales con diagnósticos desconocidos y se utilizan para hacer

Tabla 6. Clasificaciones realizadas con datos reales de dolor agudo.

Modelo	LCE(%)	OTRO(%)
M1	11172(84.02 %)	2125(15.98 %)
M2	10848(81.58 %)	2449(18.42 %)
M3	11199(84.22 %)	2098(15.78 %)
M4	12240(82.56 %)	2586(17.44 %)

LCE.- Cantidad de registros clasificados como Lesión por Causa Externa, OTRO.- Cantidad de registros clasificados como no Lesión por Causa Externa.

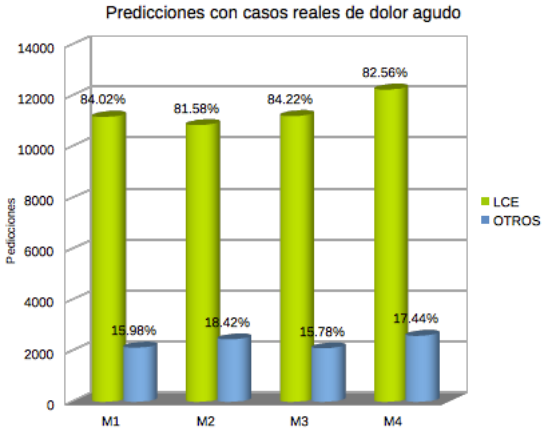

Fig. 8. Resultados de la clasificación de dolores agudos.

estimaciones de los porcentajes de LCE registradas como dolor agudo, como se puede apreciar en la gráfica de la Figura 8.

5. Conclusiones

Se presentan cuatro modelos con distintas técnicas de clasificación: 1) regresión logística con datos estructurados, 2) árboles de decisión, 3) el método boosting y 4) regresión logística con TF-IDF para el análisis de datos en formato de texto. Las pruebas realizadas y la comparación entre modelos muestran que la minería de textos es una alternativa confiable para abordar el tema de los registros inespecíficos, particularmente del subregistro de diagnósticos relativos a LCE.

Además del método tradicional de entrenamiento, validación y prueba de los algoritmos, también se utilizan los modelos validados y probados para la

clasificación de un conjunto de datos del que no se conocen las etiquetas y que representan casos reales de inespecificidad. Dada la similitud de los resultados obtenidos se puede presumir cierto grado de confianza, sin embargo, se recomienda que un porcentaje de dichos registros sean analizados y clasificados por personal experto para comparar resultados.

Considerando a la prevención como una mejor alternativa al remedio, se sugiere la revisión de catálogos y SI para evitar el uso de valores inespecificos o en su defecto, cuando la situación lo requiera, pedir al usuario una aclaración de dicho valor. Sin embargo hay que considerar que los servicios de urgencias suelen ser caóticos y se que se cuenta con poco tiempo para atención al paciente, por lo tanto los SI tendrán que ser fíciles, rápidos y asegurar la calidad de la información.

Como trabajo futuro se pretende analizar el porcentaje de diagnósticos de dolor agudo que se clasifica como LCE mediante los modelos implementados con intención de estimar a que tipo de causa externa pertenece (Accidente de tránsito, caídas, agresiones, etc.) y obtener resultados más específicos.

Agradecimientos. El presente trabajo se realizó con el apoyo del Centro Médico "Lic. Adolfo López Mateos"del Instituto de Salud del Estado de México y con la participación de la Dirección de Posgrado de la Facultad de Ingeniería de la Universidad Autónoma del Estado de México.

Referencias

1. Bhalla, K., Harrison, J., Shahraz, S., Fingerhut, L.: Availability and quality of cause-of-death data for estimating the global burden of injuries. Bulletin of the World Health Organization 88:831–838C (2010), doi: 10.2471/blt.09.068809.
2. Instituto Nacional de Salud Pública: Las lesiones por causa externa en México. Lecciones aprendidas y desafíos para el Sistema Nacional de Salud. INSP, Ciudad de México/Cuernavaca(MX) (2010)
3. Clasificación Internacional de Enfermedades (CIE-10) C OPS/OMS Colombia - OPS/OMS Colombia. En: Paho.org (Consultado 4 Apr 2017)
4. Pérez-Núñez, R., Mojarro-Íñiguez, M., Mendoza-García, M., et al.: Subestimación de la mortalidad causada por el tránsito en México: análisis subnacional. Salud Pública de México 412–420 (2016), doi: 10.21149/spm.v58i4.8021.
5. Royston, P.: Multiple imputation of missing values. Stata journal 4:227–241 (2004)
6. Híjar, M., Chandran, A., Pérez-Núñez, R. et al.: Quantifying the Underestimated Burden of Road Traffic Mortality in Mexico: A Comparison of Three Approaches. Traffic Injury Prevention 13:5–10 (2012), doi: 10.1080/15389588.2011.631065.
7. Lucini, F. S., Fogliatto, F. C., da Silveira, G. et al.: Text mining approach to predict hospital admissions using early medical records from the emergency department. International Journal of Medical Informatics 100:1–8 (2017), doi: 10.1016/j.ijmedinf.2017.01.001.
8. Ben Abacha, A., Chowdhury, M., Karanasiou, A. et al.: Text mining for pharmacovigilance: Using machine learning for drug name recognition and drug-drug interaction extraction and classification. Journal of Biomedical Informatics 58:122–132 (2015), doi: 10.1016/j.jbi.2015.09.015.

9. Kocbek, S., Cavedon, L., Martinez, D. et al.: Text mining electronic hospital records to automatically classify admissions against disease: Measuring the impact of linking data sources. Journal of Biomedical Informatics 64:158–167 (2016), doi: 10.1016/j.jbi.2016.10.008.
10. Accidentes de tránsito terrestre en zonas urbanas y suburbanas. En: Inegi.org.mx. http://www.inegi.org.mx/est/contenidos/proyectos/registros/economicas/accidentes/ (Consultado 2 Abril 2017)
11. Baldeón Calisto, M.: Análisis estadístico de accidentalidad laboral del Ecuador y comparación con la accidentalidad laboral de Colombia del año 2013. Maestría, Universidad San Francisco de Quito, Colegio de Postgrados; Quito, Ecuador (2014)
12. Secretaría de Salud/STCONAPRA: Informe sobre la Situación de la Seguridad Vial, México 2015. Secretaría de Salud, Ciudad de México (2017)
13. Bowles, M.: Machine learning in Python. 1st ed., John Wiley & Sons, Indianapolis (2015)
14. Feldman, R., Sanger, J.: The text mining Handbook. 1st ed. University Press, Cambridge (2007) ,
15. Glosario de Terminos Medicos. En: Dra Laura Cartuccia. https://auditoriamedica.wordpress.com/2009/05/24/glosario-de-terminos-medicos (Consultado 3 Abril 2017)
16. Aizawa, A.: The Feature Quantity: An Information Theoretic Perspective of Tfidf-like Measures. In: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, pp. 104–111 (2000)
17. Joachims, T.: A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization. In: Proceedings of ICML-97, 14th International Conference on Machine Learning, pp. 143–151 (1997)
18. Haibo, H., Garcia, E.: Learning from Imbalanced Data. IEEE Transactions on Knowledge and Data Engineering 21:1263–1284 (2009), doi: 10.1109/tkde.2008.239
19. GraphLab Create API Documentation GraphLab Create API 1.10 documentation. In: Turi.com. https://turi.com/products/create/docs/index.html (Consultado 4 Abril 2017)