# Redes Neuronales para la identificación de competencias legisladas

Carlos Gabriel Oblitas Villegas
*Postgrado en Informática*
*Universidad Mayor de San Andrés*
La Paz - Bolivia
oblitasss@gmail.com

***Resumen*—El modelamiento de tópicos es un campo del Procesamiento del Lenguaje Natural, que permite categorizar textos en distintos temas (tópicos), mediante algoritmos de manera supervisada y no supervisada. En la presente investigación se tratará la categorización de la normativa boliviana de los gobiernos autónomos (GA), de acuerdo a las competencias asignadas a los GA por la Constitución y la Ley Marco de Autonomías. De esta manera se utilizó word2vec para identificar el espacio semántico de los textos a clasificar y los algoritmos LDA y LSTM para el modelado de tópicos. Para mejorar los resultados obtenidos por el LDA, se buscó el número óptimo de tópicos, y se utilizó el algoritmo LDA Guiado basado en los datos de una clasificación manual previa, para extraer información significativa que permita hallar la competencia legislada en la normativa boliviana. Además, este estudio realizó una evaluación de la precisión de los algoritmos utilizados para realizar esta clasificación.**

***Palabras clave—NLP, Text mining, doc2vec, LDA, LSTM***

## I. INTRODUCCIÓN

La anterior Constitución Política del Estado (CPE) otorgaba la potestad de legislar únicamente al nivel nacional a través del Congreso de la República. Sin embargo, a partir del 7 de febrero de 2009, la actual Constitución de Bolivia otorga, en su Artículo 272°, la facultad de legislar a todos los Gobiernos Autónomos (GA), a través de sus Órganos Deliberativos, quienes pueden ahora establecer normas de acuerdo a las competencias que les fueron asignadas, mediante la nueva CPE y la Ley Marco de Autonomías y Descentralización.

A nueve años de la vigencia de la nueva Constitución, la cantidad de normativa promulgada por los 348 GA (337 Municipios, 9 Departamentos y 3 Autonomías Indígenas), es extensa y aumenta dramáticamente cada día. Un control importante previsto en la nueva Constitución, para evitar que ocurran conflictos en la normativa de los diferentes GA, es la asignación competencial, sin embargo, es necesario que toda la normativa sea analizada para evaluar el progreso que tienen los GA en normar las competencias que les fueron designadas. En este artículo, se expone el trabajo de análisis de la normativa promulgada por los GA, utilizando técnicas de minería de textos e Inteligencia Artificial, para evaluar el grado de avance de los GA en su desarrollo.

El Servicio Estatal de Autonomías (SEA), tiene la facultad de acuerdo al artículo 129 PIII, de la Ley Nacional 031, de administrar un registro de normas emitidas por las entidades territoriales autónomas y por el nivel central del Estado, en relación con el régimen autonómico. Por lo que haciendo uso de los datos públicos de dicha entidad se analizaron 17.835 normas, las cuales incluyen información de las competencias legisladas, y el GA. La definición de competencia según el SEA es: “Las Competencias son la titularidad de atribuciones ejercitables respecto de las materias determinadas por la Constitución Política del Estado y la Ley Marco de Autonomías y Descentralización” [1].

La importancia del análisis de la normativa en función de sus competencias, se basa en que, el ejercicio de la autonomía depende fundamentalmente del ejercicio efectivo de las competencias por parte del nivel central del Estado y las entidades territoriales autónomas, pues de esta manera la asignación competencial se convierte en política pública y por ende, en beneficios para la población.

## II. MARCO TEÓRICO

El Procesamiento del Lenguaje Natural - PLN es definido como: "un rango de técnicas computacionales motivadas teóricamente para analizar y representar textos naturales en uno o más niveles de análisis lingüístico con el fin de lograr un procesamiento de lenguaje similar al humano para una variedad de tareas o aplicaciones" [2]. En esta definición, se puede encontrar que el objetivo a lograr es el “procesamiento de lenguaje similar al humano”, introduciendo el hecho que las raíces del PLN, se encuentran en la Inteligencia artificial. Actualmente para el PLN existen algunos algoritmos basados en Inteligencia Artificial que serán utilizados en la presente investigación.

### *A. Paragraph Vector (doc2vec)*

El algoritmo doc2vec es una red neuronal basada en word2vec, el cual permite encontrar la representación de palabras en un vector de n-dimensiones, de acuerdo a las frecuencias y un análisis del contexto (a partir de las cercanías de palabras), lo cual permite mapear la semántica de palabras, en un espacio n-dimensional, haciendo posible posteriormente utilizar operaciones como la distancia del coseno para encontrar sus relaciones semánticas. De acuerdo a Mikolov [3], doc2vec es “un algoritmo no supervisado que aprende características de longitud fija a partir de fragmentos de textos de longitud variable, como oraciones, párrafos y documentos”, que a diferencia de word2vec, aprende las características a partir de palabras en una ventana fija.



**Para referenciar este artículo (IEEE):**
[N] C. Oblitas, «Redes Neuronales para la identificación de competencias legisladas», *Revista PGI. Investigación, Ciencia y Tecnología en Informática*, nº 8, pp. 41-44, 2020.

### B. Latent Dirichlet Allocation (LDA)

El modelado de temas (*topic modeling*) descubre automáticamente los temas ocultos de los documentos dados. El enfoque del modelado de temas considera cada documento como una colección de temas y cada tema como una colección de palabras clave, una aproximación inicial a los temas, sería que con solo mirar las palabras clave, debería poderse identificar de que trata el tema.

Una variación de este algoritmo (Guided LDA), es la incorporación de "palabras semilla" (*seed words* ó *prior words*) a los tópicos, con lo cual es posible guiar al algoritmo a encontrar ciertos tópicos buscados, pues el LDA, encuentra tópicos sin supervisión. De acuerdo a Andrzejewski and Zhu [4], estas palabras semillas son identificadas de forma manual mediante la sistematización realizada por un experto, sin embargo en esta investigación seleccionaremos priors de acuerdo a las fórmulas de Ganancia de Información y Entropía de acuerdo a lo sugerido por Jagadeesh [5].

### C. Long-Short Term Memory (LSTM)

Esta red neuronal es un tipo de Red Neuronal Recurrente, con la diferencia que las redes recurrentes, van olvidando en su estado la historia más lejana (problema del *vanishing gradient*), por lo cual este tipo de red, permite recordar en su entrenamiento, en su estado actual historias lejanas. Los LSTM están diseñados explícitamente para evitar el problema de dependencia a largo plazo. Recordar información durante largos períodos de tiempo es prácticamente su comportamiento predeterminado.

De acuerdo con Olah [6]LSTM puede entenderse considerando tratar de predecir la palabra en el texto "Crecí en Francia, hablo ------- con fluidez". La información reciente sugiere que la palabra a predecir, es probablemente el nombre de un idioma, pero si queremos deducir a qué idioma se refiere, necesitamos considerar el contexto: "Francia", que está en un estado anterior. Es completamente posible que la brecha entre la información relevante y el punto donde se necesita sea muy grande. Los LSTM están diseñados explícitamente para evitar el problema de dependencia a largo plazo.

## III. METODOLOGÍA

Dado que el objetivo del estudio es clasificar la normativa de acuerdo a las competencias legisladas por los Gobiernos Autónomos, se recurrirá a un diseño no experimental que se aplicará de manera longitudinal, a través de las gestiones, a partir del 2010.

La investigación no experimental es aquella que se realiza sin manipular deliberadamente variables. Es decir, es investigación donde no hacemos variar intencionalmente las variables independientes. Lo que hacemos en la investigación no experimental es observar fenómenos tal y como se dan en su contexto natural, para después analizarlos. Como señala Kerlinger: "La investigación no experimental o expost-facto es cualquier investigación en la que resulta imposible manipular variables o asignar aleatoriamente a los sujetos o a las condiciones" [7]. De hecho, no hay condiciones o estímulos a los cuales se expongan los sujetos del estudio. Los sujetos son observados en su ambiente natural, en su realidad.

Para el presente estudio se procesó la información textual proveniente de las normas recopiladas en formato PDF por el SEA de los GA a partir del 2010, a fin de categorizar de acuerdo a las competencias legisladas de cada una de ellas, teniendo como base el análisis previo llevado adelante por los abogados de dicha entidad (alrededor de 7 mil normas analizadas), es importante mencionar que dicho análisis es costoso en tiempo pues se requiere la lectura de cada ley para su análisis. Para la clasificación de la situación semántica de una competencia, se aplicaron los algoritmos de LDA y LDA Guiado para encontrar la asociación de documentos. Además, este estudio realizó una evaluación del desempeño de la precisión del modelado de temas, basado en redes neuronales LSTM y su eficiencia, examinando la efectividad del método propuesto. Los pasos seguidos en la presente investigación fueron:

a) Conversión de los PDFs de normas a Imágenes por página (Wand y PIL).
b) Optimización de imagen de cada hoja de los PDFs y aplicación de filtro binario, para separar el fondo del texto.
c) Aplicación de OCR sobre las imágenes procesadas, para la generación de textos en un archivo Pickle (PyTesseract).
d) La etapa principal de pre procesamiento, el tokenizado, la eliminación de stopwords y la lematización del texto. (NLTK, Gensim).
e) Generación de un diccionario con palabras utilizadas.
f) Identificación de tópicos con LDA y pruebas de coherencia.
g) Identificación de palabras clave, para LDA Guiado y pruebas de coherencia.
h) Implementación de red neuronal LSTM, para categorización de competencias y pruebas de exactitud.

Los anteriores pasos se encuentran enmarcados en el proceso estándar existente para la minería de textos CRISP-DM, que divide el proceso de Descubrimiento de Conocimiento (KD) en cinco partes: comprensión de datos, selección de datos, limpieza de datos, modelado de datos y evaluación [8].

## IV. RESULTADOS

### A. LDA

Un primer acercamiento al algoritmo LDA, fue la generación de tópicos y palabras clave para 274 tópicos, a fin de poder enlazar posteriormente cada tópico con las competencias, sin embargo, se pudo ver que la identificación de un tópico y asignación a una o varias competencias, era un proceso manual muy complicado pues muchos de los tópicos tienen palabras duplicadas, con espacios semánticos muy similares. En la Tabla 1 se puede ver palabras clave identificadas de los 274 tópicos, en las que se puede encontrar palabras comunes (como gestión) por lo que se encuentran en espacios semánticos muy cercanos.

TABLA I. TÓPICOS PALABRAS SIMILARES (7 PALABRAS POR CADA TÓPICO).

| Tópicos | Keywords |
|---|---|
| 1 | Anual, operativo, plan, presupuestar, gestión, gasto, según… |
| 18 | Documentar, detallar, consignar, aprobar, anual, operativo, gestión… |
| | …… |
| 57 | Propiedad, gestión, proponente, habilitar, gabinete, Teófilo, reconstrucción… |
| 145 | Cualquier, propiedad, gestión, organigrama, habilitar, recontratación, subalcaldesa… |
| 274 | Presupuestaria, ley, programar, traspasar, anual, gestión, incrementar… |

De los tópicos anteriores se puede ver que los mismos tienen en muchos casos palabras similares o duplicadas, por lo que la identificación incluso por parte de un ser humano es muy difícil, ya que todos ellos se encuentran en espacios semánticos muy cercanos, lo que se pudo evidenciar al hacer una visualización en dos dimensiones mediante pyLDAvis (Figura 1).

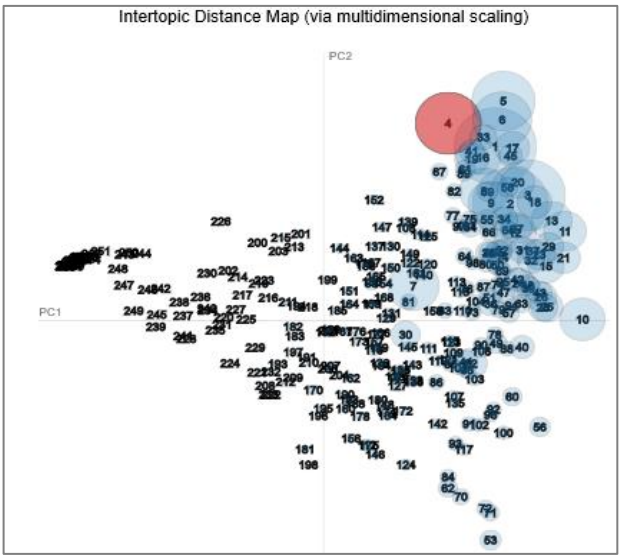


Fig. 1. Espacio semántico para 277 tópicos

De acuerdo a la figura 1 y la Tabla 1 se puede evidenciar que los espacios semánticos son muy cercanos y en muchos casos sobrepuestos, lo cual es demostrado en las palabras similares encontradas en varios tópicos, por lo que se buscó hacer correr el algoritmo LDA, de 2 a 280 tópicos, y encontrar el mejor índice de coherencia para la cantidad de tópicos, de acuerdo a los textos de las normas procesados. El índice de coherencia, sirve para indicar que tan seguro se encuentra el modelo para realizar sus predicciones, y de acuerdo a lo encontrado la mejor coherencia para la cantidad de textos procesados de las normas, es de 277 tópicos, sin embargo, existe una gran cantidad de clusters muy cercanos, con pocas coincidencias, lo cual puede evidenciarse en los espacios semánticos graficados.

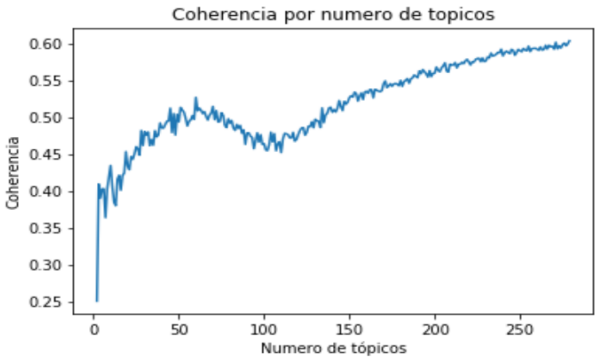


Fig. 2. La mejor coherencia está en los 277 tópicos

## B. LDA Guiado

Al observar que la coherencia del modelo LDA para 274 tópicos aún era baja y que además los tópicos encontrados no representaban las competencias sobre las cuales se requiere categorizar las normas, se decidió utilizar LDA Guiado, para encontrar las competencias de mejor manera. Para esta tarea, se utilizó la entropía y ganancia de información basados en la matriz de ocurrencias de las normas, de acuerdo a lo recomendado por Jagadeesh [5].

Para la matriz de ocurrencias se generó 4452 columnas (el vocabulario sin stopwords) y 274 filas (las competencias), donde en cada celda se colocaba la cantidad de veces que la palabra aparece en todas las normas que legisla sobre cierta competencia. Del procesamiento de dicha matriz, se consiguió armar un listado de palabras clave inicial (Tabla 2), que fue a alimentar el algoritmo LDA Guiado, con las palabras semilla encontradas.

TABLA II. ALGUNOS TÓPICOS IDENTIFICADOS MEDIANTE ENTROPÍA Y GANANCIA DE INFORMACIÓN

| Topic | Keywords |
|---|---|
| 1 | Ademaf, administrativa, anh, comando, delincuencia, administrativa, … |
| 2 | Arquitectura, huerta, momento, Pérez, agricultor, bimestral, convención, … |
| 4 | Artesanía, censo, conamaq, despoblamiento, confederación, núcleo,… |
| 5 | Atendió, burocracia, defunción, demográfico, divorcio, matrimonio… |
| 6 | Tic, electromagnéticas, radiodifusión, radio, telefonía, antena, soporte, … |
| | … |

A partir, de las palabras semilla introducidas en el algoritmo LDA, se pudieron identificar de mejor manera las competencias, sin embargo, al momento de realizar la prueba de coherencia del modelo, se vio que la perplejidad bajo, del 47% usando LDA, al 45% usando las palabras clave identificadas en el algoritmo LDA Guiado.

## C. LSTM

Para el caso de la categorización de 274 competencias de la normativa analizada, se desarrolló un modelo de red neuronal LSTM Multivariable, procesando el vocabulario de entrada por identificador de palabra (id). Se analizó todas las entradas y los párrafos de entrada ya etiquetados con las competencias, correspondientes a 7144 ejemplos, haciendo el mismo pre procesado de tokenización, eliminación de stopwords y lematización, llegando a un vocabulario de 4452 palabras. Se hizo un barrido para determinar la cantidad máxima de palabras por ejemplos, a fin de determinar el tamaño máximo de entrada a la Red Neuronal, obteniendo un máximo de entrada de 76 palabras de solo corpus. A partir de esto se generó una matriz de 7144 x 76, como entrada X y una matriz de 7144 x 274, como ejemplos etiquetados para la matriz Y.

Mediante el algoritmo LSTM, utilizando *categorical_crossentropy*, con 70% de los ejemplos para training y 30% para test, se alcanzó un 91% de *accuracy* en 100 *epoch* (Figura 3).

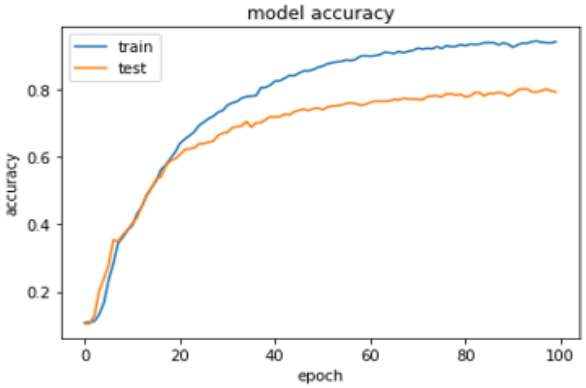


Fig. 3. La curva de exactitud alcanza al 90% en 100 epochs

## V. Discusión

De acuerdo a los resultados obtenidos por los algoritmos (Tabla 3), el mejor algoritmo para el análisis de las normas en base a las 274 competencias del presente artículo, fue la red neuronal LSTM con un 91% de precisión. Una de las razones para este resultado, fue la gran cantidad de tópicos a catalogar (274 competencias), esto se puede evidenciar en el hecho de que normalmente las palabras clave para los tópicos a categorizar son sugeridas por un experto en el área a modelar [4], sin embargo el hacerlo para el presente estudio sería una labor muy ampulosa para un experto, pues la elección de palabras para los tópicos sería muy compleja, y más considerando que se recomienda no repetir palabras para cada tópico [5], es por esa razón que se utilizó la ganancia de información, como el método de identificar palabras clave, sobre la matriz de ocurrencias.

TABLA III. Resultados alcanzados con los algoritmos utilizados

| Algoritmo | Resultados |
|---|---|
| LDA | 49% |
| Guided LDA | 45% |
| LSTM | 91% |

Una de las razones de estos resultados es que los datos de entrada son relativamente bajos a los requeridos por algoritmos que usan vectorización de palabras, para la identificación de espacios semánticos, como lo son LDA y el LDA Guiado.

En base a los resultados obtenidos en el presente estudio el Servicio Estatal de Autonomías registrará el análisis de normas de acuerdo a la información generada por el modelo desarrollado y se modificará también el registro normativo, incluyendo el texto de la norma, a fin de evitar errores al momento de procesar las normas mediante OCR. Además, con esta información sistematizada, se espera que se empiecen a realizar análisis sobre las competencias legisladas, a fin de conocer los avances en el desarrollo de los GA, pues las competencias son materializadas en la prestación de servicios a la sociedad, mediante el establecimiento de políticas públicas en la normativa.

## Referencias

[1] S. Durán, «Cartilla Ejercicio de Competencias,» Servicio Estatal de Autonomías, La Paz, 2012.

[2] D. Ferati, «Text Mining in Financial Industry Implementing Text Mining Techniques on Bank Policies,» Utrecht, Netherlands, Utrecht University, 2017.

[3] Q. Le y T. Mikolov, Distributed Representations of Sentences and Documents.

[4] D. Andrzejewski y X. Zhu, Latent Dirichlet Allocation with Topic-in-Set Knowledge, Wisconsin: Association for Computational Linguistics, 2009.

[5] J. Jagadeesh, D. I. Hal y U. Raghavendra, Incorporating Lexical Priors into Topic Models, Maryland: Association for Computational Linguistics, 2012.

[6] C. Olah, «Christopher Olah's Blog,» GitHub Pages, 27 August 2015. [En línea]. Available: https://colah.github.io/posts/2015-08-Understanding-LSTMs/. [Último acceso: 15 10 2019].

[7] Kerlinger, La investigación del comportamiento, México: Interamericana, 1979.

[8] R. Wirth, CRISP-DM: Towards a Standard Process Model for Data, Manchester, UK: Conference on the Practical Applications of Knowledge Discovery and Data Mining, 2000.

Breve CV del autor

**Carlos Gabriel Oblitas Villegas** es Licenciado en Informática por la Universidad Mayor de San Andrés (La Paz, 2005). Actualmente realiza la Maestría de Gerencia Estratégica de Sistemas de Información GETSI del Postgrado en Informática UMSA.
Ejerce profesionalmente en la Unidad de Análisis Económico Financiero del Servicio Estatal de Autonomías.
Publicó el artículo "Redes Neuronales para la identificación de competencias legisladas", PGI-Review Nro 7, Postgrado en Informática UMSA, La Paz 2019.
Sus intereses investigativos incluyen Ciencia de Datos, Aprendizaje de Maquina y Procesamiento del Lenguaje Natural.
Email: oblitasss@gmail.com.