Procesamiento de Lenguaje Natural: Herramienta para evaluar la compatibilidad de las decisiones corporativas con la legislación financiera colombiana

Elquin Cáceres Pineda

Universidad EAN

Facultad de Administración, Finanzas y Ciencias Económicas

Maestría en Gestión Financiera

Bogotá D.C., Colombia

18 de Octubre de 2022

Procesamiento de Lenguaje Natural: Herramienta para evaluar la compatibilidad de las decisiones corporativas con la legislación financiera colombiana

Elquin Cáceres Pineda

Trabajo de grado presentado como requisito para optar al título de:

Magister en Gestión Financiera

Director:

Luz Adriana Pineda Baron

Modalidad:

Monografía:

Universidad EAN

Facultad de Administración, Finanzas y Ciencias Económicas

Maestría en Gestión Financiera

Bogotá D.C., Colombia

18 de Octubre de 2022

Nota de Aprobación:

_____

_____

_____

_____

_____

_____

Firma Jurado No. 1

_____

Firma Jurado No.2

_____

Firma Jurado No.3

_____

Firma Director Trabajo de Grado

Bogotá D.C. Colombia, _____ de _____ de _____

Declaratoria

"It is strange that only extraordinary people make discoveries that then appear easily and simply"

Georg Lichtenberg

Procesamiento de Lenguaje Natural: Herramienta para evaluar la compatibilidad de las decisiones corporativas con la legislación financiera colombiana

Agradecimientos

Un trabajo de investigación es siempre fruto de ideas, proyectos y esfuerzos previos que corresponden a otras personas. A todas esas personas gracias por publicar sus conocimientos y un especial agradecimiento a la profesora Luz Adriana Pineda, por su confianza, tiempo y apoyo.

Procesamiento de Lenguaje Natural: Herramienta para evaluar la compatibilidad de las decisiones corporativas con la legislación financiera colombiana

Resumen

La estrategia corporativa es una herramienta capaz de generar valor empresarial, por tanto, las acciones oportunas en situaciones cambiantes representan un desafío, cuando el contexto además incluye un flujo creciente de nueva regulación. Considerando esta preocupación en el contexto empresarial colombiano y en la creciente regulación financiera actual, se aborda crear un método automatizado fundamentado en el modelo de procesamiento de lenguaje natural RoBERTa, en representaciones vectoriales del texto, en los conceptos de recuperación de información y en las métricas de similitud, para desarrollar una metodología que permite medir la similitud entre una decisión corporativa y su ordenamiento jurídico local. Al poner a prueba el enfoque propuesto, se observó que es posible crear representaciones semánticas del lenguaje jurídico local, con el que se pueden calcular valores de compatibilidad entre las decisiones corporativas y sus normas asociadas, puntajes que al verificarlos manualmente y de forma detallada resultan validos. Se concluye que la metodología propuesta es eficiente para una verificación propia de las acciones corporativas sin requerir la intervención de un experto legal, además se observó que estos resultados pueden robustecerse, incluyendo nuevas técnicas para mejorar los casos de extremos particulares que se encontraron durante el análisis de las observaciones iniciales obtenidas del conjunto de datos de verificación.

Palabras Clave: *PNL, Decisiones Corporativas, Ordenamiento jurídico, Similitud Semántica, Codificadores dobles, Codificadores cruzados*

Procesamiento de Lenguaje Natural: Herramienta para evaluar la compatibilidad de las decisiones corporativas con la legislación financiera colombiana

Abstract

The corporate strategy is a tool capable of generating business value, therefore, timely actions in changing situations represent a challenge, when the context also includes a growing flow of new regulations. Considering this concern in the Colombian business context and in the current growing financial regulation, it is addressed to create an automated method based on the ROBERTa natural language processing model, on vectorial representations of the text, on the concepts of information retrieval and on the metrics of similarity, to develop a methodology that allows measuring the similarity between a corporate decision and its local legal system. When testing the proposed approach, it was observed that it is possible to create semantic representations of the local legal language, with which values of compatibility between corporate decisions and their associated norms can be calculated, scores that, when verified manually and in detail, are valid. It is concluded that the proposed methodology is efficient for a proper verification of corporate actions without requiring the intervention of a legal expert, it was also observed that these results can be strengthened by including new techniques to improve the cases of particular extremes that were found during the analysis of the initial observations obtained from the verification data set.

Keywords: *NLP, Corporate Decisions, Legal System, Semantic Similarity, Bi-Encoders, Cross Encoders*

Índice

Índice de Tablas

Índice de figuras

Algoritmos Python

1. Introducción

En el ámbito corporativo, la estrategia tiene un efecto simultáneo en el relacionamiento y en el desarrollo de los negocios, en consecuencia, tomar acciones estratégicas correctas aumenta el valor empresarial y contribuye con el mejoramiento del entorno organizacional. Cada acción estratégica es el resultado de una decisión que puede afectar las operaciones, los objetivos y las actividades futuras (Stagner, 1969), al comprender su importancia y los diferentes tipos, es posible adoptarlas correctamente en diversas situaciones.

Dichas decisiones pueden ser de diversos tipos, por ejemplo; organizacionales, políticas, operativas, rutinarias, entre otras (Kownatzki y col., 2013; Lim & Chung, 2021). Todas estas, tienen un contexto, un ámbito de aplicación, un proceso para su determinación y un nivel de importancia (Belkaoui & Karpik, 1989). En este sentido, las Políticas contables, las financieras, las de riesgos, las comerciales, los planes de negocio, las ideas de nuevos productos, los planes de expansión, la modificación de la estructura corporativa y administrativa, la incursion en otros mercados, son claros ejemplos de dichas acciones, con un denominador común, cada una de ellas debe hacer parte de un documento empresarial (Cummings y col., 2002).

Para tomar una decisión corporativa que aporte valor, es de especial importancia, considerar el orden jurídico aplicable, ya que toda acción empresarial debe guardar correspondencia con la legislación de su entorno organizacional (Post, 2003). En este aspecto, el creciente flujo de documentos legales representa un desafío para la toma de decisiones, pues requiere conocimientos de dominio relacionados con la profesión legal.

Con un número de leyes cada vez mayor (Berger-Walliser & Scott, 2018; Seltzer y col., 2022), es difícil, para las instituciones, abordar oportunamente acciones estratégicas que hagan frente a un entorno empresarial en constante cambio (Silva, 2021), si la evaluación legal se limita exclusivamente al personal profesional del derecho en la forma tradicional.

A la anterior situación se adicionan los elevados costos de la consultoría legal corporativa (Aziz y col., 2021; von Philipsborn y col., 2022), dando lugar a cuestionarse sobre ¿cómo medir de

manera automatizada la similitud o compatibilidad de una decision corporativa con la legislación?, de tal manera que sea posible para una empresa actuar de manera rápida y asertiva considerando siempre su marco legal. Hallar una respuesta permitiría a cualquier empresa generar valor de forma eficiente, observando diligentemente los valores de la responsabilidad corporativa (Khaled y col., 2021).

En esta investigación, se aborda esta pregunta aplicándola al contexto colombiano y en particular a aquellas decisiones corporativas que precisan de alguna regulación financiera local para su ejecución, lo anterior, considerando el creciente número de startups y su emergente propensión al riesgo legal (Oliva y col., 2022), por ello, se aborda crear un método automatizado con base en los modelos de procesamiento de lenguaje natural, utilizando técnicas de representación vectorial de oraciones, técnicas de similitud y conceptos de recuperación de información, de tal manera que se pueda comprender el contexto normativo local y las decisiones corporativas numéricamente, para finalmente obtener un puntaje de compatibilidad entre ellas.

En las siguientes secciones de este documento se presentan apartados esenciales para este propósito como; la Metodología, donde se presenta el Enfoque General Propuesto para obtener el puntaje de compatibilidad entre una decisión corporativa y la regulación financiera colombiana, que se fundamenta en la Arquitectura del Modelo RoBERTa, el proceso realizado durante la Recolección de Información, los Principales Resultados y su Comparación con una evaluación no automatizada, los elementos de Discusión de la metodología propuesta y las Conclusiones y Trabajo Futuro sugerido.

2. Objetivos

2.1. Objetivo General

Determinar el grado de compatibilidad entre una decisión corporativa y la legislación financiera colombiana, mediante el uso de un método cuantitativo que se apoya en un modelo probabilístico para el procesamiento de lenguaje natural.

2.2. Objetivos Específicos

- Recopilar información de la legislación financiera colombiana y de diferentes decisiones corporativas, mediante una búsqueda automatizada en fuentes primarias.
- Representar la estructura semántica del texto contenido en la información recopilada, a través de un modelo de procesamiento de lenguaje natural.
- Presentar los resultados más relevantes obtenidos por el modelo de procesamiento de lenguaje natural, comparándolos con una evaluación no automatizada.

3. Justificación

La motivación de desarrollar el planteamiento de este trabajo se apoya en responder a la necesidad planteada en la sección 1Introducción, y de ello, su valor para la industria y las instituciones que la componen, al presentar un método automatizado de análisis que permite medir el grado de compatibilidad de las acciones corporativas con la legislación financiera colombiana, lo cual, aporta eficiencias en los procesos estratégicos de las organizaciones al tiempo que puede reducir significativamente los costes de la consultoría legal corporativa.

Para el contexto actual de las empresas colombianas y en particular para el creciente número de startups, resulta especialmente útil el enfoque propuesto en esta investigación, ya que, introducir análisis sistematizados, mejora las capacidades operativas para generar valor en las organizaciones, al tiempo que mitiga sus riesgos legales emergentes.

Además, se encuentra alineado con el enfoque empresarial de la universidad EAN, al tiempo que mantiene el espíritu investigativo de las líneas y campos de investigación de esta institución (Emprendimiento y Gerencia - Diseño estratégico). Se concluye que su valor teórico se centra en el uso de herramientas de última generación, aplicadas a solventar problemáticas reales y brindar alternativas útiles para las empresas Colombianas.

4. Marco Teórico

El uso de técnicas de procesamiento de lenguaje natural y otros campos de las ciencias de la computación y la inteligencia artificial en las profesiones del derecho, existen desde la década de 1960 cuando surgieron los primeros sistemas para buscar contenido legal, y aunque se fundamentaban en estructuras construidas a partir de un complejo conjunto de reglas diseñadas manualmente (Hahn, 1998). Entre 1970 y 1980, la empresa estadounidense Lexis Nexis fue pionera en la prestación de servicios de investigación jurídica, introduciendo el primer terminal del mundo que conectaba a las firmas de abogados con las bases de datos de derecho y jurisprudencia de algunas bibliotecas. Inicialmente, búsquedas de texto completo de la jurisprudencia de Ohio y Nueva York (Dale, 2019).

A partir de 1980, estos sistemas tuvieron algunas mejoras al introducirse los primeros algoritmos de aprendizaje automático, como los árboles de decisión, con estructuras de sentencias (si - entonces), muy similares a las reglas escritas manualmente (Mandal y col., 2017). Desde entonces se ha avanzado mucho en lo que se refire a tecnología aplicada al ámbito jurídico y en los conceptos bajo los cuales se denomina.

4.1. Legaltech la Tecnología Aplicada al Derecho

En materia conceptual desde 2017 han aparecido los conceptos *legaltech y Lawtech* que se utilizan según el caso y el contexto (Dubois, 2021). *Legaltech* comúnmente se entiende como el uso de la tecnología para brindar servicios legales (Munisami, 2019; Soukupova, 2021; Szostek, 2021). Por lo que se podría definir como el uso de la tecnología en servicios legales orientados a:

- Reducir o eliminar la necesidad de acudir al sector legal de forma tradicional.
- Acelerar los trámites y la gestión de tareas de los propios abogados, reduciendo el coste y el tiempo que un abogado debe invertir en sus tareas.
- Simplificar el contacto entre los profesionales del derecho y los potenciales clientes.

Lawtech se utiliza para describir varios tipos de tecnologías que tienen como objetivo apoyar, complementar o reemplazar los métodos tradicionales para brindar servicios legales, o que mejoran la forma en que opera el sistema de justicia (Webley y col., 2019), cubriendo una amplia gama de herramientas y procesos, tales como:

- Automatización de documentos.
- Chatbots y gestores de consultas.
- Contratos legales inteligentes
- Sistemas de gestión del conocimiento.

A pesar de que la validez y la generalización de cada concepto se discute ampliamente (Ashley y col., 2001; McGinnis & Pearce, 2019; R. Susskind, 2008; R. E. Susskind & Susskind, 2015), debido a que ambos reflejan evidentes similitudes, varios investigadores consideran que *Legaltech* sería el término apropiado para referirse a tecnologías aplicadas a la profesión del derecho (Salmerón-Manzano, 2021), ya que describe las actividades del sector legal, al igual que *RegTech* la tecnología que ayuda a cumplir con la regulación, *InsurTech* servicios de seguros con base tecnológica (Gramegna & Giudici, 2020), o *FinTech* finanzas y tecnología para acelerar la digitalización e inclusión del sector financiero y asegurador (Rundo y col., 2019).

4.2. Evolución del Procesamiento de Lenguaje Natural

Por su parte, la aplicación de tecnología al ámbito jurídico llegó de la mano de múltiples investigaciones para representar el lenguaje humano (Collins y col., 2017; Lehnert, 1977, 1981; Pazzani, 1983; Schank & Abelson, 1975), que originalmente se enfocaban en tareas de traducción automática. Hacia finales de la década de 1980, la mayoría de estudios en procesamiento de lenguaje, se centraron en modelos estadísticos, capaces de generar mejores representaciones del lenguaje y de tomar decisiones probabilísticas (Chowdhary, 2020; Liberman, 1991).

Durante la década de 1990, estos métodos puramente estadísticos, fueron esenciales para mantener el ritmo del enorme flujo de texto en línea. Los *N-Grams*[1] que son un tipo de modelo probabilístico para predecir el siguiente elemento en una secuencia, en forma de una cadena de Markov de orden $(n-1)$, se volvieron útiles, reconociendo y rastreando grupos de datos lingüísticos de forma numérica.

Al inicio de esta década, derivadas de las redes neuronales *feed-forward*[2] (Goldberg, 2016; Rumelhart y col., 1986), aparecieron modelos de redes neuronales recurrentes (RNN) como la que se describe en la figura 1(a), capaces de resolver ciertas tareas eficientemente (Schmidhuber, 1993). Sin embargo, a estas redes de impulso infinito les lleva demasiado tiempo aprender a almacenar información en intervalos de tiempo prolongados a traves del algoritmo de propagación hacia atrás (Leung & Haykin, 1991), la razón de esto es que los gradientes para optimizarlas, tienden a crecer o a desvanecerse con el tiempo, debido a que estos no dependen únicamente del error presente sino también de los pasados (Hochreiter & Schmidhuber, 1997).

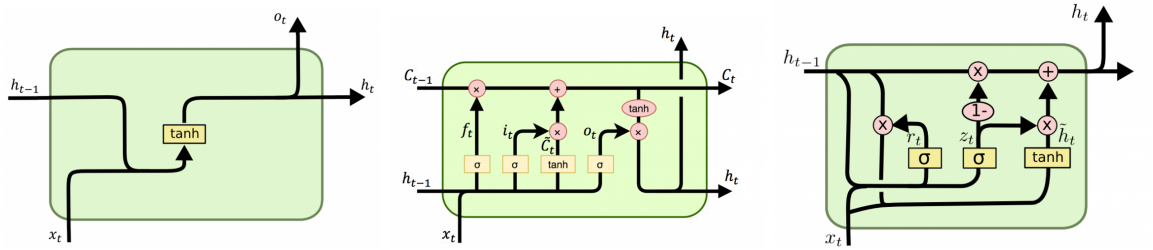
Para resolver este problema, en 1997 se introdujeron las redes de memoria a corto y largo plazo *"Long short - term memory - (LSTM)"* (Leung & Haykin, 1991), como la descrita en la figura 1(b), permitiendo que los gradientes fluyan sin cambios. Aunque estas solamente satisfacen el evento en que el gradiente converge a cero, dejando abierta la posibilidad a que este crezca infinitamente (Calin, 2020). A pesar de ofrecer avances relevantes en el campo, las RNN, sólo fueron especialmente relevantes hasta 2007, gracias a su capacidad para procesar secuencias temporales, se popularizaron en aplicaciones para el reconocimiento de voz, reconocimiento de patrones de texto y síntesis de texto a voz.

Posteriormente, en 2014 se introdujeron las unidades recurrentes cerradas *"Gated recurrent units - GRU"* (Cho y col., 2014) descritas en la figura 1(c),que son una variación de las LSTM con menor complejidad, ya que carecen de una puerta de salida (Gers y col., 2000), por lo que, pueden

[1] Un n-grama es una subsecuencia de elementos en una secuencia. Es usado en el estudio del lenguaje natural para construir los n-gramas sobre la base de distintos tipos de elementos como por ejemplo fonemas, sílabas, letras, palabras o subpalabras.

[2] Una red feedforward es la forma más sencilla de una red neuronal. En ella, la información se mueve en una única dirección, Desde los nodos de entrada, a través de los nodos ocultos y hacia los nodos de salida.

facilitar la captación de dependencias sin ignorar la información pasada de fragmentos de datos secuenciales, logrando en algunos casos generar resultados superiores (Gruber & Jockisch, 2020). Cabe mencionar que este tipo de morfologías, ha contribuido positivamente con el objetivo inicial, es decir, la calidad de los textos traducidos. La figura 1 Redes Neuronales Recurrentes (RNN), ilustra en detalle la forma más básica de los tres tipos de redes neuronales recurrentes mencionadas previamente.



(a) RNN (b) LSTM (c) GRU

$$h_t = \sigma_h(i_t) = \sigma_h(U_h x_t + V_h h_{t-1} + b_h)$$

$$y_t = \sigma_y(a_t) = \sigma_y(W_y h_t + b_h)$$

(d) Función RNN

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$$

$$\tilde{C}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c)$$

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t$$

$$h_t = o_t \odot \tanh(C_t)$$

(e) Función LSTM

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t] + b_z)$$

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t] + b_r)$$

$$\tilde{h}_t = \tanh(W_h \cdot [r_t \odot h_{t-1}, x_t] + b_h)$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t$$

(f) Función GRU

Figura 1: Redes Neuronales Recurrentes (RNN)

Nota: Adaptado de "A hybrid forecasting model using LSTM and Prophet for energy consumption with decomposition of time series data", (p. 1001) Arslan, 2022, PerlJ.

- ***RNN:*** *x_t: vector de entrada $(m \times 1)$, h_t : vector de capa oculta $(n \times 1)$, o_t : vector de salida $(n \times 1)$, b_h : vector de sesgo $(n \times 1)$, U, W : matrices de parámetros $(n \times m)$, V : matriz de parámetros $(n \times n)$, σ_h, σ_y : funciones de activación.*
- ***LSTM:*** *h_t, C_t vectores de capa oculta, x_t : vector de entrada, b_f, b_i, b_c, b_o : vector de sesgo, w_f, W_i, W_C, w_o : matrices de parámetros, $\sigma, \tanh$: funciones de activación*
- ***GRU:*** *h_t : vectores de capa oculta, x_t : vector de entrada, b_z, b_r, b_h : vector de sesgo, w_z, W_r, W_h : matrices de parámetros, $\sigma, \tanh$: funciones de activación*

Finalmente en 2014, con fundamento en las redes LSTM se plantea la arquitectura de redes neuronales *(Codificador - Decodificador)*. El codificador utiliza una LSTM para leer la secuencia de entrada, un paso de tiempo a la vez, para obtener una representación vectorial de dimensión fija, y luego el decodificador usa otra LSTM profunda para extraer la secuencia de salida. Esta segunda red es esencialmente una RNN excepto que está condicionada por la secuencia de entrada (Kalchbrenner & Blunsom, 2013; Sundermeyer y col., 2014; Sutskever y col., 2014).

Formalmente, la primera red lee una secuencia de vectores $X = (x_1, \cdots, x_{Tx})$ en un vector c con una red tal que $h_t = f(x_t, h_{t-1})$ y $c = q(\{h_1, \cdots, h_{Tx}\})$, donde $h_t \in \mathbb{R}^n$ es un estado oculto en el tiempo t, c es un vector generado a partir de la secuencia de los estados ocultos, f y q son funciones no lineales. En este caso utilizando una LSTM como f y $q(\{h_1, \cdots, h_T\}) = h_T$ (Sennrich y col., 2015).

La segunda red es entrenada para predecir la siguiente palabra $y_{t'}$ dado el vector de contexto c y todas las palabras predichas previamente $\{y_1, \cdots, y_{t'-1}\}$ (Bahdanau y col., 2014; Webber y col., 2020). En otras palabras, el decodificador define una probabilidad sobre la traducción, descomponiendo la probabilidad conjunta en los condicionales ordenados:

$$p(y) = \prod_{t=1}^{T} p(y_t | \{y_1, \cdots, y_{t-1}\}, c), \tag{1}$$

donde $y = (y_1, \cdots, y_{T_y})$. Con una red RNN, cada probabilidad condicional se modela como: $p(y_t|y_1, \cdots, y_{t-1}, c) = g(y_t, s_t, c)$, siendo g una función no lineal que genera la probabilidad de y_t, y s_t es el estado oculto de la red (Bahdanau y col., 2014).

Seguidamente en 2015, se conjetura que el uso de un vector de longitud fija es un cuello de botella para el rendimiento de esta arquitectura básica "codificador-descodificador", por lo cual, se propone dejar que el modelo busque automáticamente partes de una oración que son relevantes para predecir una palabra objetivo, sin tener que formar estas partes como un segmento rígido (Bahdanau y col., 2014).

Concretamente, se propone reemplazar el codificador por una red recurrente bidireccional

BiRNN[3] (Graves y col., 2013; Schuster & Paliwal, 1997) y redefinir la ecuación 1 del decodificador como $p(y_i|y_1, \cdots, y_{i-1}, X) = g(y_{i-1}, s_i, c_i)$ donde s_i es un estado oculto RNN para el tiempo i, calculado por $s_i = f(s_{i-1}, y_{i-1}, c_i)$ y la probabilidad está condicionada a un vector de contexto distinto c_i para cada palabra objetivo y_i y X es la primera secuencia de vectores. El vector de contexto c_i depende de una secuencia de anotaciones $(h_1, \cdots, h_{Tx})$ a las que un codificador asigna la oración de entrada, cada anotación h_i contiene información sobre toda la secuencia de entrada con enfoque en las partes que rodean la $i - esima$ palabra.

Este vector de contexto c_i calculado como una suma ponderada de las anotaciones h_i:

$$c_i = \sum_{j=1}^{T_x} \frac{exp[a(s_{i-1}, h_j)]}{\sum_{j=1}^{T_x} exp(e_{ik})} h_j. \tag{2}$$

donde $a(s_{i-1}, h_j)$ es un modelo de alineación que califica qué tan bien coinciden las entradas alrededor de la posición j y la salida en la posición i, la puntuación se basa en el estado oculto de la RNN s_{i-1} justo antes de emitir y_i en la ecuación $p(y_i|y_1, \cdots, y_{i-1}, X) = g(y_{i-1}, s_i, c_i)$ y la $j - esima$ anotación h_j de la oración de entrada y e_{ik} es la puntuación de cada una de las demás anotaciones de la secuencia.

Lo anterior, implementa un mecanismo de atención en el decodificador que le permite decidir las partes de la oración fuente a las que prestar atención, liberando al codificador de la tarea de recoger toda la información de la oración fuente en un vector de longitud fija (Bahdanau y col., 2014). Este enfoque logra un rendimiento de traducción superior al de las propuestas anteriores (Bojar y col., 2014), al tiempo que es adoptado como la tecnología central en los servicios de traducción.

Al mismo tiempo que se desarrollaron las RNN, en 1982 se presentó la primera red neuronal convolucional *"Convolutional Neural Network - CNN"* denominada *neocognitron* (Fukushima & Miyake, 1982), propuesta como mecanismo para reconocimiento de patrones visuales. Capaz de autoorganizarse mediante el aprendizaje sin un profesor, adquiriendo la habilidad de reconocer

[3]Las redes neuronales recurrentes bidireccionales (BiRNN) conectan dos capas ocultas de direcciones opuestas a la misma salida. Con esta forma de aprendizaje profundo generativo, la capa de salida puede obtener información de los estados pasados y futuros simultáneamente.

patrones de estímulo basados en la similitud geométrica sin afectarse por sus posiciones.

Esta red consta de una capa de entrada (matriz de fotorreceptores) seguida de una conexión en cascada de estructuras modulares, cada una compuesta por dos capas de celdas conectadas. La primera capa de cada módulo son células "S", que tienen características similares a las células simples, y la segunda capa consiste en células "C" similares a las células complejas, imitando el funcionamiento de las células de la corteza visual primaria de un cerebro biológico, (Gross y col., 1972; Hubel & Wiesel, 1962, 1965).

En 1998, utilizando varios mecanismos para el reconocimiento de caracteres escritos a mano, se demostró que una CNN diseñada como en la figura 2 para comprender la variabilidad de las formas 2D, superaba las demás técnicas (LeCun y col., 1998). De esta forma, se estableció un nuevo paradigma de aprendizaje, llamado redes de transformadores de gráficos (GTN) (Chellapilla y col., 2006), que permite a la red adaptarse de manera global a los múltiples módulos de los sistemas de reconocimiento de documentos de la vida real, convirtiéndose en una de las primeras redes desplegadas comercialmente (Ahlawat y col., 2020; Wu y col., 2014).

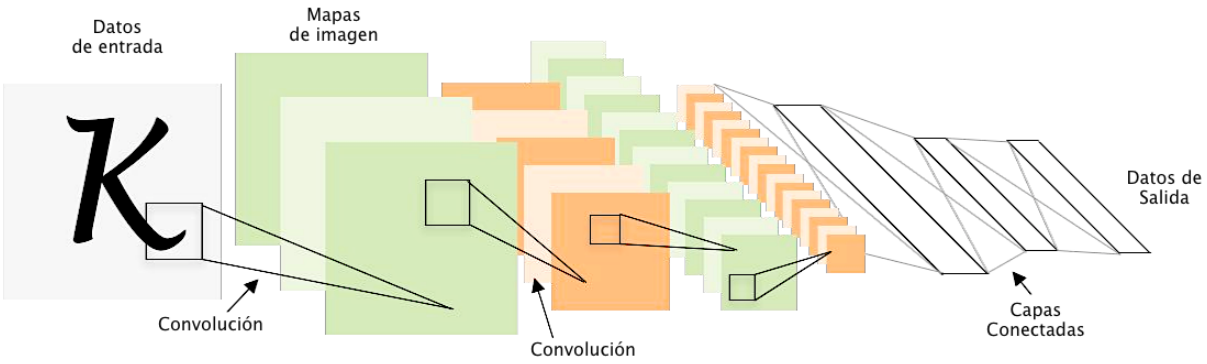


Figura 2: CNN Reconocimiento Óptico de Caracteres

Nota: Adaptado de "Comparison Of Learning Algorithms For Handwritten Digit Recognition", (p. 4) LeCun y col., 1995, Accelerating the world's research

Desde entonces, las redes CNN han sido refinadas e implementadas para entrenarse en unidades de procesamiento gráfico (GPU), convirtiéndose en el estándar para muchas tareas de

vision por ordenador y una gran cantidad de aplicaciones comerciales (dos Santos y col., 2018; Gavali & Banu, 2020; Ngo y col., 2021; Strigl y col., 2010). Lo anterior, permitió el desarrollo de modelos precisos, transferibles y eficientes para acelerar el descubrimiento y desarrollo de nuevos materiales, aplicados, por ejemplo en microscopía electrónica para la clasificación de la estructura cristalina de los materiales, como se ilustra en la figura 3, descubriendo nuevas estructuras al evaluar más de 46774 materiales (Sanyal y col., 2018; Zaloga y col., 2020).

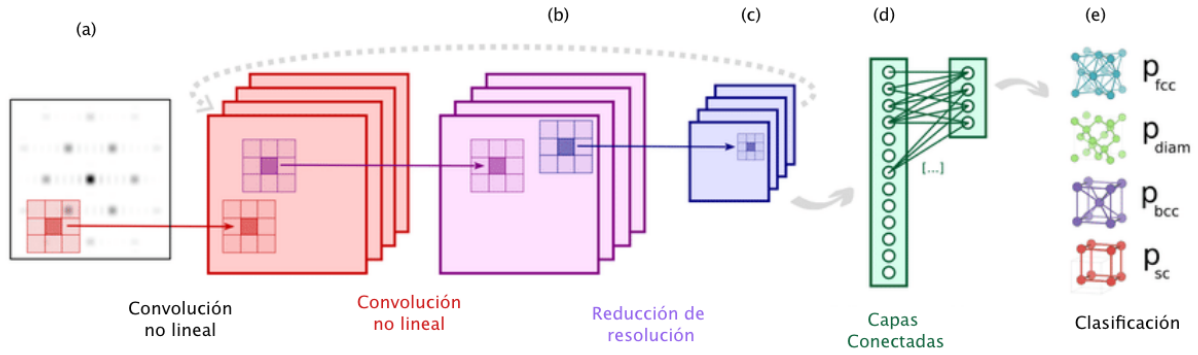


Figura 3: Ejemplo de Red CNN para Clasificación

Nota: Adaptado de "Insightful classification of crystal structures using deep learning", (p. 5) Ziletti y col., 2018, Natire Comunications

Además de esto, las CNN gradualmente han comenzado a estar presentes en el campo del procesamiento del lenguaje natural, a menudo en tareas como; el análisis de sentimiento, la recuperación de información, la clasificación de texto y documentos, y el modelado de oraciones (Dos Santos & Gatti, 2014; R. Johnson & Zhang, 2014; Shen y col., 2014; Sun y col., 2015; Weston y col., 2014; Y. Zhang & Wallace, 2015). Comúnmente con estructuras altamente eficientes y relativamente simples como la que se muestra en la figura 4. La idea fundamental en esta arquitectura es que la ventana deslizante o filtro capturará, del mismo modo que con el procesado de imágenes, características importantes del texto que luego se pueden utilizar en muchas de las tareas previamente mencionadas (Y. Liu, Fan y col., 2019; Moschitti y col., 2014).

Las anteriores aplicaciones consideran el uso exclusivo de redes CNN, sin embargo, se conocen algunos híbridos entre las RNN y las CNN que procesan secuencias de texto en las

arquitecturas de tipo codificador - decodificador (Kalchbrenner & Blunsom, 2013).

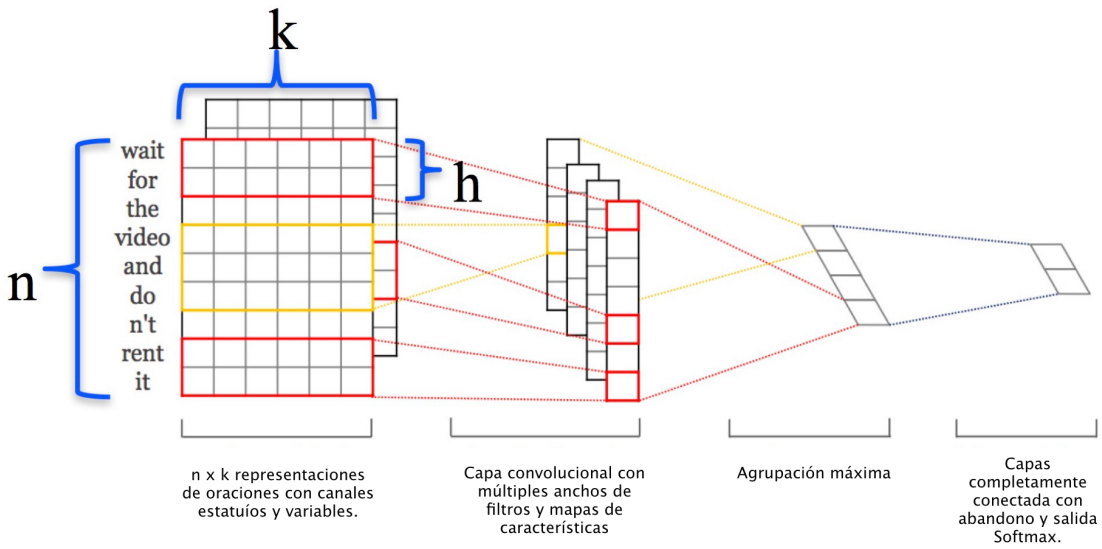


Figura 4: Red CNN para Aplicaciones de PLN

Nota: Adaptado de "A Study on Voice Command Learning of Smart Toy using Convolutional Neural Network", (p. 1211) Lee y Park, 2018, The Transactions of the Korean Institute of Electrical Engineer, Rahman y Finin, 2019

Hasta 2017, las RNN dominaban gran parte de las tareas de procesamiento de lenguaje natural, pues el lenguaje humano es precisamente una secuencia de palabras. Y estas redes se especializan en procesar este tipo de datos. Aunque eficaces en la generación de textos cortos altamente coherentes, por su memoria de corto plazo son incapaces de mantener su coherencia en secuencias extensas.

Entonces, a partir de los modelos de traducción que se fundamentan en redes CNN o RNN complejas que incluyen un codificador y un decodificador, propuestas desde 2014, se introduce la arquitectura de la figura 5 denominada Transformer o transformador de oraciones (Vaswani y col., 2017). Esta red neuronal adapta unicamente el mecanismo de atención presente en las RNN (Bahdanau y col., 2014), eliminando por completo la recurrencia y las convoluciones.

Esta red tiene una memoria de largo plazo, gracias a los mecanismos de auto-atención introducidos, al mismo tiempo, es capaz de procesar datos en paralelo, requiriendo un menor

tiempo de entrenamiento, lo que permite contraer el coste computacional de propuestas anteriores (Ghaderi, s.f.). Además, se demostró que el Transformador generaliza bien otras tareas al aplicarlo con éxito en análisis distintos a la traducción automática (Acheampong y col., 2021; Tunstall y col., 2022; Yates y col., 2021).

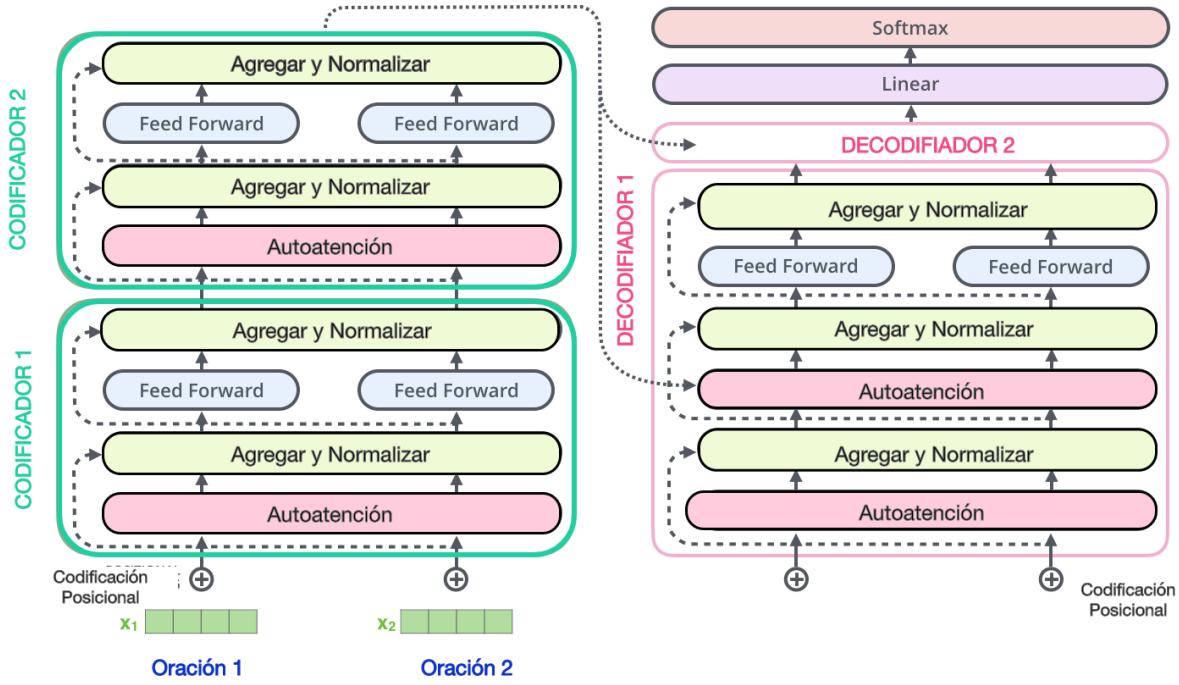


Figura 5: Arquitectura de la Red Transformer

Nota: Adaptado de "The Illustrated Transformer " Alammar, 2018, BlogPost

El Transformador utiliza capas apiladas de auto-atención totalmente conectadas entre el codificador y el decodificador, como se muestra a la izquierda y derecha de la Figura 5. El primer bloque que globalmente es un codificador, está compuesto por una pila de $N = 6$ capas idénticas, cada una con dos subcapas, (i) un mecanismo de autoatención de cabezales múltiples, y (ii) una red de retroalimentación simple completamente conectada en cuanto a la posición, adicionalmente una conexión residual alrededor de cada una de las dos subcapas (He y col., 2016), seguida de una normalización (Ba y col., 2016). Donde, la salida de cada subcapa es $LayerNorm(x + Sublayer(x))$, siendo $Sublayer(x)$ la función implementada por la propia subcapa. Finalmente,

para facilitar estas conexiones residuales, todas las subcapas y las capas incrustadas producen salidas de 512 dimensiones (Vaswani y col., 2017).

El segundo bloque que funciona como un decodificador, también se compone de una pila de $N = 6$ capas idénticas. Además de las dos subcapas en cada capa del codificador, el decodificador inserta una tercera subcapa, que calcula la atención de varios cabezales sobre la salida de la pila del codificador. De manera similar al codificador, existen conexiones residuales alrededor de cada una de las subcapas, seguidas de la normalización de capas (Ba y col., 2016; He y col., 2016), en este caso la subcapa de auto-atención esta modificada para evitar que las posiciones presten atención a las posiciones posteriores.

Este enmascaramiento, combinado con un vector de posición que compensa todas las incrustaciones de salida, asegura que las predicciones de la posición i puedan depender únicamente de las salidas conocidas sean en posiciones menores que i. Dicha codificación de la posición corresponde a una función seno $CP_{(p,2_i)} = \sin(p/10000^{\frac{2_i}{d_m}})$ para el codificador y a una función coseno $CP_{(p,2_{i+1})} = cos(p/10000^{\frac{2_i}{d_m}})$ para el decodificador, donde i es la dimensión, p es la posición y d_m son las 512 dimensiones del modelo (Gehring y col., 2017).

De este punto en adelante, dada su flexibilidad y capacidad, las redes neuronales basadas en transformers se han convertido en el estándar para la mayoría de las tareas de PNL, además de ser la precursora de diversos modelos de procesamiento de lenguaje altamente eficientes como BERT, RoBERTa, sBERT, la familia GPT y el más reciente BLOOM (Devlin y col., 2018; Lample & Conneau, 2019; Lee-Thorp y col., 2021; Y. Liu, Ott y col., 2019; Radford y col., 2018). Por lo anterior, esta es la arquitectura que se utiliza en esta investigación.

4.3. Procesamiento de Lenguaje Aplicado a la Profesión Legal

La ley tiene el lenguaje en su corazón, por lo que no sorprende que el software que procesa lenguaje natural desempeñe un papel importante en algunas áreas de la profesión legal (Dale, 2019). Pero en los últimos años se ha visto un mayor interés en aplicar técnicas modernas a una gama más amplia de problemas. Permitiendo que existan sistemas que pueden redactar

documentos legales, realizar investigaciones jurídicas, divulgar documentos en litigios, realizar procesos automáticos de debida diligencia, proporcionar orientación jurídica y resolver litigios en línea.

Esto es posible mediante tareas como; la búsqueda semántica, la clasificación de textos, el modelado de temas, la similitud textual semántica, el resumen de documentos, entre otras, derivadas de los avances en PLN mencionados en la sección 4.2 Evolución del Procesamiento de Lenguaje Natural.

Del uso de estas tareas en aplicaciones para legaltech se pueden observar diversos trabajos. Por ejemplo, en "Measuring similarity among legal court case documents" se utilizan medidas basadas en TF-IDF[4] y en similitud avanzadas como modelado de temas e incrustaciones de palabras y documentos, para calcular la similitud entre dos documentos legales para identificar precedentes relevantes para un litigio. Demostrando que el uso de incrustaciones funciona mejor que otros enfoques. (Mandal y col., 2017).

En "Effective deep learning approaches for summarization of legal texts" se proponen técnicas que utilizan redes neuronales para resumir documentos judiciales. La principal ventaja del enfoque propuesto es que no se basan en características hechas a mano, o conocimiento específico del dominio, ni su aplicación está restringida a un subdominio en particular, lo que los hace aptos para extenderse también a otros dominios. Las evaluaciones establecen una mayor efectividad en comparación con otros enfoques. (Anand & Wagh, 2019)

Para mitigar el riesgo de daños ocasionados a las empresas por litigios estratégicos, en el artículo *"A semantic analysis approach for identifying patent infringement based on a product – patent map"* los autores proponen el método semántico producto - patente basado en la similitud tecnológica sujeto - acción - objeto (SAO) para generar mapas de infracción de patentes y sugieren varios índices y métodos de subagrupación para interpretar el mapa. Particularmente, explotan datos sobre tecnología y productos relacionados con la lámpara de diodos emisores de luz (LED).

[4]TF-IDF, frecuencia de término - frecuencia inversa de documento, es una medida numérica que expresa cuán relevante es una palabra para un documento en una colección. Esta medida se utiliza a menudo como un factor de ponderación en la recuperación de información y la minería de texto.

(Park & Yoon, 2014)

Legal Judgement Prediction (LJP) aplica técnicas de procesamiento de lenguaje natural para predecir el resultado de un juicio en función de los hechos de un caso, utilizando un marco de aprendizaje de las dependencias topológicas entre las subtareas del proceso legal. Al poner a prueba este método en casos penales en el sistema de derecho civil, se obtienen mejoras consistentes sobre otros métodos que usan una única tarea para la predicción de fallos judiciales (Zhong y col., 2018).

5. Hipótesis

H1. Mediante los modelos transformer aplicados a tareas de búsqueda y similitud textual, es posible medir la compatibilidad de una decisión corporativa con la regulación financiera local.

H2. La métrica de similitud del coseno es una medida valida para calcular la compatibilidad de las decisiones corporativas con el orden jurídico financiero.

6. Variables

6.1. Leyes, Jurisprudencia y Doctrina

Las normativas aplicables a la actividad financiera están organizadas jerárquicamente así: (i) la Constitución Política de Colombia; (ii) las leyes marco, expedidas por el Congreso de la República, las leyes ordinarias, las resoluciones y cartas circulares que expide el Banco de la República en desarrollo de sus funciones, y los decretos con fuerza de ley que expide el Gobierno con base en facultades extraordinarias, como el Estatuto Orgánico del Sistema Financiero (Alesina, 2005). (iii) Los decretos reglamentarios que expide el Gobierno en desarrollo de las leyes marco y, (iv) las cartas circulares y las resoluciones que expide la Superintendencia Financiera en ejercicio de su actividad de inspección y vigilancia (Cárdenas y col., 2008). Estas normas, en conjunto, conforman el ordenamiento jurídico financiero de Colombia.

Por otra parte, la doctrina se refiere al conjunto de opiniones, conceptos y aclaraciones que emite la Superintendencia Financiera, que dan resolución a posibles controversias que no se encuentren legisladas de manera particular (Lax, 2011). Es decir, son una guía de cómo aplicar e interpretar las leyes emitidas por el orden jurídico Financiero (Rubin & Feeley, 1995; Tiller & Cross, 2006). De modo que, se podrían definir como aquellos elementos a los que se acude para tomar decisiones de una manera objetiva.

Finalmente, la jurisprudencia, es el conjunto de sentencias y demás resoluciones judiciales emitidas en un mismo sentido por los órganos judiciales del ordenamiento jurídico financiero (Taruffo, 2007). Tiene un valor fundamental como fuente de conocimiento del derecho, con el cual se procura evitar que una misma situación sea interpretada en forma distinta por otros (Vidal, 1991).

6.2. Decisiones Corporativas

Cada acción que lleva a cabo una empresa es el resultado de una decisión que puede afectar sus operaciones, sus objetivos y sus actividades futuras (Stagner, 1969), al comprender su

importancia y los diferentes tipos, esta puede asegurarse de adoptarlas correctamente en diversas situaciones y momentos.

Las decisiones corporativas tomadas por una empresa pueden ser de diversos tipos, por ejemplo; estratégicas, políticas, operativas, organizacionales, rutinarias, entre otras (Kownatzki y col., 2013; Lim & Chung, 2021). Todas ellas, tienen un contexto, un ámbito de aplicación, un proceso para su determinación y un cierto nivel de importancia relativa (Belkaoui & Karpik, 1989). En este sentido, las políticas contables, financieras, de riesgos, comerciales, los planes de negocio, las ideas de nuevos productos, los planes de expansión, la modificación de la estructura corporativa y administrativa, la incursión en otros mercados, por mencionar algunos, son ejemplos en los que convergen dichas acciones, comúnmente plasmados en diversos documentos empresariales.

Al momento de evaluar sus alternativas en el proceso de toma de decisiones, una empresa debe considerar limitaciones como el mercado objetivo, el tamaño, las capacidades y su regulación particular aplicable (Arrow, 1974), condiciones que pueden variar entre una y otra por el objeto social que cada cual desarrolla.

Una limitación de especial importancia en este proceso es el orden jurídico, ya que toda acción debe guardar correspondencia con dicho marco aplicable. En efecto, para esta investigación, las decisiones corporativas, serán todas aquellas acciones que tome una empresa para desarrollar un objeto social y que su desarrollo sea impactado por un precepto normativo financiero.

6.3. Tratamiento de las Variables

Para el modelo propuesto en esta investigación, el ordenamiento jurídico, la jurisprudencia y la doctrina financiera se denominarán el *Corpus o Contexto*. Por otra parte, las acciones o decisiones corporativas se denominarán *Query o Consulta* debido a que es la variable a la cual se le medirá su compatibilidad con el *Corpus*. No obstante, el Corpus mantendrá sus tres divisiones y la consulta se tratará individualmente por cada decisión corporativa que se requiera validar, en detalle su tratamiento se describe en la sección 7 Metodología.

7. Metodología

Esta investigación adopta un enfoque mixto, al desarrollar un método cuantitativo que permite medir el grado de compatibilidad de las decisiones corporativas con el ordenamiento jurídico financiero colombiano. Concretamente, el proceso consiste en generar una representación de documentos y oraciones en un espacio vectorial, también conocidas como incrustaciones de oraciones y documentos o *embedings*, para el texto contenido en la legislación financiera local, incluyendo la jurisprudencia y doctrina asociada. Para posteriormente, utilizando el mismo espacio, incrustar el texto de una decisión corporativa, con lo cual, es posible medir la similitud semántica entre ambas representaciones del texto. En la Figura 6 se ilustra el enfoque general propuesto.

Dichas incrustaciones son generadas por un transformador de oraciones llamado RoBERTa *"Robustly optimized BERT approach"* (Y. Liu, Ott y col., 2019), este modelo consta de una arquitectura particular de redes neuronales dispuestas en capas con un mecanismo de autoatención similar al que se ilustra en la figura 5, capaz de generar representaciones de palabras, oraciones o documentos en vectores densos de 768 dimensiones, ricos en información del lenguaje y su contexto, esta arquitectura se aborda más en detalle en la sección 7.3.

La codificación de contexto implica que, a diferencia de los modelos tradicionales basados en representaciones dispersas del lenguaje que producen el mismo vector para, por ejemplo, la palabra *"banco"*, ya sea *"un banco cubierto de nieve"* o *"El Banco de la República"*, este modelo modifica la codificación de *"banco"* en función del contexto circundante.

La similitud semántica entre palabras, oraciones y documentos, se puede obtener utilizando la métrica del coseno, como se detalla en la sección 7.4, esta medida aprovecha la representación del texto como un vector en un espacio de alta dimensión para calcular la concurrencia entre ellos, y es valida por su amplio uso en el campo de PLN (Kang y col., 2020; Nguyen & Bai, 2010; Rahutomo y col., 2012; Zhou y col., 2020).

7.1. Enfoque General Propuesto

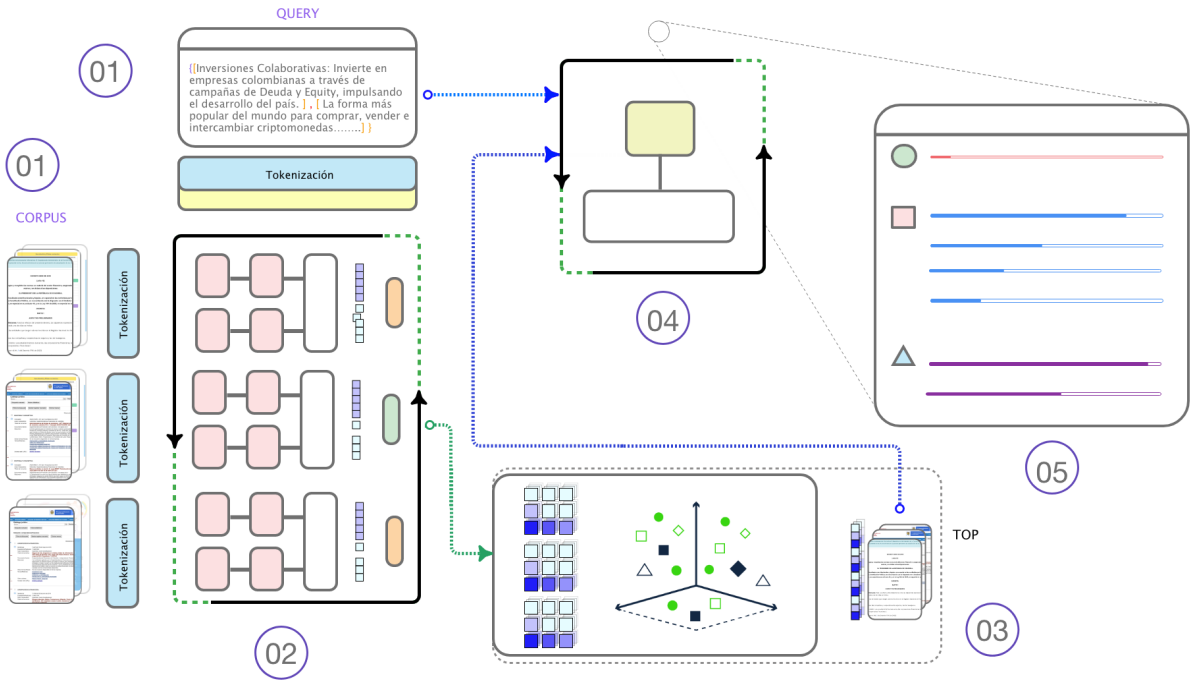


Figura 6: Arquitectura del Modelo

Nota: Adaptado de "Optimizing Bi-Encoder for Named Entity Recognition via Contrastive Learning", (p. 3) S. Zhang y col., 2022, Microsoft Research y "Trans Encoder Unsupervided Sentence Pair Modelling Through Self and Manual Destillations", (p. 2) F. Liu y col., 2021, University of Cambridge Amazon Research

① *En este primer paso, el texto del ordenamiento jurídico financiero colombiano y el de la decisión corporativa objeto de validación, se convierte a vectores de números reales con el método que se describe en la sección 7.2.*

② *Este paso toma los vectores generados en el paso ① y utilizando la arquitectura descrita en la sección 7.3 y el método (Bi-codificador) de la sección 7.3.1 se codifica y empareja el texto en un espacio vectorial denso de 768 dimensiones.*

③ *Con la salida del paso ② se calcula la similitud del coseno descrita en la sección 7.4.*

④ *En este paso se ejecuta un codificador cruzado como se describe en la sección 7.3.2 sobre las salidas del paso ③ y se calcula una nueva similitud de coseno.*

⑤ *Este último paso toma la salida del paso ④, promedia los resultados de similitud de acuerdo a la sección 7.5, presenta resultado de similitud y traduce los vectores generados al texto original de entrada.*

Las secciones ②, ③ y ④ de la figura 6 consideran una tarea general de coincidencia semántica entre un contexto y una consulta (Legislación y Decision Corporativa) (Ye y col., 2022). Lo anterior, tiene un alto interés práctico en un amplio espectro de aplicaciones empresariales, como la búsqueda web y la respuesta automatizada a preguntas (Ferrucci & Lally, 2004; Lewis & Young, 2019; Masson & Paroubek, 2020). En este caso se puede considerar como el aprendizaje de una función de puntuación $f : C \times L \rightarrow \mathbb{R}$, donde C es un conjunto de consultas y L es un conjunto de candidatos. La función f asigna un par de consultas y candidatos $(S_c, S_l) \in C \times L$ a una puntuación de relevancia p_{cl}. La consulta S_c está representada por n palabras $S_c = [c_1, \ldots, c_n]$ y el candidato S_l está representado por m palabras $S_l = [l_1, \ldots, l_m]$.

7.2. Tratamiento inicial del texto de la Legislación Financiera Local y de las Decisiones Corporativas - (Contexto y Consulta)

7.2.1. Algoritmo de Codificación de Pares de Bytes (BPE)

Como se menciona en la sección 7.3 RoBERTa adopta la codificación BPE para el conjunto de datos de entrenamiento[5] (Bowman y col., 2015; Conneau y col., 2018; Gururangan y col., 2018; Williams y col., 2017), este método es originalmente un algoritmo para compresión de información mediante la búsqueda de combinaciones comunes de pares de bytes (Gage, 1994). Sin embargo, actualmente se usa en PNL para hallar la forma más eficiente de representar texto en forma de tokens.

Tokenizar un texto es dividirlo en unidades más pequeñas, que luego se convierten en identificadores únicos (Ids) (Dai y col., 2019; Eyre y col., 2021; Graën y col., 2018), el algoritmo BPE utilizado en RoBERTa se fundamenta en unidades de subpalabras a nivel de byte (Y. Liu, Fan y col., 2019), que se extraen realizando un análisis estadístico del corpus de entrenamiento (Sennrich y col., 2015), formalmente el procedimiento es el siguiente:

1. BPE, toma un contexto base C que se normaliza para obtener un vocabulario de tamaño k.

[5] ESXNLI: solo la parte en español, SNLI y MultiNLI: traducido automáticamente

2. $V_k \leftarrow$ es el vocabulario con todos los tokens o $n-gramas$ únicos en C.

3. En seguida se toma el par de tokens (t_m, t_r) más frecuente en C

4. El par (t_m, t_r) genera un nuevo token $(t_m, t_r) \rightarrow t_N$.

5. Este nuevo token se agrega al vocabulario $t_N + V_k \rightarrow V_{k+1}$. Donde V_k es el vocabulario inicial del numeral 1 y V_{k+1} es el vocabulario inicial más en nuevo token t_N

6. Cada ocurrencia de (t_m, t_r) en C se reemplaza con el nuevo token t_N.

7. Finalmente, se repite el proceso desde 3, hasta no hallar nuevos tokens para agregar al vocabulario V_k.

8. El tamaño final del vocabulario $V_{kRoBERTa}$ es igual al tamaño del vocabulario inicial, más el número de operaciones de combinación.

7.2.2. Codificación del Texto de Entrada - Tokenización

Para codificar los nuevos datos, en este caso la legislación financiera local y la decisión corporativa (Contexto y Consulta) el proceso es el mismo de la sección 7.2.1, como resultado se obtiene una lista de tokens V_{kL} y V_{kC} que ya están presentes en el diccionario inicial $V_{kRoBERTa}$, si quedan algunos $n-gramas$ que el algoritmo BPE de RoBERTa no vio en el entrenamiento, estos son reemplazadas por tokens desconocidos $[UNK]$.

De esta forma, las oraciones del contexto y la consulta serán representadas en vectores de números de 512 tokens, los valores numéricos corresponderán al identificador de cada token en el diccionario $V_{kRoBERTa}$.

7.2.3. Decodificación

é í $V_{kRoBERTa}$ ó

7.3. Incrustación de Legislación y Decisiones Corporativas - Embeding

Para mapear el par $(S_l, S_c) \in L \times C$ se utiliza un transformador de oraciones, previamente entrenado en una tarea de inferencia de lenguaje natural (NLI) del idioma español denominado ROBERTa (Y. Liu, Fan y col., 2019; Radford y col., 2019). A pesar de conservar la estructura original de transformer, este adopta 12 capas $(L = 12)$ de codificador y decodificadores, un tamaño de capa oculta $(H = 768)$ y $(A = 12)$ cabezas de autoatención en total $110M$ parámetros, adicionalmente, adopta el método de codificación descrito en la sección 7.2 *"Byte Pair Encoding - BPE"* para el texto de entrada (Yates y col., 2021).

El modelo transformer consiste en varias capas de autoatención apiladas con conexiones residuales. Cada capa de atención propia recibe n incrustaciones $\{x_i\}_{n=1}^{n}$ correspondientes a tokens de entrada únicos y genera n incrustaciones $\{z_i\}_{n=1}^{n}$, conservando las dimensiones de entrada. El $i - esimo$ token se asigna a través de transformaciones lineales a una clave k_i, una consulta q_i y un valor v_i. La $i - esima$ salida de la capa de autoatención se obtiene ponderando los valores v_j por el producto escalar normalizado entre la consulta q_i y otras claves k_j, dividido por la raíz de la dimensión de los vectores clave $\sqrt{d_k}$:

$$z_i = \sum_{j=1}^{m} softmax(\{\frac{\langle q_i, k_{j\prime}\rangle}{\sqrt{d_k}}\}_{j\prime=1}^{n})_j \cdot v_j. \tag{3}$$

7.3.1. Codificador Doble (Bi-Encoder)

En la sección ② de la figura 6 se emplean codificadores dobles para cada uno de los elementos del corpus (3 Bi-Codificadores), cada par C_L y C_C codifican separadamente la consulta y la legislación en el espacio como:

$$v_l = Poling(C_L(S_l)), v_c = Poling(C_C(S_c)) \tag{4}$$

Luego se usa la distancia de coseno para medir la relevancia entre s_l y s_c. La función $Pooling()$ selecciona la primera muestra de $C_L(S_l)$ y $C_C(S_c)$ como sus incorporaciones finales.

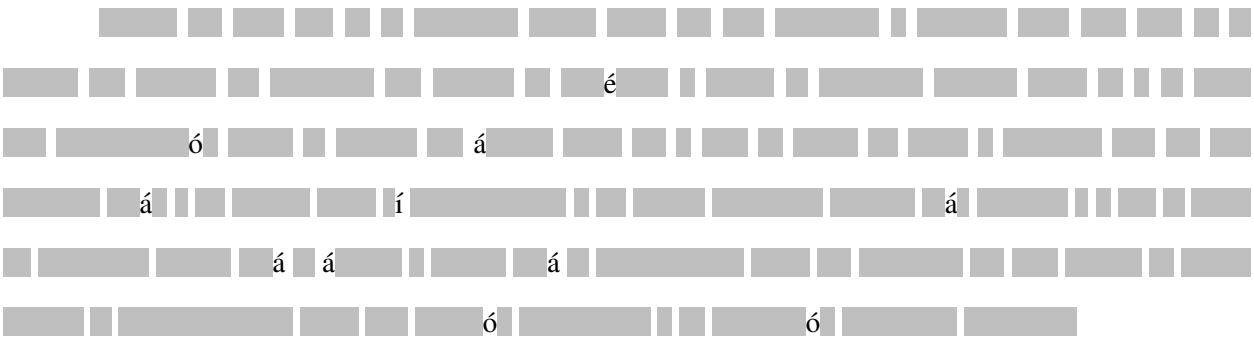
7.3.2. Codificador Cruzado (Cross-Encoder)

En la sección ④ de la figura 6 se utiliza un codificador cruzado que es un método basado en la interacción, que aplica RoBERTa en la concatenación de S_c y S_l como:

$$P_{cl} = RoBERTa([[CLS]; S_c; [SEP]; S_l; [SEP]]) \tag{5}$$

Donde $[CLS]$ es un token de entrada adicional para agregar la incrustación de salida y $[SEP]$ es una notación para la separación. Este método logra una mayor precisión que otros métodos basados en RoBERTa debido a la codificación contextual de alta calidad generada por la autoatención total.

7.4. Similitud del Coseno



7.5. Obtención de Resultados

8. Trabajo de Campo

8.1. Recolección de Información

8.1.1. Regulacion Financiera (Corpus o Contexto)

Debido al principio de democracia participativa contenido en la Constitución Política de Colombia, todo el texto que compone el ordenamiento jurídico, la doctrina y la jurisprudencia de la actividad financiera colombiana se puede hallar y extraer con cierta facilidad de las páginas web de las entidades administrativas del estado, descritas ee la tabla 1.

Entidad	Tipo de Documento	Clasificación
Secretaria General del Senado	Constitución Política	Legislación
Función Pública	Leyes	Legislación
	Decretos	Legislación
Superintendencia Financiera	Circulares Externas	Legislación
	Cartas Circulares	Legislación
	Resoluciones	Legislación
	Doctrina y Conceptos	Doctrina
	Fallos Jurisdiccionales	Jurisprudencia
	Jurisprudencia Financiera	Jurisprudencia
Banco de la República	Cartas Circulares	Legislación
	Resoluciones	Legislación
Corte Constitucional	Boletines	Jurisprudencia
Corte Suprema de Justicia	Boletines	Jurisprudencia

Tabla 1: Fuentes de Datos

Nota: Se incluye como fuente de información a la web de la Función Pública, ya que esta, es una entidad técnica, estratégica y transversal del Gobierno Nacional que agrupa un gran número de leyes, decretos, cartas circulares, resoluciones, entre otros documentos emitidos por las entidades que hacen parte del ordenamiento jurídico financiero.

Para obtener el texto mencionado anteriormente se utiliza un algoritmo de web scraping con la librería BeautifullSoup (L. Richardson, 2007), primero se obtienen todas los url's contenidos en un sitio web y posteriormente se extrae el texto de la regulación, un ejemplo de como realizar esta extracción desde una url se muestra en el anexo B.

8.1.2. Desciciones Corporativas (Query Consulta)

Para la validación del modelo propuesto se toman diferentes muestras de políticas, planes de negocio y otras decisiones corporativas, como las descritas en la tabla 2, sobre las cuales se conocen ampliamente las normas asociadas y que cualitativamente se reconocen como decisiones acordes con la legislación financiera local.

Tipo	**Decisión**
Objetivo del FIC	El objetivo del Fondo de Inversión Colectiva es proporcionar a los inversionistas un instrumento de inversión de renta fija de baja duración, con el perfil de riesgo conservador, cuyo propósito es la estabilidad del capital en un horizonte de inversión de corto plazo (Davivienda, 2022).
Código de Ética	12. Compromiso frente al Riesgo de Lavado de Activos y de la Financiación del Terrorismo Los directores, administradores y funcionarios de Corficolombiana mantienen la cultura de prevenir, detectar y controlar que la Corporación sea utilizada como instrumento para el lavado de activos y la financiación del terrorismo (LA/FT). Por tal motivo, se ha implementado el Sistema de Administración de Riesgo de Lavado de Activos y Financiación del Terrorismo (SARLAFT), el cual contiene las políticas de ética que orientan la actuación de los directores, administradores y funcionarios para el cumplimiento del mismo, las políticas de vinculación y conocimiento de clientes y de sus transacciones con la Corporación, los procedimientos y metodologías para la identificación, evaluación, control y monitoreo de riesgos, la capacitación al personal y la colaboración con las autoridades contribuyendo al aseguramiento de la confianza del público en la Corporación y en el sistema financiero colombiano (Corficolombiana, 2021).
Alcance Política Tratamiento de Datos	El tratamiento que se realice por parte de la entidad. se basará en la autorización otorgada por el titular y tomará en cuenta las finalidades expresamente informadas. Así mismo, en desarrollo de su actividad y gestión, y con el fin de brindar colaboración empresarial entre las empresas del grupo, durante la ejecución de sus actividades podrá efectuar el tratamiento de datos personales de forma conjunta con las entidades que pertenezcan o llegaren a pertenecer al GRUPO, o a quien represente sus derechos u ostente en el futuro la calidad de acreedor, cesionario, o cualquier calidad frente a los titulares de la información (Bancolombia, 2022).

(*continua en la página siguiente*)

Tipo	Decisión
Alcance Política Tratamiento de Datos	Se entenderán que son parte del GRUPO las entidades que pertenezcan o puedan llegar a pertenecer al Grupo de acuerdo con la ley, sus filiales y/o subsidiarias, o las entidades en las cuales estas, directa o indirectamente, tengan participación accionaria o sean asociados, domiciliadas en Colombia y/o en el exterior.
Términos y Condiciones	Actividad de Financiación Colaborativa - Crowdfunding La Bolsa de Valores de Colombia S.A. (en adelante "bvc") administra una plataforma que realiza la actividad de financiación colaborativa. Tal Plataforma de Financiación Colaborativa se denomina a2censo (en adelante a2censo o la Plataforma). La administración de la actividad de financiación colaborativa realizada por bvc se desarrolla básicamente a través de una infraestructura tecnológica, que puede incluir interfaces, plataformas, páginas de internet u otro medio de comunicación electrónica o digital, a través del cual se pone en contacto un número plural de aportantes con receptores que solicitan financiación en nombre propio para destinarlo a un proyecto productivo. La financiación colaborativa se materializa a través de la adquisición de valores de financiación colaborativa y es realizada directamente por los Aportantes en favor de los Receptores (bvc - bolsa de valores de colombia, 2020).

Tabla 2: Decisiones Corporativas

8.2. Transformación

Para agilizar los procesos de lectura y escritura del texto recuperado del ordenamiento jurídico colombiano y para facilitar su utilización en el modelo, los marcos de datos recuperados se transforman al formato Feather con compresión estándar (LZ4) mediante la librería Pyarrow, este es un formato de archivo portátil para almacenar tablas Arrow, utilizando el formato Arrow IPC internamente.

8.3. Estructura del Modelo

La estructura del modelo descrito en la figura 6 se implementó utilizando las bibliotecas Transformer, Pytorch, Pandas, Numpy, y Pyarrow (Harris y col., 2020; pandas development team, 2022; Paszke y col., 2019; N. Richardson y col., 2022; Wolf y col., 2020) en el lenguaje python (Van Rossum & Drake, 2009) como se muestra en el anexo A.

8.4. Análisis de Resultados

Las validaciones del modelo se ejecutaron utilizando los métodos descritos en las secciones 7.2 a 7.5, con los conjuntos de datos recuperados de las páginas web descritas en la tabla 1 para el contexto, y 100 decisiones corporativas como las descritas en la tabla 2 para las consultas de validación, los resultados más relevantes observados para el modelo utilizado se presentan en los apartados a continuación.

8.4.1. Estructura Semántica del Texto

Las normas de textualidad señalan que la cohesión se refiere a la estabilidad de un texto que se mantiene gracias a la continuidad de los elementos que lo conforman (De Beaugrande & Dessler, 1997). Esta noción de continuidad se basa, en la suposición de que existe una relación entre los diferentes elementos lingüísticos que configuran el texto, mientras que la coherencia es algo que va más allá de lo que se encuentra en la superficie del texto, ya que es un juego entre el texto mismo y los conocimientos que tiene el lector, que a diferencia de la cohesión, alude a elementos intangibles (Hernández Osuna & Ferreira Cabrera, 2016; Sleimi y col., 2018).

Aunque en la literatura la discusión de los anteriores conceptos es amplia (Parsing, 2009), en el campo de la lingüística computacional, la cohesión se refiere a la forma en que las unidades textuales son enlazadas, y la coherencia se refiere a las relaciones de significados entre dos unidades léxicas (Gardner y col., 2018), estas dos normas, conforman la estructura semántica del texto en el procesamiento de lenguaje natural, es decir, que dicha semántica implica el uso y el significado de palabras o frases en un contexto (Gabrilovich & Markovitch, 2009).

El transformer RoBERTa utilizado en este trabajo, gracias a los mecanismos de atención, logra representar dicha estructura semántica de forma adecuada para el ordenamiento jurídico colombiano, en la figura 8 se muestra una representación aproximada en dos dimensiones del espacio vectorial denso generado por el modelo, esta representación captura la relación entre oraciones que se encuentran dispersas en todo el contexto y genera grupos en regiones puntuales del espacio cuando se hace referencia a elementos particulares, como se puede observar en los gráficos 8(a) y 8(b) respectivamente.

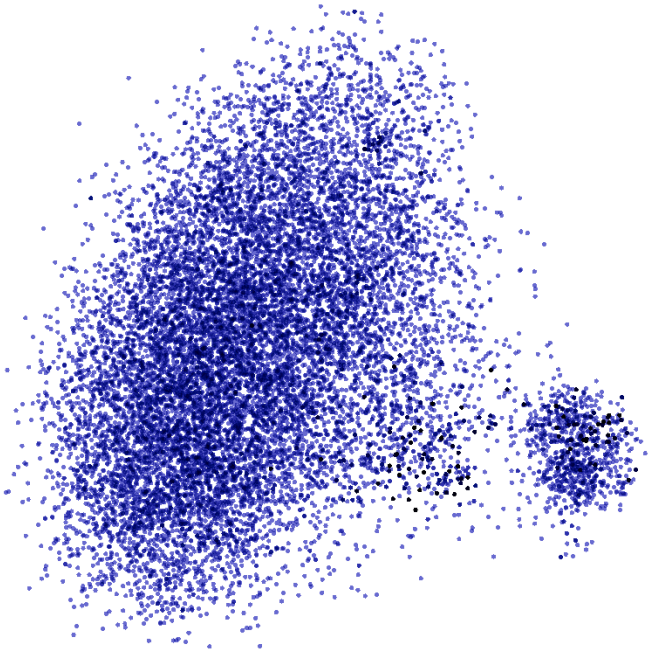
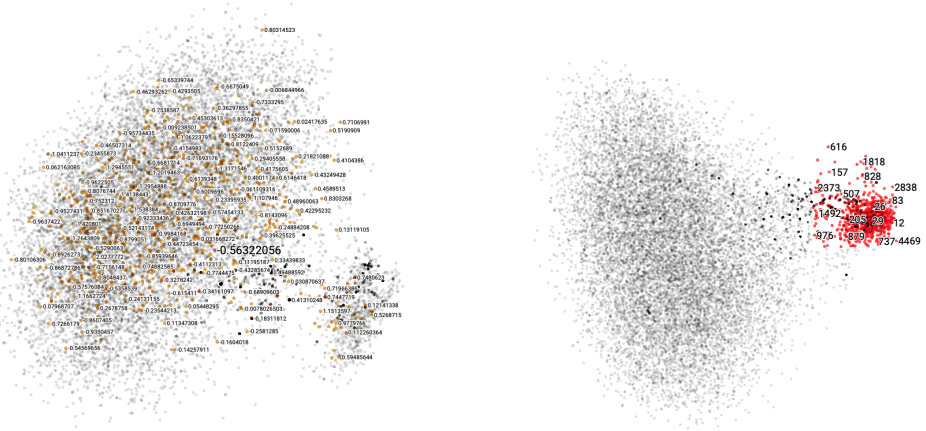


Figura 7: Representación del Contexto



(a) Relación Extensa (b) Relación Local

Figura 8: Relaciones del Contexto

Nota: Gráficos generados desde de la web Embeding Projector, para los vectores del modelo RoBERTa utilizado

8.4.2. Principales Resultados y su Comparación

Como se mencionó previamente, el modelo ROBERTa hace uso de la estructura semántica que incluye la desambiguación del sentido, es decir, que deriva el significado de una oración en función del contexto, lo cual representa una ventaja para la medida de similitud propuesta, al poner a prueba esto, los resultados observados demuestran que, se recuperan adecuadamente las normas asociadas a la decisión corporativa objeto de validación, con su respectivo puntaje de compatibilidad, resultando típicos, valores entre 0.73 y 0.94 de similitud media, del el grupo de 100 decisiones evaluadas como se muestra en la tabla 3.

%	**Rango Puntaje de Similitud Media**	
2 %	0.05	0.20
5 %	0.21	0.60
7 %	0.61	0.72
33 %	0.73	0.80
52 %	0.81	0.94
1 %	$\geq$	0.95

Tabla 3: Resultados Generales

Al entrar en detalle en los resultados anteriores se encuentra, por ejemplo, que al tomar el texto completo de la cuarta decisión corporativa detallada en la tabla 2 (Términos y condiciones), el modelo recupera, además de otras normas relacionadas, las detalladas en la tabla 4 con una similitud media de 0.83, valor que indica que la política es altamente compatible con el orden jurídico aplicable.

No.	**Norma**
1	*La actividad de financiación colaborativa es aquella desarrollada por entidades autorizadas por la Superintendencia Financiera de Colombia, a partir de una infraestructura electrónica, que puede incluir interfaces, plataformas, páginas de internet u otro medio de comunicación electrónica, a través de la cual se ponen en contacto un número plural de aportantes con receptores que solicitan financiación en nombre propio para destinarlo a un proyecto productivo de inversión.*

(*continua en la página siguiente*)

No.	Norma
2	*La actividad de financiación colaborativa será desarrollada por sociedades anónimas de objeto exclusivo que tengan como propósito poner en contacto a un número plural de aportantes con receptores que solicitan financiación en nombre propio para destinarlo a un proyecto productivo, las cuales se denominarán sociedades de financiación colaborativa. Las bolsas de valores y los sistemas de negociación o registro de valores autorizados por la Superintendencia Financiera de Colombia, también podrán realizar la actividad de financiación colaborativa.*
3	*Para efectos de la actividad de financiación colaborativa, se denomina genéricamente como aportante, a las personas que intervienen en cualquier operación de financiación que se lleve a cabo a través de las entidades autorizadas para realizar la actividad de financiación colaborativa con el fin de financiar proyectos productivos.*

Tabla 4: Resultado Términos y Condiciones

Algoritmo Python 1: Ressultado Terminos y Condiciones

```
search_and_Similarity_q(consulta = 'La Bolsa de Valores de Co ...')
```

```
0.87 La actividad de financiacion colaborativa es aquella desarrollada...
...
...
...
--------------------------------
Puntaje Medio Similitud total: 0.83
```

El extracto de las anteriores normas corresponden al libro 41 Actividad de Financiación Colaborativa, Título 1 del Decreto 2555 de 2010, artículos 2.41.1.1.1 al 2.41.1.1.5, adicionado por el artículo 1 del Decreto 1357 de 2018. Al verificarlo detalladamente, en efecto, corresponde al marco normativo aplicable a la política de términos y condiciones de la plataforma a2censo (bvc - bolsa de valores de colombia, 2020), plataforma que cumple adecuadamente las disposiciones aplicables, por lo cual, es razonable el puntaje de compatibilidad observado previamente.

Ahora bien, si se toma la siguiente descripción del objeto social de otra plataforma de financiación colaborativa, el modelo retorna un resultado de compatibilidad media de 0.67 y un marco normativo similar al de la tabla 4.

Vaki es una plataforma de crowdfunding o financiamiento colectivo, donde puedes crear campañas a las cuales llamamos Vakis. Una campaña de crowdfunding básicamente es una "vaca en línea", donde se busca recaudar fondos de diferentes personas que comparten los mismos ideales y quieren llevar a cabo un proyecto juntos (Vaki, 2020).

Algoritmo Python 2: Ressultado Vaki

```
search_and_Similarity_q(consulta = 'Vaki es una plataforma de crow...')
```

```
...
...
...
...
...
------------------------------
Puntaje Medio Similitud total: 0.67
```

En este caso, al realizar una validación exhaustiva del resultado del modelo, se observa que aunque las plataformas son similares en su funcionamiento, "LaVaquinha S.A.S - Vaki", no es una entidad vigilada y autorizada por la Superintendencia Financiera de Colombia, el tipo de sociedad no corresponde con las definidas en la norma, entre otras, lo cual, indica que el valor de compatibilidad obtenido es adecuado, dado que el orden jurídico financiero es el correcto para la entidad, sin embargo, su objeto social no satisface cabalmente los requerimientos de dichas normas. Cabe mencionar que lo discutido anteriormente no implica que la entidad Vaki no cumpla con la regulación colombiana, pues este estudio se limita al ordenamiento jurídico financiero, es decir, que la entidad puede operar bajo regulaciones particulares que no son estrictamente financieras.

Otro resultado observado en el caso de Vaki, es que el modelo calcula algunos valores de similitud negativos, lo cual, indicaría que la decisión corporativa es opuesta a la regulación, este resultado podría ser contradictorio y conducir a mediciones inconsistentes en algunos casos como se menciona en la sección 9 más adelante. Por parte de la jurisprudencia, el modelo no recupera resultados, mientras que en la doctrina se recupera los conceptos 019009738 - 001 del 14 de febrero de 2019, 2019111966 - 002 del 22 de agosto de 2019, 2018126630 - 001 del 8 de noviembre de 2018 y 2017008080 - 001 del 24 de febrero de 2017 de la Superintendencia Financiera, que contribuyen con la explicación de los resultados ya mencionados.

Otros resultados relevantes surgen de la evaluación de los textos de la tabla 5, donde se detallan dos políticas asociadas al patrimonio mínimo con el que debe contar cada fondo de inversion colectiva, tomadas de los reglamentos de dos vehículos de inversión, pertenecientes a la misma categoría y tipo, administrados por dos sociedades fiduciarias reconocidas del sistema financiero colombiano.

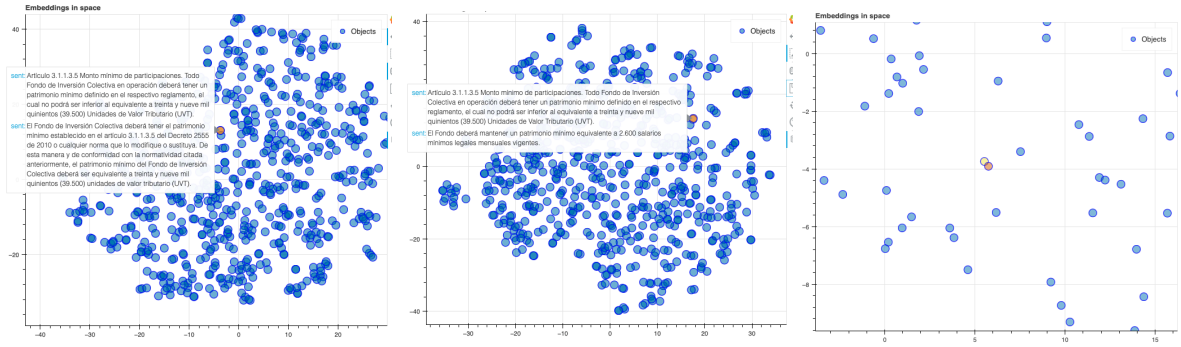
Texto	Descripción	Score

(*continua en la página siguiente*)

Texto	Descripción	Score
FIC_1	El Fondo de Inversión Colectiva deberá tener el patrimonio mínimo establecido en el artículo 3.1.1.3.5 del Decreto 2555 de 2010 o cualquier norma que lo modifique o sustituya. De esta manera y de conformidad con la normatividad citada anteriormente, el patrimonio mínimo del Fondo de Inversión Colectiva deberá ser equivalente a treinta y nueve mil quinientos (39.500) unidades de valor tributario (UVT).	0.9250
FIC_2	El Fondo deberá mantener un patrimonio mínimo equivalente a 2.600 salarios mínimos legales mensuales vigentes.	0.7230
$Decreto_{2555}$	Artículo 3.1.1.3.5 Monto mínimo de participaciones. Todo Fondo de Inversión Colectiva en operación deberá tener un patrimonio mínimo definido en el respectivo reglamento, el cual no podrá ser inferior al equivalente a treinta y nueve mil quinientos (39.500) unidades de valor tributario (UVT).	1

Tabla 5: Patrimonio Mínimo y Regulación Aplicable

Aunque, el texto de cada política de patrimonio mínimo es substancialmente diferente, ambas cumplen con los requisitos normativos del artículo 3.1.1.3.5, lo anterior es verificable si se tiene en cuenta que los reglamentos de estos tipos de fondos debe ser autorizados por el regulador, y en efecto se encuentran con operación vigente, por lo cual, de manera simple se entendería que los valores de compatibilidad son adecuados. Sin embargo, considerando la notable diferencia en su puntuación de similitud, es necesario comprender como se ubican estas dos políticas en el espacio vectorial del orden jurídico financiero, lo cual se muestra en la figura 9.



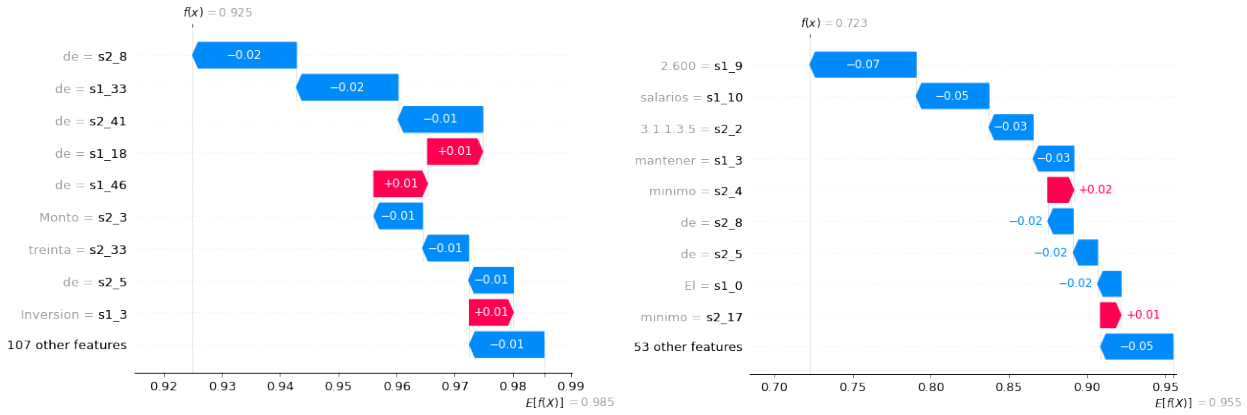
(a) FIC_1 Vs $Decreto_{2555}$ (b) FIC_2 Vs $Decreto_{2555}$ (c) FIC_1 Vs FIC_2

Figura 9: Representación Gráfica

Nota: (●) Representa el contexto de orden jurídico financiero, (●) Corresponde a la norma recuperada por el modelo en función de la decisión corporativa de consulta, y (●) es la incrustación de la decisión corporativa en el mismo espacio vectorial del contexto.

Al fijar la atención en la figura 9(c), se puede notar que la política de patrimonio mínimo del fondo 1 y 2 se ubican cerca en el espacio vectorial, con un puntaje de similitud entre sí de 0.9133, indicando esto que su puntaje individual en relación con el artículo que los gobierna, debiera ser un valor más cercano.

Utilizando los valores de Shapley, que son un enfoque ampliamente utilizado de la teoría de juegos cooperativos, que permite saber cuánto ha contribuido a la predicción cada una de las características (Aumann & Shapley, 2015; Ethayarajh & Jurafsky, 2021), en este caso para observar la contribución al puntaje de similitud, de cada elemento en la oración, se obtienen los resultados que se muestran en la figura 10.



(a) FIC_1 Vs $Decreto_{2555}$ (b) FIC_2 Vs $Decreto_{2555}$

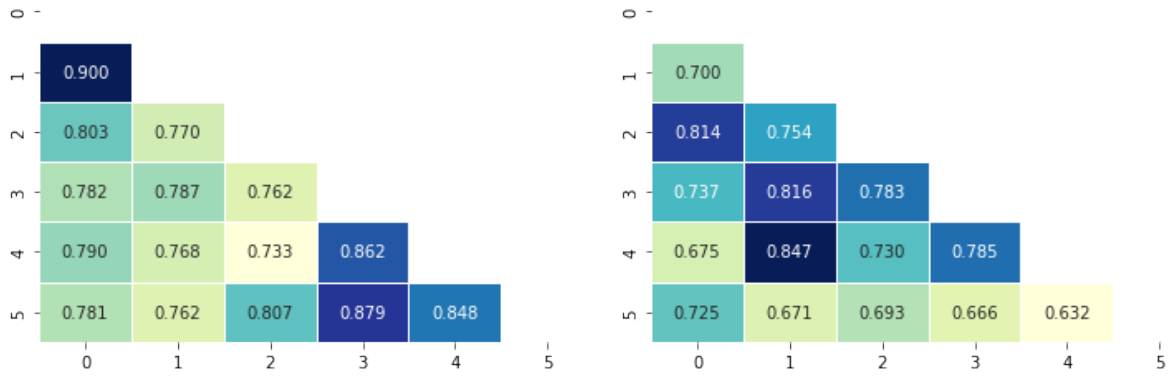
Figura 10: Valores de Shapley

Los valores de Shapley permiten identificar que el texto resaltado en Rojo en la tabla 5, influye de forma importante en el valor de compatibilidad de la política del FIC_1, pues mientras que elementos como los resaltados en verde, le permiten al modelo identificar adecuadamente el marco normativo aplicable, el texto en Rojo al ser una copia literal de la norma, acerca de manera importante la decisión corporativa al contexto particular.

Lo anterior, no es estrictamente un error, sin embargo, genera un efecto de sobre ajuste en el modelo que debe ser tratado, y es que al verificar el conjunto de decisiones corporativas, se identifica una propensión de los administradores a transcribir fragmentos literales de las normas aplicables, y por consiguiente se requiere establecer una forma de manejo que se discutirá en la sección 9.

Finalmente, continuando con los resultados de la tabla 5, al explorar los valores de similitud del coseno recuperados por el codificador cruzado en la metodología propuesta, para el top 5 de normas más

compatible, se tienen los resultados de la figura 11, presentados en forma de una matriz.



(a) FIC_1 Vs TOP_5 (b) FIC_2 Vs TOP_5

Figura 11: Similitud del Coseno

Nota: 0 corresponde a la política de patrimonio mínimo y 1-5 corresponde al top 5 de normas compatibles

Al tomar la primera columna de la figura 11(a) se observa que la fila 1 con puntaje 0.90, pertenece a la legislación discutida previamente, lo mismo sucede, en la columna 1 de la figura 11(b) fila 2 con puntaje 0.81, lo que corresponde a un resultado correcto.

No obstante, al entrar en detalle de las normas recuperadas con los valores entre 0.78 y 0.80 de la figura 11(a), y los valores 0.67 a 0.73 de la figura 11(b), son validas en el contexto general de los fondos de inversión colectiva, pero se refieren a un tipo de fondo en específico, en este sentido, para el enfoque aquí propuesto, los resultados recuperados son correctos. No obstante, para un resultado más preciso de la relación de contexto, sería necesario indexar el orden jurídico en clusters más pequeños, correspondiendo esto a un enfoque futuro de investigación que se menciona en la sección 10 de este documento.

9. Discusión

Aunque el enfoque general propuesto logra buenos resultados, también puede llegar a presentar ciertas imposiciones, como las mencionadas previamente en la sección 8.4.2 Principales Resultados y su Comparación, para comprender dichas observaciones inconsistentes, se pueden abordar en tres categorías.

1. Cuando se trató el caso de la plataforma de financiación colaborativa (Vaki) y la legislación normativa aplicable de la tabla 4, se observaron ciertos valores de similitud negativos, lo cual, es potencialmente probable que tenga origen en diseño conceptual del modelo RoBERTa, pues este no es sensible a la polaridad de los fragmentos de texto (Ferreira y col., 2014; Schulder y col., 2017), ocasionando colisiones semánticas entre oraciones que naturalmente no están relacionadas, pero que generalmente los modelos de PNL juzgan como similares (Song y col., 2020).

2. En los resultados de la tabla 5, se resalta un sobre ajuste en las medidas de similitud, comprensible por la inclinación a redactar políticas con contenidos literales de las normas, pese a ello, no son efectos deseables en aplicaciones del mundo corporativo, dado que a presencia lleva falsos positivos, no obstante, incrementar un mecanismo de resumen basado en PNL (Adhikari y col., 2020; LeClair y col., 2019), para la entrada de la decisión corporativa en el modelo aquí propuesto, mitigaría estos efectos.

3. Un resultado que se considera relevante y potencialmente contradictorio para el enfoque propuesto, es la capacidad interna del modelo para generar categorías y subcategorías de contextos muy particulares, a partir del orden jurídico global, al suponer que las normas colombianas son exhaustivas y detalladas, observando valores inconsistentes en el caso de las políticas de patrimonio de los Fondos de Inversión. Particularidad en la cual se puede profundizar adoptando enfoques de indexación en la arquitectura propuesta (Bast y col., 2016; J. Johnson y col., 2019).

En definitiva, la metodología propuesta para medir la compatibilidad de las decisiones corporativas con el ordenamiento jurídico financiero colombiano, tiene potenciales factores de mejora, que podrían permitirle al modelo desenvolverse correctamente en aplicaciones del ámbito corporativo, en este sentido, una modificación que se considera prudente en el enfoque propuesto, una vez validados en detalle los resultados obtenidos durante el trabajo de campo y análisis, consiste en establecer umbrales en la puntuación

de similitud para facilitar la diferenciación entre aquellas decisiones compatibles y no compatibles con la regulación financiera.

En todo caso, los resultados iniciales permiten aproximarse en la dirección correcta a un método de validación automatizada que mitiga los costes de acudir a expertos de la profesión legal para la validación de los documentos corporativos.

10. Conclusiones y Trabajo Futuro

10.1. Conclusiones

En esta investigación se explora cómo determinar el grado de compatibilidad entre una decisión corporativa y la legislación financiera colombiana, en específico, como medir de forma automatizada la similitud entre estos dos elementos de texto, haciendo uso de las técnicas más recientes para el procesamiento de lenguaje natural, desarrolladas en el campo de la lingüística computacional. Expresamente, empleando un modelo probabilístico comúnmente conocido como transformer o transformador de oraciones, nombrado RoBERTa por sus creadores (Y. Liu, Ott y col., 2019).

Para ello, se propone la arquitectura de la figura 6 que tiene fundamento en el transformador RoBERTa, dicha estructura computacional, aprovecha los conceptos de estudios previos en métodos de recuperación de información, incluyendo tres codificadores dobles y un codificador cruzado, en combinación con la métrica de similitud del coseno, para dar respuesta a los objetivos planteados.

Los resultados obtenidos discutidos en la sección 8.4.2, demuestran la capacidad de la metodología para obtener una puntuación de compatibilidad entre una decisión corporativa y el orden jurídico financiero aplicable, en este caso mediante la similitud del coseno, comparable a una evaluación no automatizada realizada por un evaluador humano, resolviendo de este modo, el problema y objetivo central de esta investigación.

Así mismo, las representaciones estructurales generadas y discutidas en la sección 8.4.1, permiten concluir que el núcleo del enfoque propuesto, representado por el modelo de lenguaje RoBERTa, posee la capacidad para representar de forma adecuada la estructura semántica del contexto del orden Jurídico financiero colombiano, y naturalmente, durante el desarrollo de la investigación y el trabajo de campo realizado, fue posible obtener la información necesaria para el modelo, a traves de un proceso automatizado conocido como *"web scraping"*.

En consecuencia, el desarrollo de cada una de las etapas de esta investigación, ha permitido cumplir rigurosamente con los objetivos específicos propuestos al comienzo de este documento, descritos en la sección 2.2 y que fueron necesarios para alcanzar plenamente el objetivo principal.

Finalmente, aunque los resultados obtenidos precisan refinamiento, resultan interesantes para continuar

investigando y robusteciendo el enfoque propuesto, de tal forma que pueda ser útil en aplicaciones del ámbito corporativo.

10.2. Trabajo Futuro

El método propuesto en el enfoque de esta investigación se puede continuar investigando en profundidad y potencialmente robustecerse para mitigar las posibles distorsiones en sus resultados, mediante:

- El uso de los enfoques basados en gradientes y otros métodos de mitigación para tratar las colisiones semánticas observadas en la sección 9.
- El preprocesado de las decisiones corporativas, con un método que permita sintetizarlas, sin perder sus características principales.
- La integración de extracción de ontologías, para construir índices automáticos que conservan la semántica y permitan segmentar el contexto global para obtener resultados más precisos.
- Establecer capas finales de clasificación con umbrales de puntuación de similitud, para facilitar la comprensión de los resultados a los usuarios finales no experimentados.

A. Anexo 1 Código Python del Modelo

Algoritmo Python 3: Script del Modelo

```
#from xxxxxxxx_xxxxxxxxxxx import xxxxxxxxxxxxxxx
#from xxxxxxxxxx import xxxxxxxxxx
#from xxxxxxxx_xxxxxxxxxxx import xxxxx, xxxxxxxxxx
#import xxxxxx.xxxxxx as xxxxxx
#import xxxxxx as xxxx
#import xxxx as xxxx
#import xxxxx
#import xxxx

#xxxx_xxx = r'xxxxxx/xxxxxx'
#xxxxx = xxx.xxxxx.xxxxx(xxxxxx_xxxx, 'xxxx')
#xxxx_xxx = "xxxxxxxxxx.arrow"
#xxxxx_xxx = xxxxxxx.xxxx_xxxxxxx(xxxxx+xxxx.xxxxxxxxxx)
#xxxxxxx = xxxxxx(xxxx_xxxxx['xxxxxxx'])
#xxxxxxx = [xxxxxx for xxxxxxx in xxxxx if xxx(xxxxxx) !=x]

#xxxxxxxxx= 'xxxxx_xxxx'

#xx_xxxxxx = xxxxxxxxxxxxx(xxxxxxxxxx)
#xx_xxxxxxx.xxxx_xxx_xxxxxxx = xxxxxxx
#xxxxxx_xx = xxxxx

#xxxx_xxxxxxx = xxxxxxxxxxxx(xxxxxxxxxx)

#xxxxxx_xxxxxxxxx = xxxxxxx.Xxxxxxx(xxxxx,
#                        xxxxxxxxxxx=xxxxx,
#                        xxxxxxxxxxx=xxxxxx)

def search_and_Similarity_query(query):
    #xxxx("xxxxxxxxxxxxx xxxxx:", xxxxxx)

    #xxxxx_xxxxxxx = xxxxxxxxx.xxxxxx(xxxxxx,
    #                         xxxx_xx_xxxxxr=Xxxxx)
    #xxxxxxxxxxxxx = xxxxxxxxxxx.xxxx()
    #xxxx = xxx.xxxxxX_xxxxxxxx(xxxxx_xxxxxxxxx,
    #                              xxxxxxxx_xxxxxxxx,
    #                              xxx_xxx=xxxxxx_xxxx)
    ##xxxx = XXXX[###]
    #XXXX_XXXXX = [[XXXX, xxxxx[xxxx['xxxxxx_xxx']]] for xxxxx in xxxx]
    #xxxx_xxxxxxxx = xxxxx_xxxxxxx.xxxxxxx(xxx_xxxx)
    #for xxx_xx in xxxx(xxx(xxxx_xxxx)):
    #    xxxxx[xxxx]['xxxxx-xxxxxx'] = xxxxxxxx[xxxxx]

```

```
    #xxxxxx("xxxxxxxxxxxxxxxxxxxxxxxxxxxx")

    #xxxxx("xxx xxxxx_xxxxxxx ")
    #xxxxx = xxxxxxx(xxxxx,
    #             xxxx=xxxxxxx x: x['xxxxxx'],
    #             xxxxx=Xxxx)
    #xxxxxx_xxx = xxxxx()
    #for xxx in xxxx[##:###]:
    #    xxxx_xxxxx += xxxxx['Xxxxx']
    #    xxxxx("\t{:.#f}\t{}".xxxxx(xxxxxx['xxxx'],
    #             xxxx_xxx[xxxx['xxx_xxxxx']].xxxxxx("\n", " ")))
    #xxxxxx("\nxxxxxxxxxnxxxxxxxxxxxxxxx")
    #xxxx("xxxx xxxxx xxx xx.xxxxxxxx:",
    #             "{:.#f}".xxxxxxx(xxxxxxx/#))

    #xxxx("xxxx xxxxx xxx xx.xxxxxxxx:",

    #xxxx("xxx xxxxx Xxxxxxxx")
    #xxxxx = xxxxxx(hxxxx,
    #             xxxxx=xxx x: x['xxxxxxxe'],
    #             xxxxx=xxxx)
    #xxxxx_x = fxxxxx()
    #for xxxx in xxxxx[#:#]:
    #    xxxx_xxxx += xxx['xxxxxx-xxxxx']
    #    xxxxxx("\t{:.#f}\t{}".xxxxx(xxx['xxx-xxxxx'],
    #             xxxx[xxxxx['xxx_xxxx']].xxxx("\n", " ")))

    #xxxxxx("\nxxxxxxxxxxxxxxxxxxxn")
    #xxx("xxxxxx xxxxx xxxxx xxxxx xxxxx:",
    #        "{:.#f}".xxxxx(xxxx_xxxx/####))
    #xxxxxx("\nxxxxxxxxxxxxxxxxxxxn")
    #xxxxxx("xxxxx xxxxxx xxxxxxx xxxxxx:",
    #       "{:.#f}".xxxxxxx(xxxxxxx/######))
    pass

#################################################################
qyery = '' #define query
search_and_Similarity_query(query)
```

Para obtener más información y el código fuente póngase en contacto con el autor de este documento al correo electrónico: ecaceres3941@universidadean.edu.co

B. Anexo 2 Recolección de Datos

Algoritmo Python 4: Busqueda de Información

```
#import xxxxxx.xxxxxxxx as xxxxxxx
#from xxxx import xxxxxxxxx
#import xxxx as xxxxxx
#import xxxxxx as xxxxx
#import xxxxxxxx
#import xxxx
#import xxxx

#xxxxxxx = xxx.xxxx.xxxxxxx(__file__)
#xxxxx = xxx.xxxxx.xxxxx(xxxxx_xxx, 'XXX')

def xx_xxxx_Xxxx_Xxxx(xxxx):
    #return xxxx.xxxxxxx_xxx("XXXXX")
    pass

def xxx_xxxxxxxx_xxxxx(XXXXX):
    #xxxx = xxxxxxxxx.xxxxx(xxxxx)
    #xxxx = xxxxxxxxxxxx(xxxxx.xxxxx, "xxxx.xxxxxxx")
    #xxxxx = xxxxx.xxxx_xxxx('xxxx',{'xxxxx':'xxx'}, xxxxx=###)
    #xxx_xxxx =[k.xxx_xxxxx() for k in xxxxx ]
    #xxxxxx = xxxx.xxxx_xxxx('xxx',{'xxxxx':'xxxxx'})
    #xxxx_xxxx = [v.xxxx_xxxx() for v in xxxxxx]
    #xxxx = xxx.xxxx__xxxxxx(xxx_xxxxx, xxx(xxx_xxxxx)/#####)
    #xxx_xxxx = [a.xxxx("xxxxx") for a in xxxxxx]
    #xxx = xxxx.xxxxxx(xxxxx,xxxxxx=xxx_x_xxx)
    #xxxx = xxxx.xxxxxxx('=$&([^ ]*)',xxxxxx)
    #xxxxxx.xxxxxx_Xxxx(xxx,
    #                   xxxxx+xxxx.xxxx+xxxx+'.arrow',
    #                   xxxxxx='xxxx')

    #with xxxxx(xxxxx + xxxx.xxxxx + xxxxx + '_xxxx.txt',"w") as f:
    #    for xxxx in xxxx_xxxxxx:
    #        f.xxxxxx(xxxxxxx + "\r\n\n")

    pass

def xxx_xxxx_xxxxx_xxxx(XXX):
    #xxxx = xxxxxx.xxxxx(XXX)
    #xxxx = xxxxxxxxxx(xxxxx.xxxxxx,"xxxx.xxxxxx")

    #xxxxx = xxxx.xxxx_x_xxx("xxx",
    #                   {'xxxxx':'xxxxxxx-xxxxxxxx'})
    #xxxxx_xxxx = [p.xxxxx_xxx() for p in xxxxxx]
```

```
    #xxX_xxxx = re.xxxxxx('?i=([^ ]*)', url)

    #with xxxxx(xxxxx + xxx.xxxxx + xxx_xxxx+'.xxxx',"w") as f:
    #    for xxxx in xxxx_xxxxx:
    #        f.xxxxxx(xxxxxx + "\r\n\n")

    #xxxx= xxxx.xxxxx(xxx_xxxx, xxxxxx=['xxxxxx'])
    #xxxxx['xxxx']= xxxxx
    #xxxx.xxxxx_xxxxxxx(xxx,
    #                   xxx+xxxx.xxxx+xxx_xxx+'.arrow',
    #                   xxxxxxxx='xxxxx')
    pass
```

Para obtener más información y el código fuente póngase en contacto con el autor de este documento al correo electrónico: ecaceres3941@universidadean.edu.co

Referencias

Acheampong, F. A., Nunoo-Mensah, H., & Chen, W. (2021). Transformer models for text-based emotion detection: a review of BERT based approaches. *Artificial Intelligence Review*, *54*(8), 5789-5829.

Adhikari, S., y col. (2020). Nlp based machine learning approaches for text summarization. *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, 535-538.

Ahlawat, S., Choudhary, A., Nayyar, A., Singh, S., & Yoon, B. (2020). Improved handwritten digit recognition using convolutional neural networks (CNN). *Sensors*, *20*(12), 3344.

Alammar, J. (2018). https://jalammar.github.io/illustrated-transformer/

Alesina, A. (2005). *Institutional reforms: The case of Colombia*. MIT press.

Anand, D., & Wagh, R. (2019). Effective deep learning approaches for summarization of legal texts. *Journal of King Saud University-Computer and Information Sciences*.

Arrow, K. J. (1974). *The limits of organization*. WW Norton & Company.

Arslan, S. (2022). A hybrid forecasting model using LSTM and Prophet for energy consumption with decomposition of time series data. *PeerJ Computer Science*, *8*, e1001.

Ashley, K., Branting, K., Margolis, H., & Sunstein, C. R. (2001). Legal Reasoning and Artificial Intelligence: How Computers "Think"Like Lawyers. *University of Chicago Law School Roundtable*, *8*(1), 1-28.

Aumann, R. J., & Shapley, L. S. (2015). *Values of non-atomic games*. Princeton University Press.

Aziz, H. M., Sorguli, S., Hamza, P. A., Sabir, B. Y., Qader, K. S., Ismeal, B. A., Anwar, G., Gardi, B., y col. (2021). Factors affecting International Finance Corporation. *International Journal of Humanities and Education Development (IJHED)*, *3*(3), 148-157.

Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). Layer normalization. *arXiv preprint arXiv:1607.06450*.

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Bancolombia, G. (2022). https://www.bancolombia.com/

Bast, H., Buchhold, B., Haussmann, E., y col. (2016). Semantic search on text and knowledge bases. *Foundations and Trends® in Information Retrieval*, *10*(2-3), 119-271.

Belkaoui, A., & Karpik, P. G. (1989). Determinants of the corporate decision to disclose social information. *Accounting, Auditing & Accountability Journal*, *2*(1), 0–0.

Berger-Walliser, G., & Scott, I. (2018). Redefining corporate social responsibility in an era of globalization and regulatory hardening. *American Business Law Journal*, *55*(1), 167-218.

Bojar, O., Buck, C., Federmann, C., Haddow, B., Koehn, P., Leveling, J., Monz, C., Pecina, P., Post, M., Saint-Amand, H., y col. (2014). Findings of the 2014 workshop on statistical machine translation. *Proceedings of the ninth workshop on statistical machine translation*, 12-58.

Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

bvc - bolsa de valores de colombia. (2020). *a2censo*. https://a2censo.com

Calin, O. (2020). *Deep learning architectures*. Springer.

Cárdenas, M., Junguito, R., & Pachón, M. (2008). Political institutions and policy outcomes in Colombia: The effects of the 1991 constitution. *Policymaking in Latin America: how politics shapes policies*, 199-242.

Chellapilla, K., Puri, S., & Simard, P. (2006). High performance convolutional neural networks for document processing. *Tenth international workshop on frontiers in handwriting recognition*.

Cho, K., Van Merriënboer, B., Bahdanau, D., & Bengio, Y. (2014). On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.

Chowdhary, K. (2020). Natural language processing. *Fundamentals of artificial intelligence*, 603-649.

Collins, A., Brown, J. S., & Larkin, K. M. (2017). Inference in text understanding. En *Theoretical issues in reading comprehension* (pp. 385-408). Routledge.

Conneau, A., Lample, G., Rinott, R., Williams, A., Bowman, S. R., Schwenk, H., & Stoyanov, V. (2018). XNLI: Evaluating cross-lingual sentence representations. *arXiv preprint arXiv:1809.05053*.

Corficolombiana. (2021). https://www.corficolombiana.com

Cummings, K. M., Morley, C., Horan, J., Steger, C., & Leavell, N.-R. (2002). Marketing to America's youth: evidence from corporate documents. *Tobacco control*, *11*(suppl 1), i5-i17.

Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q. V., & Salakhutdinov, R. (2019). Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*.

Dale, R. (2019). Law and word order: NLP in legal tech. *Natural Language Engineering*, *25*(1), 211-217.

Davivienda, F. (2022). *Fondo de Inversion Colectiva Renta Fija*. https://fidudavivienda.davivienda.com

De Beaugrande, R., & Dessler, W. (1997). *Introducciona la linguistica del texto*. Barcelona: Ariel Linguistica.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dos Santos, C., & Gatti, M. (2014). Deep convolutional neural networks for sentiment analysis of short texts. *Proceedings of COLING 2014, the 25th international conference on computational linguistics: technical papers*, 69-78.

dos Santos, F. F., Pimenta, P. F., Lunardi, C., Draghetti, L., Carro, L., Kaeli, D., & Rech, P. (2018). Analyzing and increasing the reliability of convolutional neural networks on GPUs. *IEEE Transactions on Reliability*, *68*(2), 663-677.

Dubois, C. (2021). How do lawyers engineer and develop legaltech projects?: A story of opportunities, platforms, creative rationalities, and strategies. *Law, Technology and Humans*, *3*(1), 68-81.

Ethayarajh, K., & Jurafsky, D. (2021). Attention flows are shapley value explanations. *arXiv preprint arXiv:2105.14652*.

Eyre, H., Chapman, A. B., Peterson, K. S., Shi, J., Alba, P. R., Jones, M. M., Box, T. L., DuVall, S. L., & Patterson, O. V. (2021). Launching into clinical space with medspaCy: a new clinical text processing toolkit in Python. *AMIA Annual Symposium Proceedings*, *2021*, 438.

Ferreira, J. Z., Rodrigues, J., Cristo, M., & de Oliveira, D. F. (2014). Multi-entity polarity analysis in financial documents. *Proceedings of the 20th Brazilian Symposium on Multimedia and the Web*, 115-122.

Ferrucci, D., & Lally, A. (2004). UIMA: an architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*, *10*(3-4), 327-348.

Fukushima, K., & Miyake, S. (1982). Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. En *Competition and cooperation in neural nets* (pp. 267-285). Springer.

Gabrilovich, E., & Markovitch, S. (2009). Wikipedia-based semantic interpretation for natural language processing. *Journal of Artificial Intelligence Research*, *34*, 443-498.

Gage, P. (1994). A new algorithm for data compression. *C Users Journal*, *12*(2), 23-38.

Gardner, M., Grus, J., Neumann, M., Tafjord, O., Dasigi, P., Liu, N., Peters, M., Schmitz, M., & Zettlemoyer, L. (2018). Allennlp: A deep semantic natural language processing platform. *arXiv preprint arXiv:1803.07640*.

Gavali, P., & Banu, J. S. (2020). Bird species identification using deep learning on GPU platform. *2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)*, 1-6.

Gehring, J., Auli, M., Grangier, D., Yarats, D., & Dauphin, Y. N. (2017). Convolutional sequence to sequence learning. *International conference on machine learning*, 1243-1252.

Gers, F. A., Schmidhuber, J., & Cummins, F. (2000). Learning to forget: Continual prediction with LSTM. *Neural computation*, *12*(10), 2451-2471.

Ghaderi, S. (s.f.). Transformers in Action: Attention Is All You Need A brief survey, illustration, and implementation.

Goldberg, Y. (2016). A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research*, *57*, 345-420.

Graën, J., Bertamini, M., Volk, M., Cieliebak, M., Tuggener, D., & Benites, F. (2018). Cutter: a universal multilingual tokenizer. *CEUR Workshop Proceedings*, (2226), 75-81.

Gramegna, A., & Giudici, P. (2020). Why to buy insurance? an explainable artificial intelligence approach. *Risks*, *8*(4), 137.

Graves, A., Jaitly, N., & Mohamed, A.-r. (2013). Hybrid speech recognition with deep bidirectional LSTM. *2013 IEEE workshop on automatic speech recognition and understanding*, 273-278.

Gross, C. G., Rocha-Miranda, C. d., & Bender, D. (1972). Visual properties of neurons in inferotemporal cortex of the Macaque. *Journal of neurophysiology*, *35*(1), 96-111.

Gruber, N., & Jockisch, A. (2020). Are GRU cells more specific and LSTM cells more sensitive in motive classification of text? *Frontiers in artificial intelligence*, *3*, 40.

Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S. R., & Smith, N. A. (2018). Annotation artifacts in natural language inference data. *arXiv preprint arXiv:1803.02324*.

Hahn, T. B. (1998). Text retrieval online: historical perspective on web search engines.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., ... Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357-362. https://doi.org/10.1038/s41586-020-2649-2

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770-778.

Hernández Osuna, S., & Ferreira Cabrera, A. (2016). Procesamiento semantico automatico, enfocado en la coherencia textual, para apoyar la produccion escrita de noticias. *Estudios filologicos*, (58), 97-122.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735-1780.

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, *160*(1), 106.

Hubel, D. H., & Wiesel, T. N. (1965). Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat. *Journal of neurophysiology*, *28*(2), 229-289.

Johnson, J., Douze, M., & Jégou, H. (2019). Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, *7*(3), 535-547.

Johnson, R., & Zhang, T. (2014). Effective use of word order for text categorization with convolutional neural networks. *arXiv preprint arXiv:1412.1058*.

Kalchbrenner, N., & Blunsom, P. (2013). Recurrent continuous translation models. *Proceedings of the 2013 conference on empirical methods in natural language processing*, 1700-1709.

Kang, Y., Cai, Z., Tan, C.-W., Huang, Q., & Liu, H. (2020). Natural language processing (NLP) in management research: A literature review. *Journal of Management Analytics*, *7*(2), 139-172.

Khaled, R., Ali, H., & Mohamed, E. K. (2021). The Sustainable Development Goals and corporate sustainability performance: Mapping, extent and determinants. *Journal of Cleaner Production*, *311*, 127599.

Kownatzki, M., Walter, J., Floyd, S. W., & Lechner, C. (2013). Corporate control and the speed of strategic business unit decision making. *Academy of Management Journal*, *56*(5), 1295-1324.

Lample, G., & Conneau, A. (2019). Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Lax, J. R. (2011). The new judicial politics of legal doctrine. *Annual Review of Political Science*, *14*, 131-157.

LeClair, A., Jiang, S., & McMillan, C. (2019). A neural model for generating natural language summaries of program subroutines. *2019 IEEE/ACM 41st International Conference on Software Engineering (ICSE)*, 795-806.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278-2324.

LeCun, Y., Jackel, L., Bottou, L., Brunot, A., Cortes, C., Denker, J., Drucker, H., Guyon, I., Muller, U., Sackinger, E., y col. (1995). Comparison of learning algorithms for handwritten digit recognition. *International conference on artificial neural networks*, *60*(1), 53-60.

Lee, K.-M., & Park, C.-W. (2018). A Study on Voice Command Learning of Smart Toy using Convolutional Neural Network. *The transactions of The Korean Institute of Electrical Engineers*, *67*(9), 1210-1215.

Lee-Thorp, J., Ainslie, J., Eckstein, I., & Ontanon, S. (2021). Fnet: Mixing tokens with fourier transforms. *arXiv preprint arXiv:2105.03824*.

Lehnert, W. G. (1977). A conceptual theory of question answering. *Proceedings of the 5th international joint conference on Artificial intelligence-Volume 1*, 158-164.

Lehnert, W. G. (1981). Plot units and narrative summarization. *Cognitive science*, *5*(4), 293-331.

Leung, H., & Haykin, S. (1991). The complex backpropagation algorithm. *IEEE Transactions on signal processing*, *39*(9), 2101-2104.

Lewis, C., & Young, S. (2019). Fad or future? Automated analysis of financial text and its implications for corporate reporting. *Accounting and Business Research*, *49*(5), 587-615.

Liberman, M. Y. (1991). The trend towards statistical models in natural language processing. En *Natural Language and Speech* (pp. 1-7). Springer.

Lim, M.-H., & Chung, J. Y. (2021). The effects of female chief executive officers on corporate social responsibility. *Managerial and Decision Economics*, *42*(5), 1235-1247.

Liu, F., Jiao, Y., Massiah, J., Yilmaz, E., & Havrylov, S. (2021). Trans-Encoder: Unsupervised sentence-pair modelling through self-and mutual-distillations. *arXiv preprint arXiv:2109.13059*.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Liu, Y., Fan, B., Xiang, S., & Pan, C. (2019). Relation-shape convolutional neural network for point cloud analysis. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8895-8904.

Mandal, A., Chaki, R., Saha, S., Ghosh, K., Pal, A., & Ghosh, S. (2017). Measuring similarity among legal court case documents. *Proceedings of the 10th annual ACM India compute conference*, 1-9.

Masson, C., & Paroubek, P. (2020). Nlp analytics in finance with dore: A french 250m tokens corpus of corporate annual reports. *Proceedings of the 12th Language Resources and Evaluation Conference*, 2261-2267.

McGinnis, J. O., & Pearce, R. G. (2019). The great disruption: How machine intelligence will transform the role of lawyers in the delivery of legal services. *Actual Probs. Econ. & L.*, 1230.

Moschitti, A., Pang, B., & Daelemans, W. (2014). Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Munisami, K. (2019). Legal Technology and the Future of Women in Law. *Windsor Yearbook of Access to Justice/Recueil annuel de Windsor d'accès à la justice*, *36*, 164-183.

Ngo, T. D., Bui, T. T., Pham, T. M., Thai, H. T., Nguyen, G. L., & Nguyen, T. N. (2021). Image deconvolution for optical small satellite with deep learning and real-time GPU acceleration. *Journal of Real-Time Image Processing*, *18*(5), 1697-1710.

Nguyen, H. V., & Bai, L. (2010). Cosine similarity metric learning for face verification. *Asian conference on computer vision*, 709-720.

Oliva, F. L., Teberga, P. M. F., Testi, L. I. O., Kotabe, M., Del Giudice, M., Kelle, P., & Cunha, M. P. (2022). Risks and critical success factors in the internationalization of born global startups of industry 4.0: A social, environmental, economic, and institutional analysis. *Technological Forecasting and Social Change*, *175*, 121346.

pandas development team, T. (2022). *pandas-dev/pandas: Pandas* (Ver. v1.5.2) [If you use this software, please cite it as below.]. Zenodo. https://doi.org/10.5281/zenodo.7344967

Park, I., & Yoon, B. (2014). A semantic analysis approach for identifying patent infringement based on a product – patent map. *Technology Analysis & Strategic Management*, *26*(8), 855-874.

Parsing, C. (2009). Speech and language processing.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., ... Chintala, S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. En *Advances in Neural Information Processing Systems 32* (pp. 8024-8035). Curran Associates, Inc. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

Pazzani, M. J. (1983). INTERACTWE SCRIPT INSTANTIATION.

Post, F. R. (2003). A response to "the social responsibility of corporate management: a classical critique". *American Journal of Business*.

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., y col. (2018). Improving language understanding by generative pre-training.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., y col. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

Rahman, M. M., & Finin, T. (2019). Unfolding the Structure of a Document using Deep Learning. *arXiv preprint arXiv:1910.03678*.

Rahutomo, F., Kitasuka, T., & Aritsugi, M. (2012). Semantic cosine similarity. *The 7th international student conference on advanced science and technology ICAST*, *4*(1), 1.

Richardson, L. (2007). Beautiful soup documentation. *April*.

Richardson, N., Cook, I., Crane, N., Dunnington, D., François, R., Keane, J., Moldovan-Grünfeld, D., Ooms, J., & Apache Arrow. (2022). *arrow: Integration to 'Apache' 'Arrow'* [https://github.com/apache/ https://arrow.apache.org/docs/r/].

Rubin, E., & Feeley, M. (1995). Creating Legal Doctrine. *S. Cal. L. Rev.*, *69*, 1989.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *nature*, *323*(6088), 533-536.

Rundo, F., Trenta, F., di Stallo, A. L., & Battiato, S. (2019). Machine learning for quantitative finance applications: A survey. *Applied Sciences*, *9*(24), 5574.

Salmerón-Manzano, E. (2021). Legaltech and Lawtech: Global Perspectives, Challenges, and Opportunities. *Laws*, *10*(2), 24.

Sanyal, S., Balachandran, J., Yadati, N., Kumar, A., Rajagopalan, P., Sanyal, S., & Talukdar, P. (2018). MT-CGCNN: Integrating crystal graph convolutional neural network with multitask learning for material property prediction. *arXiv preprint arXiv:1811.05660*.

Schank, R. C., & Abelson, R. P. (1975). Scripts, plans, and knowledge. *IJCAI*, *75*, 151-157.

Schmidhuber, J. (1993). Habilitation thesis: System modeling and optimization. *Page 150 ff demonstrates credit assignment across the equivalent of 1,200 layers in an unfolded RNN*.

Schulder, M., Wiegand, M., Ruppenhofer, J., & Roth, B. (2017). Towards bootstrapping a polarity shifter lexicon using linguistic features. *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 624-633.

Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, *45*(11), 2673-2681.

Seltzer, L. H., Starks, L., & Zhu, Q. (2022). *Climate regulatory risk and corporate bonds* (inf. téc.). National Bureau of Economic Research.

Sennrich, R., Haddow, B., & Birch, A. (2015). Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Shen, Y., He, X., Gao, J., Deng, L., & Mesnil, G. (2014). A latent semantic model with convolutional-pooling structure for information retrieval. *Proceedings of the 23rd ACM international conference on conference on information and knowledge management*, 101-110.

Silva, S. (2021). Corporate contributions to the Sustainable Development Goals: An empirical analysis informed by legitimacy theory. *Journal of Cleaner Production*, *292*, 125962.

Sleimi, A., Sannier, N., Sabetzadeh, M., Briand, L., & Dann, J. (2018). Automated extraction of semantic legal metadata using natural language processing. *2018 IEEE 26th International Requirements Engineering Conference (RE)*, 124-135.

Song, C., Rush, A. M., & Shmatikov, V. (2020). Adversarial semantic collisions. *arXiv preprint arXiv:2011.04743*.

Soukupova, J. (2021). AI-based Legal Technology: A Critical Assessment of the Current Use of Artificial Intelligence in Legal Practice. *Masaryk UJL & Tech.*, *15*, 279.

Stagner, R. (1969). Corporate decision making: An empirical study. *Journal of Applied Psychology*, *53*(1p1), 1.

Strigl, D., Kofler, K., & Podlipnig, S. (2010). Performance and scalability of GPU-based convolutional neural networks. *2010 18th Euromicro conference on parallel, distributed and network-based processing*, 317-324.

Sun, Y., Lin, L., Tang, D., Yang, N., Ji, Z., & Wang, X. (2015). Modeling mention, context and entity with neural networks for entity disambiguation. *Twenty-fourth international joint conference on artificial intelligence*.

Sundermeyer, M., Alkhouli, T., Wuebker, J., & Ney, H. (2014). Translation modeling with bidirectional recurrent neural networks. *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 14-25.

Susskind, R. (2008). *The end of lawyers*. Oxford: Oxford University Press.

Susskind, R. E., & Susskind, D. (2015). *The future of the professions: How technology will transform the work of human experts*. Oxford University Press, USA.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, *27*.

Szostek, D. (2021). The Concept of Legal Technology (LegalTech) and Legal Engineering. *Legal Tech*, 19-28.

Taruffo, M. (2007). Precedente y jurisprudencia. *Precedente. Revista Juridica*, 86-99.

Tiller, E. H., & Cross, F. B. (2006). What is legal doctrine. *Nw. UL Rev.*, *100*, 517.

Tunstall, L., von Werra, L., & Wolf, T. (2022). *Natural language processing with transformers*. .ºReilly Media, Inc.".

Vaki. (2020). Consultado en 2022, desde https://vaki.co/es/

Van Rossum, G., & Drake, F. L. (2009). *Python 3 Reference Manual*. CreateSpace.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Vidal, F. M. C. (1991). *La jurisprudencia: fuente del derecho* (Tesis doctoral). Universidad de Valladolid.

von Philipsborn, P., Geffert, K., Klinger, C., Hebestreit, A., Stratil, J., Rehfuess, E. A., Consortium, P., y col. (2022). Nutrition policies in Germany: a systematic assessment with the Food Environment Policy Index. *Public Health Nutrition*, *25*(6), 1691-1700.

Webber, B., Cohn, T., He, Y., & Liu, Y. (2020). Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Webley, L., Flood, J., Webb, J., Bartlett, F., Galloway, K., & Tranter, K. (2019). The profession (s)'engagements with lawtech: Narratives and archetypes of future law. *Law, Tech. & Hum.*, *1*, 6.

Weston, J., Chopra, S., & Adams, K. (2014). # tagspace: Semantic embeddings from hashtags. *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1822-1827.

Williams, A., Nangia, N., & Bowman, S. R. (2017). A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., y col. (2020). Transformers: State-of-the-art natural language processing. *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, 38-45.

Wu, C., Fan, W., He, Y., Sun, J., & Naoi, S. (2014). Handwritten character recognition by alternately trained relaxation convolutional neural network. *2014 14th International Conference on Frontiers in Handwriting Recognition*, 291-296.

Yates, A., Nogueira, R., & Lin, J. (2021). Pretrained transformers for text ranking: BERT and beyond. *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, 1154-1156.

Ye, W., Liu, Y., Zou, L., Cai, H., Cheng, S., Wang, S., & Yin, D. (2022). Fast semantic matching via flexible contextualized interaction. *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, 1275-1283.

Zaloga, A. N., Stanovov, V. V., Bezrukova, O. E., Dubinin, P. S., & Yakimov, I. S. (2020). Crystal symmetry classification from powder X-ray diffraction patterns using a convolutional neural network. *Materials Today Communications*, *25*, 101662.

Zhang, S., Cheng, H., Gao, J., & Poon, H. (2022). Optimizing Bi-Encoder for Named Entity Recognition via Contrastive Learning. *arXiv preprint arXiv:2208.14565*.

Zhang, Y., & Wallace, B. (2015). A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. *preprint arXiv:1510.03820*.

Zhong, H., Guo, Z., Tu, C., Xiao, C., Liu, Z., & Sun, M. (2018). Legal judgment prediction via topological learning, 3540-3549.

Zhou, M., Duan, N., Liu, S., & Shum, H.-Y. (2020). Progress in neural NLP: modeling, learning, and reasoning. *Engineering*, *6*(3), 275-290.

Ziletti, A., Kumar, D., Scheffler, M., & Ghiringhelli, L. M. (2018). Insightful classification of crystal structures using deep learning. *Nature communications*, *9*(1), 1-10.