Sistema de Soporte para la Recuperación de Normativas en la Ingeniería Legal

Luciano Perezzini[1,2], Ana Casali[1,2], Claudia Deco[1]

[1] Facultad de Cs. Exactas, Ingeniería y Agrimensura,
Universidad Nacional de Rosario, Rosario, Santa Fe, Argentina
lperezzini@dcc.fceia.unr.edu.ar
{acasali,deco}@fceia.unr.edu.ar
[2] Centro Internacional Franco Argentino de Cs. de la Información y de Sistemas
(CIFASIS: CONICET-UNR)

Resumen En la ingeniería legal se necesitan de agentes artificiales con la capacidad de extraer conocimiento y patrones dentro de documentos de índole legal con el propósito de diseñar aplicaciones que asistan a profesionales a realizar determinadas tareas. Muchas de éstas se deben realizar periódicamente, en donde se recuperan y analizan cientos de normativas publicadas por numerosos boletines oficiales del gobierno. Principalmente en la industria, la búsqueda de normativas relevantes para una determinada actividad se realiza manualmente e involucra numerosos profesionales en diversas materias. Con la finalidad de semi-automatizar lo anterior, en este trabajo se propone el diseño de un sistema de soporte para recuperar, de forma periódica, normativas potencialmente relevantes con respecto a las actividades realizadas por una empresa.

Palabras claves Sistemas de apoyo·Ingeniería legal·Clasificación de normativas.

1. Introducción

En la era de la información, grandes volúmenes de datos se encuentran constantemente a nuestra disposición para analizarlos con el fin de tomar decisiones. Descubrir información relevante dentro de largas cadenas de datos, por lo general no estructurados y provenientes de diversas fuentes, se ha convertido en una de las disciplinas más importantes dentro de cualquier entidad. La capacidad de aprovechar toda la información que se dispone es considerada una habilidad crítica para el éxito de una organización. Una empresa que dedica parte de sus ingresos a desarrollar esta capacidad, logra tomar decisiones más inteligentes y de forma más rápida [1]. Pero la anterior tarea no se considera sencilla, ya que por lo general es un trabajo que involucra muchos perfiles: tecnólogos, científicos de datos, informáticos y profesionales de los negocios, entre otros.

Ahora bien, el tipo de datos que juega uno de los principales roles en la era de la información es el texto: nos comunicamos utilizando lenguaje natural, producimos y consumimos enormes cantidades de datos textuales tales como páginas web, e-mails y noticias, todos los días y en diferentes circunstancias [2].

Los documentos legales, tales como normativas y fallos judiciales, son textos escritos por el ser humano para la propia emisión en distintos ámbitos de un gobierno.

La práctica de la Ley necesariamente involucra el análisis y la interpretación del lenguaje natural. Con el objeto de dar soporte al análisis de documentos legales, surge la denominada ingeniería legal (IL). Como se detalla en [3], la IL aplica nociones de las ciencias de la información, ingeniería de software e inteligencia artificial para asistir a profesionales de la Ley, y a individuos en general, en tareas de toma de decisión que involucren a la legislación. Dado que la interpretación del lenguaje natural no escapa del análisis de textos legales, no es sorprendente que muchas herramientas de la inteligencia artificial y del procesamiento del lenguaje natural (PLN), como los clasificadores de texto, sean fundamentales para la mayoría de las aplicaciones existentes de la IL. En los últimos años, el interés en dicha área ha aumentado considerablemente en el ámbito científico. A modo de ejemplo, en [4] se propone un sistema de aprendizaje automático con la intención de detectar cláusulas injustas dentro de largos contratos electrónicos, en [5] se han estudiado métodos de resumen de documentos legales, en [6] se presenta un enfoque para modelar argumentos legales para la defensa o por la fiscalía durante un juicio y en [7] se mejora el acceso a una base de datos de legislación utilizando detección automática de entidades.

En especial, las empresas necesitan del continuo análisis de normativas con el fin de corroborar su cumplimiento acorde a las actividades que éstas realizan y así, prevenir sanciones del Estado. Nuevas normativas potencialmente relevantes para una empresa son publicadas periódicamente en diferentes boletines oficiales tales como el Boletín Oficial de la República Argentina, los boletines oficiales provinciales y los boletines oficiales municipales, los cuales disponen de sus propios portales web. Cada boletín publica diariamente un considerable número de documentos, por lo cual el análisis humano de cada una de las normativas se convierte en una tarea realmente engorrosa que involucra numerosos y diversos profesionales en distintas materias. En la industria, ésta tarea es usualmente denominada *matricería legal* (ML).

Si bien el Sistema Argentino de Información Jurídica[1] (SAIJ) realiza un gran esfuerzo con respecto al etiquetado diario y manual, utilizando terminología del Tesauro del Derecho Argentino[2] (TDA), de documentos legales publicados por ciertos boletines oficiales del gobierno, al ser éste un trabajo realizado por un conjunto reducido de individuos, se puede observar que dicho sistema no se encuentra actualizado en su totalidad. Dicha problemática dificulta que ciudadanos e industriales realicen una búsqueda periódica y más inteligente de normativas, publicadas por cualquier boletín oficial del gobierno, mediante términos específicos.

En este trabajo, realizado con fines académicos durante el año 2019, con el objeto de brindar soporte a los expertos encargados en una empresa de desarrollar la mencionada ML, se propone un sistema de recomendación, en tiempo real, de normativas potencialmente relevantes con respecto a la actividad de una empresa utilizando técnicas del PLN y web crawling. Se debe destacar que, durante el período de tiempo

[1] Base de datos de documentación jurídica dependiente de la Secretaría de Justicia del Ministerio de Justicia y Derechos Humanos de la Nación. Su portal web se encuentra en `http://www.saij.gob.ar`

[2] Creación propia del SAIJ, el cual asiste en la búsqueda temática en dicho sistema. El mencionado tesauro se puede encontrar en `http://datos.jus.gob.ar/dataset/tesauro-saij-de-derecho-argentino`

en el cual se redactó este trabajo, no se encontraron servicios informáticos comerciales en la Argentina que solventen la problemática de la ML de forma automática o semi-automática; aunque se han localizado empresas particulares que realizan el mencionado procedimiento de forma manual.

El resto de esta publicación se organiza como sigue. En la siguiente sección, se realiza una breve introducción a la representación y clasificación de texto, en donde se presentan las técnicas y herramientas que se utilizan a lo largo del trabajo. En la Sección 3 se describe, primeramente, una arquitectura base de sistema de soporte para tareas de la IL. Esta arquitectura es la que se utiliza para implementar la propuesta de un sistema en línea de recomendación de normativas para asistir a profesionales encargados de la ML en empresas. Luego, en la Sección 4 se exhiben resultados de experimentación. Finalmente, en la Sección 5 se brindan conclusiones y trabajo futuro con el propósito de refinar la semi-automatización de la matricería legal.

2. La representación y el análisis de documentos de texto

Durante décadas se han estudiado diversos modelos de representación de texto. Uno de los más utilizados es el modelo algebraico denominado Espacio-Vectorial [8]. En pocas palabras, sea $\mathscr{C}$ es una colección de documentos, dicho modelo representa un documento d mediante un vector $\vec{v}_d$ de pesos en donde cada uno de éstos corresponde a un término del vocabulario de $\mathscr{C}$. La forma más aceptada en la práctica de calcular dichos pesos es mediante el producto de un peso local y un peso global. El esquema de pesaje *tf-idf* es considerado uno de los más atractivos de emplear en el modelo Espacio-Vectorial [9] y es el que se aplica a lo largo del presente trabajo.

2.1. Clasificación de texto

La clasificación de texto se considera una de las principales técnicas de análisis de texto. Como se explica en [10], la clasificación de texto es la actividad de etiquetar documentos escritos en lenguaje natural con categorías o clases temáticas pertenecientes a un conjunto predefinido. En especial, puede resolverse mediante una técnica de aprendizaje supervisado [9]: a partir de un conjunto de entrenamiento de documentos etiquetados $\mathscr{D}$ y un algoritmo de aprendizaje Γ, se aprende un modelo clasificador $\gamma = \Gamma(\mathscr{D})$.

En la literatura, se pueden distinguir dos formas de clasificación: la clasificación multi-clase y la clasificación multi-valor [11]. En la clasificación multi-clase todo documento es miembro de exactamente una única clase; mientras que en la multi-valor todo documento puede pertenecer a una única clase, a varias clases de forma simultánea o a ninguna. En la recuperación de información es común el problema multi-valor, ya que las clases, por lo general, no son mutuamente excluyentes.

Dado que la representación del texto en el modelo Espacio-Vectorial se encuentra en un espacio ralo y de gran dimensionalidad, los modelos lineales de clasificación se consideran particularmente atractivos de aplicar [9] y en especial, la máquina de vectores soporte con kernel lineal ($linSVM$) es uno de los métodos más utilizados y que brinda mejores resultados a la hora de aprender clasificadores de texto [12] [11] [9].

3. Propuesta de sistema soporte a la matricería legal

Con la finalidad de presentar la propuesta de un sistema asistente a la ML, a continuación se propone una arquitectura conceptual de un asistente artificial en línea denominado *Sistema de Soporte a la Ingeniería Legal* (SiSIL), el cual se ilustra en la Figura 1. El mencionado asistente analiza periódicamente normativas de boletines

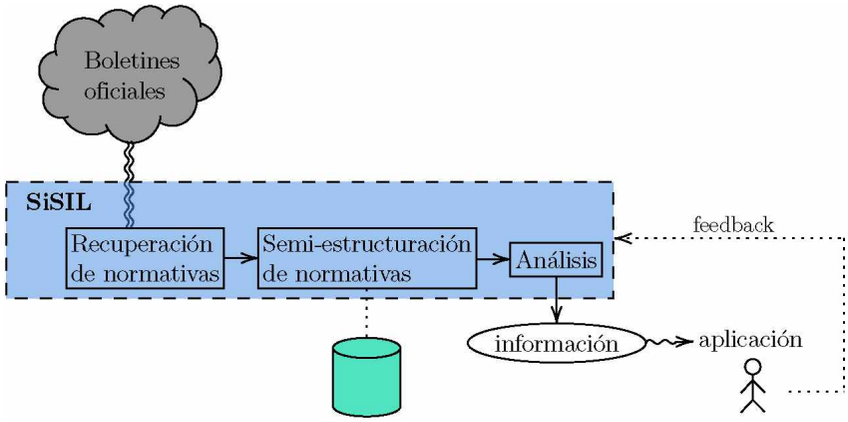

Figura 1. Propuesta de arquitectura conceptual de un asistente en línea a la ingeniería legal

oficiales del gobierno con el fin de recabar información relevante para afrontar una problemática en especial y así, asistir a los usuarios encargados de la misma. SiSIL se compone de los siguientes tres módulos:

1. **Recuperación de normativas**. Como se ha detallado anteriormente, cada boletín oficial posee su correspondiente portal web. Éstos publican toda normativa en formato HTML pero no comparten una estructura común; es decir, cada boletín emplea su propia forma de publicar documentos utilizando el mencionado lenguaje de marcado. No existe hasta el momento un estándar entre todos los boletines oficiales sobre cómo estructurar normativas para que sean publicadas en la web. Además, estos portales no poseen interfaces de programación de aplicaciones (APIs) y por lo tanto, la recuperación debe ser realizada mediante la ejecución diaria de distintos web crawlers, construidos específicamente para cada fuente de información.
2. **Semi-estructuración de normativas**. Se puede notar que algunos de estos portales emplean una estructuración HTML más fina y mejor lograda, mientras que otros implementan una estructuración considerablemente menos robusta implicando que la extracción de información mediante web crawlers sea una tarea compleja. Entre los metadatos de importancia a extraer, se pueden nombrar el tipo de normativa (ley, decreto, etc.), el número, el título, el organismo emisor, fecha de sanción o de emisión, fecha de promulgación, fecha de publicación y principalmente, el texto completo. Esta información, por lo general, no se encuentra etiquetada

específicamente en el código HTML y por lo tanto se deben definir expresiones regulares con el fin de encontrar patrones en cadenas de caracteres. Las normativas semi-estructuradas son almacenadas en una base de datos.

3. **Análisis**. Aquí, dependiendo de la problemática a afrontar, se aplican diversas técnicas del PLN a las normativas semi-estructuradas para obtener información útil que pueda asistir al experto a realizar la correspondiente tarea. Este módulo es considerado el núcleo de SiSIL, dado que emplea técnicas inteligentes para descubrir patrones los cuales serán luego explotados.

El usuario, como se observa en la Figura 1, puede brindar su feedback acerca de la calidad de los resultados retornados por la aplicación final. Esta devolución explícita o implícita, es eventualmente utilizada por SiSIL para obtener una mejor comprensión de la necesidad de información de la problemática y así, refinar la calidad de la información retornada (por ejemplo, reentrenando modelos de aprendizaje automatizado aplicados en el módulo de análisis). Como se puede observar, SiSIL deja abierta la posibilidad de incluir distintas formas de análisis de documentos normativos con el fin de asistir en diversas aplicaciones de usuario relacionadas a la toma de decisión legal.

Con el objeto de semi-automatizar la matricería legal, en este trabajo se propone utilizar SiSIL como arquitectura conceptual base para la recuperación de normativas con la finalidad de realizar un poblado continuo del repositorio de las matrices legales pertenecientes a una empresa. Particularmente en este caso, además de recuperar diariamente numerosas normativas de distintos boletines, se plantea que SiSIL determine cuáles de éstas son potencialmente relevantes con respecto a las actividades que realiza una empresa mediante la aplicación en cadena de múltiples clasificadores binarios de texto en el módulo de análisis del sistema de soporte.

Ahora bien, no todas las normativas publicadas por los boletines oficiales son relevantes para una empresa e: conocer las actividades que ésta realiza cumple un rol primordial en la selección de documentos normativos. Para ésto, se propone la construcción de un perfil de empresa Ω_e, el cual es utilizado por SiSIL durante los procedimientos de recuperación y análisis de normativas. La información que se incluye en dicho perfil se describe brevemente como sigue. Por un lado, la actividad de la ML involucra ciertas ramas del Derecho de interés en el proceso de selección de documentos normativos: la entidad suele interesarse sólo por algunas ramas dependiendo de las actividades que usualmente ejecuta. Se plantea que el experto seleccione dichas ramas mediante un diálogo con el sistema soporte, ilustrado en la Figura 2. En este trabajo se consideró subdividir el Derecho en ramas siguiendo la especificación del Tesauro del Derecho Argentino (TDA), construido por SAIJ. Luego, a partir de las principales ramas del Derecho elegidas por el usuario, el sistema presenta tópicos más específicos dentro del dominio de cada una de ellas, a fin de que éste pueda continuar refinando la necesidad de información. Por otra parte, la ubicación territorial de la empresa determina los portales web de los cuales extraer normativas periódicamente (boletín nacional, boletines provinciales y boletines municipales). Por lo tanto, las URL de éstos portales son incluidas en Ω_e.

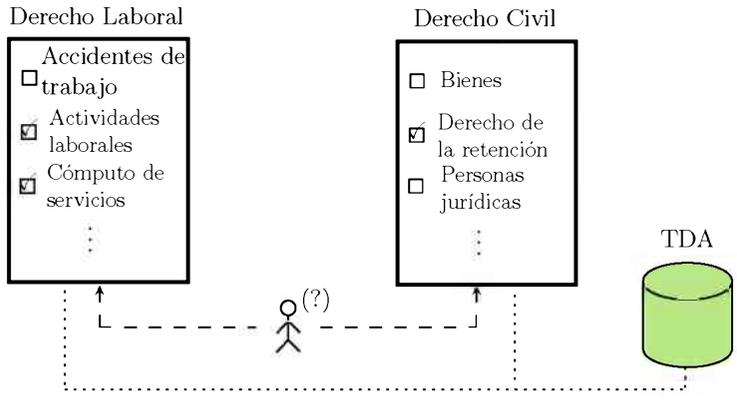

Figura 2. Diálogo entre el usuario experto y el sistema asistente

3.1. Clasificación artificial de normativas

Como se detalló en la Sección 2, la clasificación de texto es una técnica de aprendizaje automatizado. Por lo cual, aprender clasificadores de texto requiere principalmente obtener un conjunto de documentos etiquetados. A dicho conjunto de datos se le aplica un algoritmo de aprendizaje supervisado el cual, finalmente, produce un modelo matemático que toma un vector representación de un documento como entrada y devuelve información que ayuda a deducir la etiqueta de dicho documento.

Sea $\mathbb{D}$ el conjunto de las 16 principales ramas del Derecho según el criterio de subdivisión del TDA, en donde se descartó el Derecho Canónico debido a la escasez de normativas en la base de datos del SAIJ. Trivialmente, éstas no son mutuamente excluyentes (por ejemplo, el decreto nacional 720/2014 es considerado por el SAIJ como perteneciente al Derecho Ambiental y al Derecho Civil). Luego, la clasificación de documentos normativos en ramas del Derecho es un problema de clasificación multivalor: se deben aprender $|\mathbb{D}|$ clasificadores binarios; es decir, un clasificador binario de texto γ_r para cada $r \in \mathbb{D}$. La metodología que se propone durante el presente trabajo para aprender cada uno de éstos es la misma y se describe brevemente a continuación.

Como primer medida, se recuperan normativas para cada clase de $\mathbb{D}$. A lo largo de este artículo, al conjunto anterior se lo nota Λ. Λ es un conjunto de pares (d,t) donde d es un documento normativo y $t \in \mathbb{D}$ es una rama del Derecho a la cual d pertenece y $\bigcap_{r \in \mathbb{D}} \Lambda_r = \emptyset$ se satisface, con $\Lambda_r = \{d \mid (d,r) \in \Lambda\}$.

Sea $r \in \mathbb{D}$, con el fin de aprender un clasificador binario de texto γ_r tal que determine si un documento d pertenece a dicha rama r del Derecho, es decir $\gamma_r(d) \in \{r,\bar{r}\}$, se procede con la construcción de un conjunto de observaciones positivas (las normativas categorizadas mediante la rama r) y observaciones negativas (muestras aleatorias de documentos de los restantes $|\mathbb{D}|-1$ conjuntos de normativas Λ_i, con $i \neq r$). De esta forma, se define un conjunto de observaciones $\mathscr{D}_{\Lambda_r} = \mathscr{D}^{+}_{\Lambda_r} \cup \mathscr{D}^{-}_{\Lambda_r}$, donde *rand* es una

función aleatoria específicamente diseñada para este problema:

$$\mathscr{D}^{+}_{\Lambda_r} = \{(d,r) \mid d \in \Lambda_r\},$$

$$\mathscr{D}^{-}_{\Lambda_r} = \{(d,\bar{r}) \mid d \in \bigcup_{i \in \mathbb{D}-\{r\}} rand(\left\lfloor \frac{|\Lambda_r|}{|\mathbb{D}|-1} \right\rfloor, \Lambda_i)\}$$

Esta propuesta de construcción del conjunto $\mathscr{D}_{\Lambda_r}$ tal que posea todas las observaciones positivas y un subconjunto aleatorio de observaciones negativas, tiene como propósito obtener un conjunto aproximadamente balanceado en un escenario de clases desbalanceadas. La Figura 3 intenta ilustrar dicha propuesta. La estrategia anterior deriva de la técnica denominada submuestreo aleatorio[3].

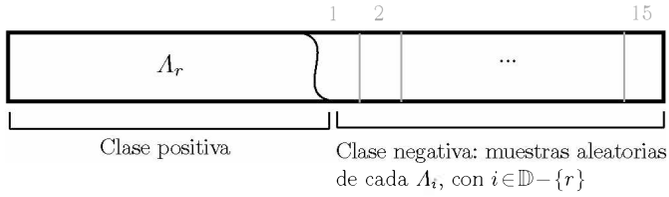

Figura 3. Construcción del conjunto aproximadamente balanceado de observaciones $\mathscr{D}_{\Lambda_r}$

Una vez construido $\mathscr{D}_{\Lambda_r}$, se procede con el aprendizaje y validación del clasificador γ_r, aplicando $linSVM$ como algoritmo de aprendizaje. Particularmente, se realizan los siguientes pasos:

1. **Partición del conjunto total**. Se procede con la partición de $\mathscr{D}_{\Lambda_r}$, de forma aleatoria y estratificada, en conjuntos disjuntos de entrenamiento y validación. Éstos se notan $\mathscr{T}$ y $\mathscr{V}$, respectivamente. $\mathscr{V}$ es totalmente aislado del proceso de aprendizaje y se utiliza para estimar el error de generalización del modelo aprendido.
2. **Ajuste del hiperparámetro de regularización**. $linSVM$ posee un hiperparámetro de regularización denominado C. Sea V_C un conjunto de posibles configuraciones de C. Utilizando cada configuración del anterior conjunto, se aplica la técnica de remuestreo k-fold CV estratificado en el conjunto de entrenamiento $\mathscr{T}$ resultando en múltiples modelos y estimaciones de errores de generalización. El procedimiento completo, en la literatura, se denomina *grid search*.
3. **Entrenamiento**. Tomando la configuración del hiperparámetro C con mejor resultado de generalización, se aplica el algoritmo de aprendizaje $linSVM$ en el conjunto de observaciones de entrenamiento $\mathscr{T}$ para obtener el clasificador γ_r; es decir, $\gamma_r = linSVM(\mathscr{T})$.
4. **Validación**. Utilizando el conjunto de observaciones $\mathscr{V}$ apartado al comienzo, se evalúa la generalización del modelo aprendido γ_r.

Esta propuesta de aprendizaje de clasificadores binarios en ramas del Derecho se utiliza para aprender todo γ_r con $r \in \mathbb{D}$.

[3] La técnica de submuestreo aleatorio equilibra las distribuciones de clase descartando, al azar, instancias de la clase mayoritaria [13]

Detección de tópicos relevantes No toda normativa clasificada mediante alguna rama del Derecho de interés para la empresa es del todo relevante para la misma. Por ejemplo, una empresa puede estar principalmente interesada en el tópico *desechos peligrosos* dentro del Derecho Ambiental. Por lo tanto, cada normativa clasificada con alguna rama del Derecho de interés para la empresa debe continuar siendo evaluada por el sistema de soporte con la finalidad de que éste logre una selección de normativas más específica. Para afrontar esta problemática, se propone que el experto de la empresa explore el dominio de cada rama del Derecho de interés mediante la interfaz del sistema, la cual fue ilustrada en la Figura 2. De esta forma, el sistema refina su noción acerca de la necesidad de información y mejora la capacidad de recuperar aquellas normativas que potencialmente traten sobre términos más específicos de las ramas del Derecho detectadas. Esta etapa no es sencilla y posiblemente requiera aplicar una combinación de técnicas de aprendizaje automatizado con otras herramientas semánticas.

Para realizar una primera prueba de viabilidad, se propuso aprender clasificadores binarios de texto: es decir, por cada rama $r \in \mathbb{D}$ de interés para la empresa, se plantea aprender un clasificador de texto por cada término más específico t escogido por el usuario en el diálogo de solicitud de requerimientos. Cada uno de estos modelos se aprende de manera análoga a los clasificadores de ramas del Derecho, a partir de la construcción de un conjunto de observaciones $\mathscr{D}_t$ utilizando sólo los documentos del conjunto Λ_r, dado que se busca aprender sobre un término específico del dominio del Derecho r. Nuevamente, se debe enfrentar un problema de clases desbalanceadas: la cardinalidad del conjunto de todos los documentos dentro de la rama r del Derecho que además están etiquetados mediante la clase t (observaciones positivas) es mucho menor que la cardinalidad del conjunto de los restantes documentos normativos de ese conjunto (observaciones negativas). Para solventar lo anterior, se propone aplicar la técnica de submuestreo aleatorio a la clase mayoritaria; es decir, a la clase negativa. El aprendizaje y la validación de estos clasificadores de texto se realiza tal como se describió anteriormente para los clasificadores en ramas del Derecho aplicando, nuevamente, $linSVM$ como algoritmo de aprendizaje.

3.2. Aplicación de usuario

En la Figura 4 se presenta la aplicación de usuario diseñada a partir de una implementación específica de la arquitectura SiSIL para dar soporte a la matricería legal en una empresa. Allí, se ilustra la interacción entre el sistema de soporte propuesto y el encargado de realizar la ML en una empresa e, donde B es el conjunto de boletines oficiales preseleccionados y M es el repositorio de las matrices legales las cuales los expertos en el contexto de e actualizan de forma diaria. SiSIL, mediante la información del perfil Ω_e, recupera diaria y automáticamente conjuntos de normativas publicadas por los boletines oficiales que se encuentran en el conjunto B. Luego, aplicando los clasificadores de texto en ramas del Derecho y los clasificadores de tópicos de interés aprendidos de forma offline, analiza cada una de las anteriores normativas y detecta aquellas que considera potencialmente relevantes para las actividades realizadas por e. Esta aplicación en cadena de clasificadores de texto se ilustra en la Figura 5: el sistema de soporte considera relevante a una normativa si además de ser clasificada primeramente mediante alguna rama del Derecho de interés $r \in \mathbb{D}$ especificada por

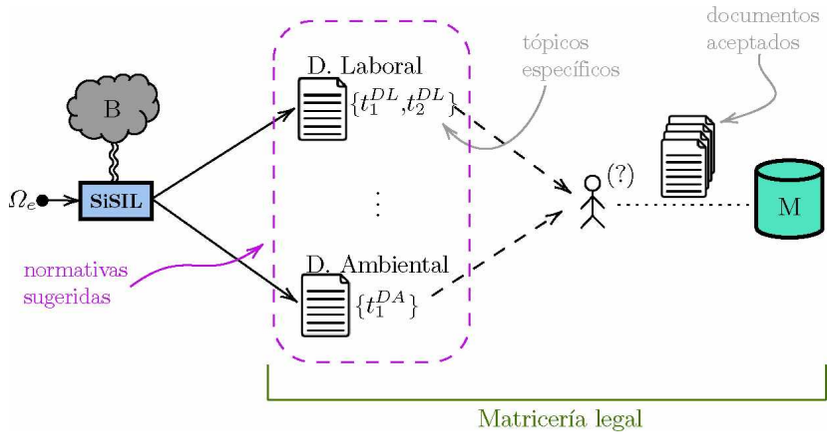

Figura 4. Ejemplo ilustrativo de la aplicación de la matricería legal en una empresa e utilizando SiSIL como agente de soporte artificial

e, se considera que trata sobre algún tópico más especifico del dominio de r. Las normativas filtradas son finalmente presentadas al experto, quien es el encargado de tomar la última decisión sobre aceptarlas o rechazarlas.

Como se puede notar, el usuario sólo debe analizar las normativas sugeridas por el sistema. Sin esta asistencia, los encargados de la ML deberían manualmente recuperar todas las normativas de los boletines oficiales seleccionados y analizar detalladamente el contenido textual de cada una de ellas.

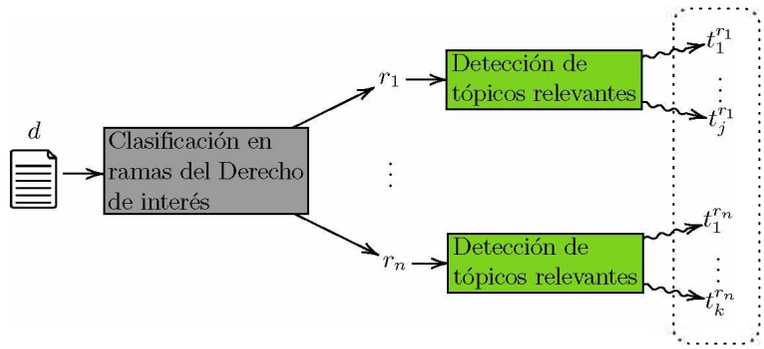

Figura 5. Análisis de una normativa d realizado por SiSIL específicamente para la problemática de la ML. Las hojas del árbol son la respuesta de clasificación de d con respecto a cada rama del Derecho y tópico de interés para una empresa en particular

Implementación En cuanto a la implementación del prototipo, se utilizaron las siguientes herramientas. Para el desarrollo de cada web crawler se decidió utilizar la

librería Scrapy[4], escrita en el lenguaje de programación Python. Con respecto al aprendizaje de los modelos clasificadores de texto, se usó la librería dedicada al aprendizaje automatizado llamada Scikit-learn[5], escrita también en el mencionado lenguaje.

4. Experimentación

Para analizar esta propuesta se realizó una experimentación en dos etapas. En primer lugar se evaluó el comportamiento de los clasificadores aprendidos para las distintas ramas del Derecho. En segundo lugar, se hizo una primera prueba del sistema de soporte a la ML propuesto. Esta última consiste desde el crawling de normativas de seleccionados boletines oficiales hasta la clasificación de éstas en ramas del Derecho y detección de tópicos de interés.

4.1. Evaluación de clasificadores de texto en ramas del Derecho y tópicos de interés

Primeramente, se entrenaron y validaron los 16 clasificadores de ramas del Derecho sobre un corpus obtenido de la base de datos del SAIJ. En la Figura 6 se exhibe información sobre dicho corpus, con un total de 26520 normativas extraídas. A continuación, se presentan resultados obtenidos de la validación de los mencionados clasificadores, en donde se detalla la validación del clasificador de normativas del Derecho Ambiental. Luego, se presentan los resultados de los 15 restantes clasificadores. Cada uno de éstos fueron entrenados y validados tal como se describió en la Sección 3.1. Gracias a la metodología propuesta para entrenar y validar cada clasificador de texto, todo conjunto de datos de validación resulta aproximadamente balanceado. Por lo tanto, se considera que el resultado de la medición de exactitud (ACC)[6] en el umbral de decisión por defecto de $linSVM$ ($\theta = 0$) es la más adecuada para resumir el rendimiento de cada clasificador.

Ahora bien, utilizando las normativas recuperadas del SAIJ, con respecto al clasificador de normativas del Derecho Ambiental, el conjunto total de observaciones construido consta de 1814 observaciones positivas y 1797 observaciones negativas. En la Figura 7 se muestra la matriz de confusión del clasificador aprendido, en donde se utilizó alrededor del 30% de las 3611 observaciones totales como observaciones de prueba. En dicha figura, el término DA refiere al Derecho Ambiental. De la mencionada matriz es posible estimar una exactitud $ACC = 0.91$, el cual se considera un muy buen resultado.

Finalmente, en la Tabla 1 se exhiben los resultados de validación de los restantes 15 clasificadores binarios de texto. En ésta, $\mathscr{V}$ representa al conjunto de observaciones utilizadas para validar cada modelo ($\sim |\mathscr{V}|$ representa el porcentaje aproximado utilizado para validación). En la mencionada tabla, se puede notar que las mediciones de la métrica ACC arrojan resultados alentadores, los cuales varían entre 0.82 y 0.95.

Como se destacó al principio de la presente sección, durante este trabajo, además, se aprendieron clasificadores de tópicos específicos aplicando la metodología brevemente comentada en la Sección 3.1. En particular, se seleccionaron los tópicos *desechos*

[4] `https://scrapy.org/`
[5] `https://scikit-learn.org/`
[6] Fracción de predicciones que el modelo realizó correctamente

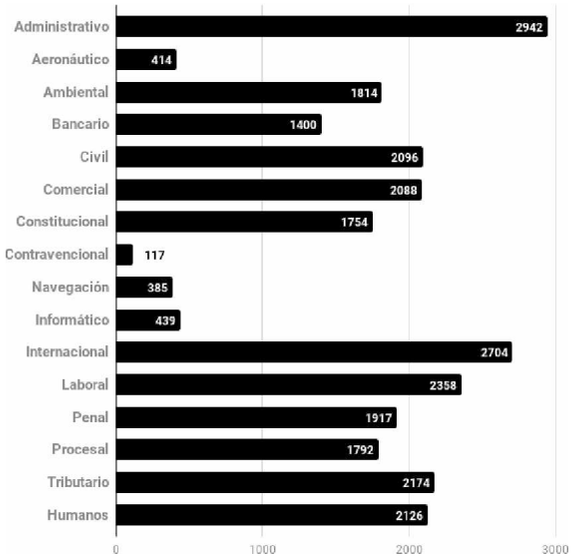

Figura 6. Número de normativas recuperadas para cada rama

		Predicho		
		$\overline{DA}$	DA	total
Real	$\overline{DA}$	494	45	539
	DA	52	493	545
	total	546	538	

Figura 7. Matriz de confusión del modelo clasificador de texto del Derecho Ambiental en el umbral de decisión $\theta = 0$

peligrosos y *accidentes de trabajo* pertenecientes al dominio del Derecho Ambiental y al Derecho Laboral, respectivamente. La validación del clasificador de *desechos peligrosos* resultó en un $ACC = 0.87$, y un $ACC = 0.88$ con respecto a la validación del tópico *accidentes de trabajo*.

4.2. Análisis del sistema de soporte

Con el objeto de evaluar el sistema de soporte propuesto para asistir a la matricería legal, se recopilaron 4766 documentos normativos del Boletín Oficial de la República Argentina para realizar una prueba offline de la aplicación. Estas normativas no etiquetadas datan desde principios del mes de enero hasta fines de mayo de 2019. A partir de ésta simulación offline, a continuación se exhibe información del resultado

Tabla 1. Resultados de validación de los restantes 15 clasificadores en ramas del Derecho

Derecho	# obs. pos.	# obs. neg.	$\sim\lvert\mathcal{V}\rvert$	ACC
Administrativo	2942	2861	30% del conj. total	0.85
Aeronáutico	414	405	20% del conj. total	0.87
Bancario	1400	1395	30% del conj. total	0.93
Civil	2096	2063	30% del conj. total	0.82
Comercial	2088	2063	30% del conj. total	0.87
Constitucional	1754	1740	30% del conj. total	0.82
Contravencional	117	105	20% del conj. total	0.87
Navegación	385	375	20% del conj. total	0.93
Informático	439	435	20% del conj. total	0.90
Internacional	2704	2637	30% del conj. total	0.95
Laboral	2358	2315	30% del conj. total	0.88
Penal	1917	1895	30% del conj. total	0.87
Procesal	1792	1783	30% del conj. total	0.92
Tributario	2174	2133	30% del conj. total	0.90
Humanos	2126	2091	30% del conj. total	0.93

de la clasificación del conjunto de normativas recuperadas en Derecho Ambiental y Derecho Laboral, e información acerca de las normativas candidatas a poblar las matrices legales de dichas ramas del Derecho de una posible industria aceitera. Los resultados se encuentran en la Tabla 2 y 3. Dado que el sistema sólo selecciona

Tabla 2. Flujo de clasificación artificial de normativas en el dominio del Derecho Ambiental

# total de normativas	# Derecho Ambiental	# desechos peligrosos
4766	185	**54**

Tabla 3. Flujo de clasificación artificial de normativas en el dominio del Derecho Laboral

# total de normativas	# Derecho Laboral	# accidentes de trabajo
4766	1049	**177**

aquellos documentos que tratan sobre términos más específicos de los dominios de las mencionadas dos ramas del Derecho, el resultado final está compuesto por 54 normativas del Derecho Ambiental y 177 del Derecho Laboral. Estos documentos fueron parcialmente evaluados por los profesionales expertos y los resultados fueron prometedores, destacando la alta reducción del número de normativas a revisar y el muy buen rendimiento, principalmente, con respecto a la detección artificial de normativas correspondientes al Derecho Ambiental y al Derecho Laboral.

5. Conclusiones y trabajo futuro

La ingeniería legal tiene la misión de crear herramientas capaces de consumir, analizar y obtener patrones de documentos legales para asistir tanto a profesionales de la Ley como a individuos en una sociedad en la toma de decisión y en la búsqueda de información legal específica. En particular, una gran cantidad de documentos es publicada diariamente de forma digital a través de distintos boletines oficiales. Las empresas en especial, necesitan del continuo análisis de estos documentos normativos con el fin de corroborar su cumplimiento acorde a las actividades que realizan y de esta manera, cumplir con todas la leyes vigentes y prevenir posibles sanciones del Estado. Esta actividad se denominó matricería legal a lo largo del presente trabajo.

En este artículo, primeramente, se propuso una arquitectura conceptual denominada SiSIL con el fin de asistir, en tiempo real, a diversas tareas relacionadas a la IL. Esta arquitectura base permite el desarrollo de nuevas y diversas aplicaciones de soporte a la toma de decisión legal. Estas aplicaciones pueden apoyar no sólo a empresas, sino también a la comunidad en general para que esta última se involucre en una Justicia más abierta, moderna e inclusiva en el marco del programa *Justicia 2020*[7] del Ministerio de Justicia y Derechos Humanos de la República Argentina.

Además, se ha propuesto el diseño e implementación de un asistente soporte a la matricería legal utilizando SiSIL como arquitectura base. En este caso, el módulo de análisis de la mencionada arquitectura, asume la responsabilidad de evaluar todo documento normativo recuperado de determinados boletines oficiales y sugerir aquellos que potencialmente se consideren relevantes con respecto a las actividades realizadas por una determinada empresa, predefinida con anterioridad. Con el propósito de obtener los requisitos de la misma, el sistema y el experto establecen un diálogo exploratorio del dominio de las principales ramas del Derecho de interés por la entidad, utilizando el Tesauro del Derecho Argentino como fuente de información externa. El análisis empleado por SiSIL a toda normativa recuperada consta de una evaluación de contenido textual mediante la aplicación en cadena de distintos clasificadores binarios de texto, los cuales intentan inferir cuándo una normativa es relevante para el contexto de una empresa. El usuario encargado de la matricería legal interactúa con los documentos sugeridos por el sistema y es quien finalmente, decide la verdadera relevancia de dichos documentos. Los documentos aceptados por el usuario son los que la aplicación termina incorporando al repositorio de las matrices legales de la empresa. Esta aplicación tiene como principal finalidad asistir y facilitar la ardua tarea de los profesionales encargados de la selección diaria de normativas exigibles a una empresa, provenientes de múltiples fuentes de información, mediante la sugerencia de normativas potencialmente relevantes. En este trabajo se ha logrado un primer prototipo de dicha aplicación con resultados preliminares promisorios, los cuales fueron evaluados por un conjunto de profesionales de la Ley y expertos en matricería legal, quienes principalmente destacaron la alta reducción de normativas a analizar diariamente.

Como trabajo futuro, se propone refinar el aprendizaje de los clasificadores de tópicos específicos, situación crítica cuando los requerimientos del experto son cada vez más estrictos. Dada las dificultades encontradas para aprender estos clasificadores

[7] `http://www.justicia2020.gob.ar`

de texto durante el presente trabajo (causada principalmente por el incremento de clases desbalanceadas debido a la escasez de normativas clasificadas en dichos tópicos por el SAIJ), se plantea explorar y utilizar otras alternativas como, por ejemplo, el reconocimiento de entidades nombradas, aprendizaje en línea utilizando el algoritmo *Passive-Aggressive*, y el sobremuestreo de texto. Estas alternativas se introducen en [14], [15] y [16], respectivamente.

Agradecimientos Los autores agradecen el apoyo recibido en la evaluación de los resultados del presente trabajo a voluntarios profesionales de la Ley, y a *SmartLegal* en el marco de un Servicio Tecnológico de Alto Nivel (STAN) brindado por el CIFASIS (CONICET-UNR).

Referencias

1. Olszak, C.M.: Toward better understanding and use of business intelligence in organizations. Information Systems Management **33**(2), 105–123 (2016)
2. Baggio, B.G.: Analyzing Digital Discourse and Human Behavior in Modern Virtual Environments. IGI Global (2016)
3. Shimazu, A., Le Nguyen, M.: Legal Engineering and Its Natural Language Processing. In: Knowledge and Systems Engineering, pp. 7–7. Springer (2014)
4. Lippi, M., Pałka, P., Contissa, G., Lagioia, F., Micklitz, H.W., Sartor, G., Torroni, P.: Claudette: an automated detector of potentially unfair clauses in online terms of service. Artificial Intelligence and Law **27**(2), 117–139 (2019)
5. Yamada, H., Teufel, S., Tokunaga, T.: Building a corpus of legal argumentation in japanese judgement documents: towards structure-based summarisation. Artificial Intelligence and Law **27**(2), 141–170 (2019)
6. Neil, M., Fenton, N., Lagnado, D., Gill, R.D.: Modelling competing legal arguments using bayesian model comparison and averaging. Artificial Intelligence and Law pp. 1–28 (2018)
7. Cardellino, F., Cardellino, C., Haag, K., Soto, A., Teruel, M., Alonso Alemany, L., Villata, S.: Mejora del acceso a infoleg mediante técnicas de procesamiento automático del lenguaje. In: XVIII Simposio Argentino de Informática y Derecho (2018)
8. Nath, J., Sanjay, S., Dwivedi, K.: A comparative study on approaches of vector space model in information retrieval. International Journal of Computer Applications **975**, 8887 (2013)
9. Manning, C.D., Raghavan, P., Schutze, H.: Introduction to Information Retrieval. Cambridge University Press, New York, NY, USA (2008)
10. Sebastiani, F.: Machine learning in automated text categorization. ACM computing surveys (CSUR) **34**(1), 1–47 (2002)
11. Joachims, T.: Learning to classify text using support vector machines. Springer Science & Business Media, Boston, MA (2002)
12. Aggarwal, C.C.: Machine learning for text. Springer (2018)
13. Borovicka, T., Jirina, M., Kordik, P., Jiri, M.: Selecting Representative Data Sets. In: Advances in Data Mining Knowledge Discovery and Applications. InTech (2012)
14. Nadeau, D., Sekine, S.: A survey of named entity recognition and classification. Lingvisticae Investigationes **30**(1), 3–26 (2007)
15. Crammer, K., Dekel, O., Keshet, J., Shalev-Shwartz, S., Singer, Y.: Online passive-aggressive algorithms. Journal of Machine Learning Research **7**, 551–585 (2006)
16. Iglesias, E.L., Vieira, A.S., Diz, M.L.B.: An hmm-based over-sampling technique to improve text classification. Expert Syst. Appl. **40**, 7184–7192 (2013)