

Sistemas de Apoyo a la Ingeniería Legal

Tesina de grado

Luciano Francisco Perezzini



Lic. en Cs. de la Computación

Octubre de 2019

Directora: Dra. Ana Casali

Co-directora: Dra. Claudia Deco

Contenido ☰

1. Introducción

La recuperación de información

El problema: la matricería legal

2. Sistemas de información de texto

Preprocesamiento y representación de texto

Descubrimiento de información

3. Propuesta de asistente a la matricería legal

Sistema de soporte a la ingeniería legal (SiSIL)

Análisis: clasificación de normativas

La matricería legal como aplicación de usuario

4. Experimentación

5. Conclusiones y trabajo futuro

1. Introducción

La recuperación de información (IR)

- ▶ Área sumamente **importante** en la **Era de la Información**

La IR es la **tarea de encontrar**, dentro de grandes conjuntos, **material** que **satisfaga** (potencialmente) una **necesidad de información**



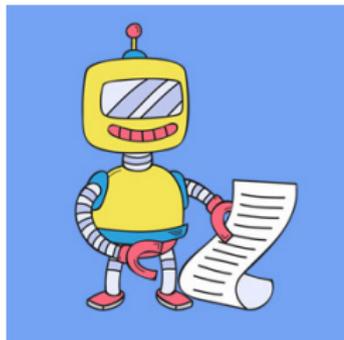
La información de texto

Nuestra forma preferida para expresar información

- ▶ **Rico** en contenido **semántico** (información valiosa: opiniones y preferencias)
- ▶ Por lo general, **no estructurado** (*texto libre*)
- ▶ **Protagonista** principal de la **web** actual

Necesidad de **agentes artificiales** para:

- ▶ **Procesar** y
- ▶ **descubrir información relevante**

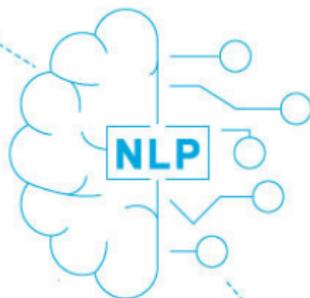


La información de texto

Procesamiento del lenguaje natural (NLP)

- ▶ Área de estudio dentro de la **lingüística** y las **ciencias de la computación**

Recuperación de
información



Clasificación de
texto

El problema – Introducción

Boletines oficiales

Documentos normativos:

- ▶ Dictan el **comportamiento** que toda **persona u organización** debe atender para el **ordenamiento** de una comunidad
- ▶ **Publicados electrónicamente** mediante distintos **boletines oficiales** (nacional, provinciales, municipales)

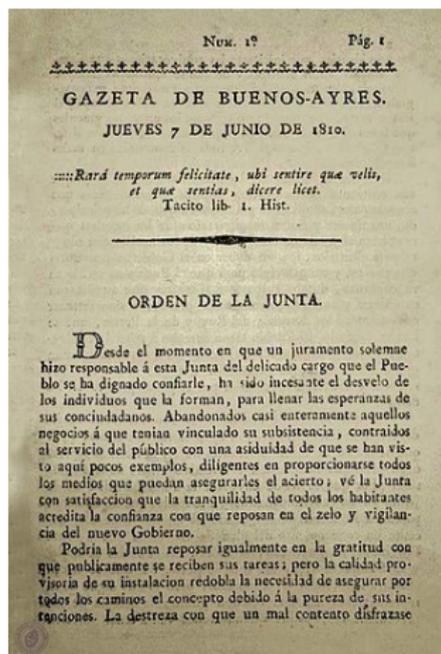


Figura: Gazeta de Buenos Ayres (Primera Junta de Gobierno – 1810)

El problema – Introducción

Boletines oficiales

The screenshot displays the BORA website interface. At the top left is the logo of the Boletín Oficial de la República Argentina. The top navigation bar includes links for INSTITUCIONAL, PRODUCTOS Y SERVICIOS, PREGUNTAS FRECUENTES, and CONTACTO. The main content area is titled 'Legislación y Avisos Oficiales' and is divided into sections: 'Primera sección' (highlighted in blue), 'DECISIONES ADMINISTRATIVAS', 'MINISTERIO DE SEGURIDAD' (with a link to 'Decisión Administrativa 771/2019'), 'SUPERINTENDENCIA DE SERVICIOS DE SALUD' (with a link to 'Decisión Administrativa 772/2019'), 'RESOLUCIONES', 'JEFATURA DE GABINETE DE MINISTROS - SECRETARÍA DE EMPLEO PÚBLICO' (with a link to 'Resolución 294/2019'), and another 'JEFATURA DE GABINETE DE MINISTROS - SECRETARÍA DE EMPLEO PÚBLICO' (with a link to 'Resolución 295/2019'). On the right side, there is a search bar for the current date (September 11, 2019), buttons for 'BÚSQUEDA AVANZADA', 'HISTORIAL DE SOCIEDADES COMERCIALES', and 'PUBLICAR AVISOS'. Below these is a calendar for September 2019, with the 11th highlighted. A 'Biblioteca de Normativas' link is also present.

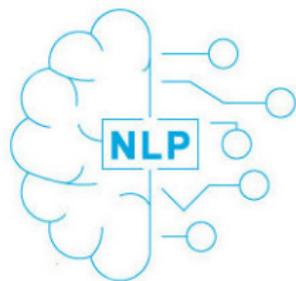
Figura: Portal web del Boletín Oficial de la República Argentina (BORA). El BORA publica alrededor de 60 normativas diarias.

El problema – Introducción

Acerca de la ingeniería legal

Ingeniería legal

Rama de las ingenierías que aplica **ciencias de la información a documentos legales** con la finalidad de **asistir en tareas de toma de decisión legal**



- ▶ Técnicas de NLP: **atractivas** de aplicar para **(semi) automatizar** tareas de la ingeniería legal

El problema: la matricería legal

Matricería legal (ML)

Actividad que se encarga de la **compilación** de **normativas exigibles** a una **entidad** acorde a su **actividad productiva**

- ▶ **⚠ Diaria.** Normativas **nuevas** y potencialmente relevantes se publican **diariamente**.
- ▶ **⚠ Trabajosa.** Se debe **analizar** cada normativa para evaluar su **relevancia** con respecto a la **actividad productiva** de la empresa.
- ▶ **⚠ Cautelosa.** Un **error** puede causar desde **penalizaciones** del Estado hasta **malas decisiones** industriales.

También conocido como **Tecnología Regulatoria** (*RegTech*)

El problema: la matricería legal

- ▶ Carecer de una matriz legal puede ocasionar problemas...

ANMAT prohibió comercializar una serie de alfajores, galletitas, budines y turrónes Nevares



- ▶ No cumplía con una normativa publicada en el **año 2010**

Figura: Noticia 27/09/19

El problema: la matricería legal

Poblado diario de matriz legal en la actualidad

1. Se **recuperan** todas las normativas del **día de la fecha**
2. Se **escogen** las pertenecientes a determinadas **ramas** del Derecho **de interés**.
3. **Expertos** en **distintas áreas** de la empresa realizan una **evaluación más refinada**.
4. Las normativas seleccionadas son **almacenadas** en la **matriz legal**.

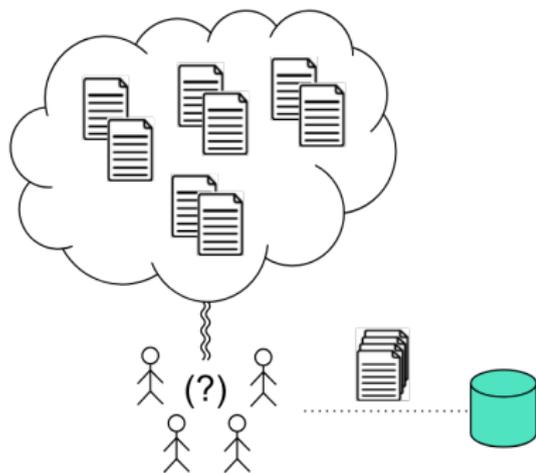


Figura: La ML en la actualidad.

Problema:

↑↑ **magnitud** de la empresa \implies ↑↑ **costo** de la ML

Propuesta: semi-automatización de la ML

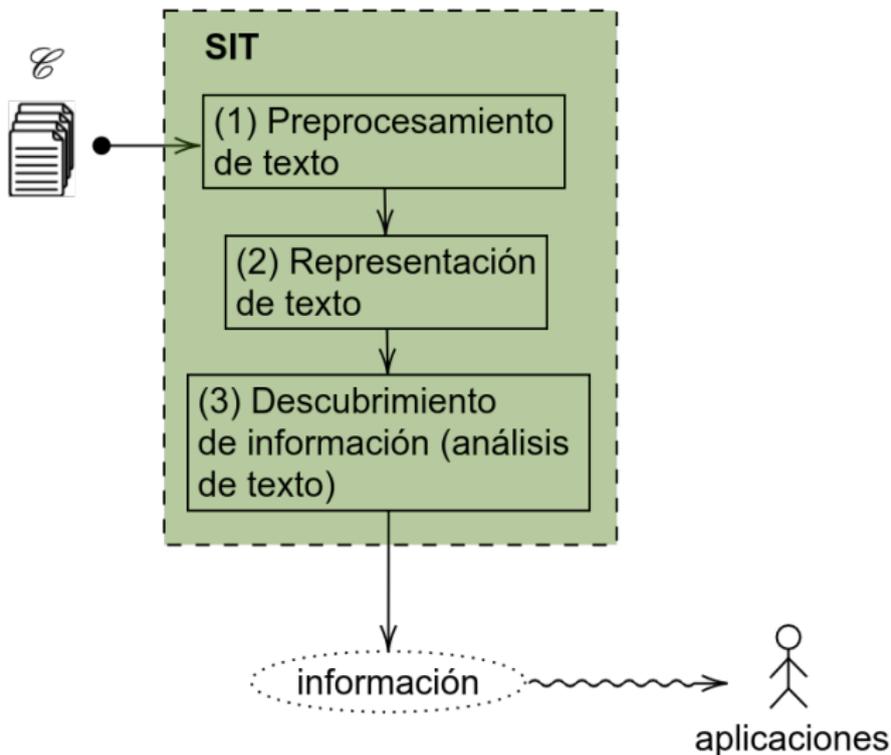
- ▶ Sistema **en línea** de **recuperación** y **sugerencia** de normativas **potencialmente relevantes** para el **poblado continuo** de una **matriz legal**
 - ▶ **Revisión** experta **final**

2. Sistemas de información de texto

Procesar y analizar grandes conjuntos de datos de lenguaje natural

Sistemas de información de texto (SITs)

3 etapas principales



1. Preprocesamiento de texto

Objetivo: determinar el **vocabulario de términos** (palabras) de la **colección**

- ▶ Composición de operaciones lingüísticas

1. **Tokenización** (*tok*). Cortar cadenas de caracteres en pedazos, llamados *tokens* o términos.
2. **Normalización** (*norm*). Transformación de cada término a una forma canónica
3. **Supresión de palabras vacías** (*rem*). Remoción de términos extremadamente frecuentes del lenguaje.
4. **Stemming** (*stem*). Reducción de términos a una forma base.

$$\text{preproc} = \text{stem} \circ \text{rem} \circ \text{norm} \circ \text{tok}$$

1. Preprocesamiento de texto

Ejemplo

d = «A veces sentimos que lo que estamos haciendo es sólo una gota en el océano. Pero si esa gota no estuviera en el océano, el océano sería menos por no tenerla»

preproc(d) = ['sent', 'got', 'ocean', 'got', 'ocean', 'ocean', 'ten']

$$preproc(d) \forall d \in \mathcal{C} \xrightarrow{\text{crearVocab()}} V = \{t_1, t_2, \dots, t_n\}$$

2. Representación de texto

El Modelo Espacio-Vectorial (VSM)

- ▶ $d \mapsto \vec{v}_d = (w_{t_1,d}, w_{t_2,d}, \dots, w_{t_{|V|},d}) \in \mathbb{X}^{|V|}$, donde:

$$w_{t,d} = w_{t,d}^{local} \times w_t^{global}$$

- ▶ Esquema de pesaje **time frequency – inverse document frequency (tf.idf)**:

$$w_{t,d}^{local} = \mathbf{tf}(t, d) = \mathit{num}(t, d)$$

$$w_t^{global} = \mathbf{idf}(t) = \log \frac{|\mathcal{C}|}{|\{d \in \mathcal{C} \mid t \in d\}|}$$

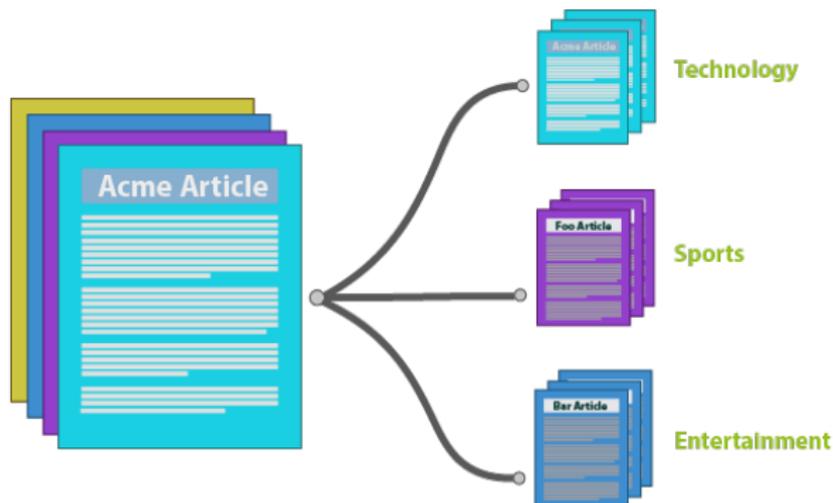
$$\therefore \vec{v}_d = (\mathit{tf.idf}_{t_1,d}, \mathit{tf.idf}_{t_2,d}, \dots, \mathit{tf.idf}_{t_{|V|},d}) \in \mathbb{R}_0^{|V|}$$

- ▶ \vec{v}_d es de **alta dimensionalidad y ralo**

3. Descubrimiento de información

El problema de la clasificación de texto

Actividad de **etiquetar** documentos con **clases** temáticas pertenecientes a un conjunto **predefinido**



3. Descubrimiento de información

El problema de la clasificación de texto

Naturaleza de alta dimensionalidad del texto

\Rightarrow $\uparrow\uparrow$ probabilidad de separar linealmente ambas clases \Rightarrow **clases cuasi-linealmente separables**

\therefore Aprender frontera de decisión lineal

- ▶ **Clasificador de vectores soporte (SVC)**
 - ▶ Gran desempeño en la clasificación de texto

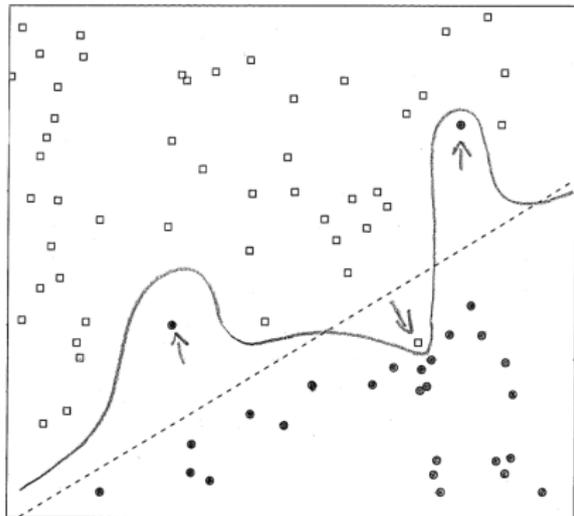
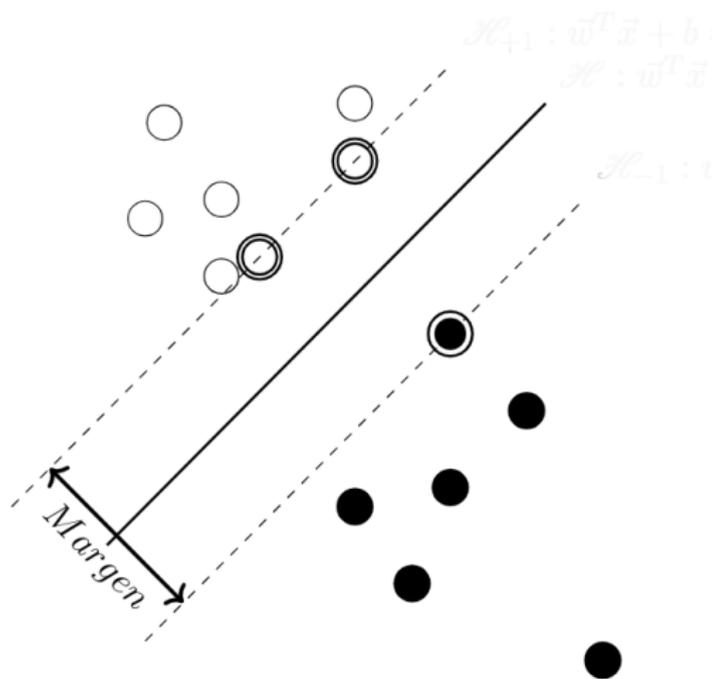


Figura: Los modelos más flexibles no suelen brindar mejores resultados en la clasificación de texto.

3. Descubrimiento de información

Clasificación lineal: clasificador de vectores soporte (SVC)

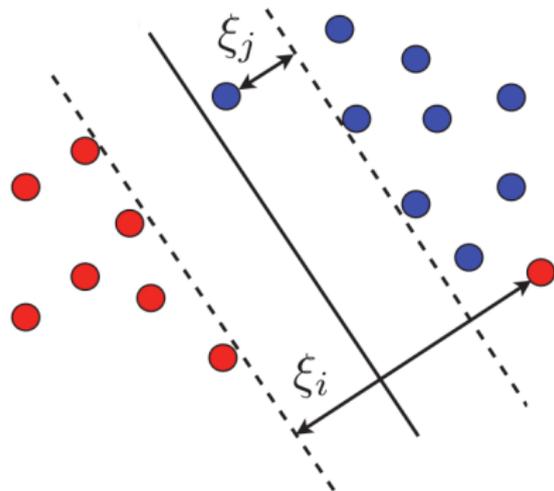
- ▶ Hiperplano **separador de margen máximo**



3. Descubrimiento de información

Clasificación lineal: clasificador de vectores soporte (SVC)

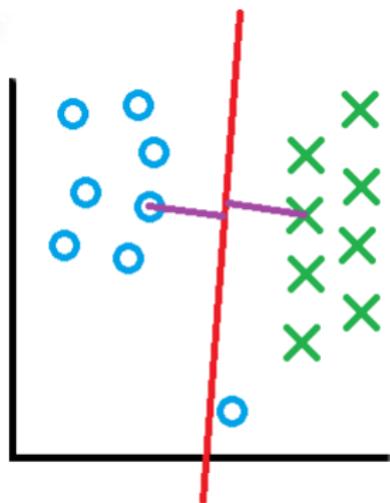
- ▶ Se permiten **violaciones** del margen (ξ_i)
 - ▶ **penalizadas** por un hiperparámetro C
- ▶ C controla el **tradeoff** entre la **maximización del margen** y la **minimización del error** ($\sum_i \xi_i$)



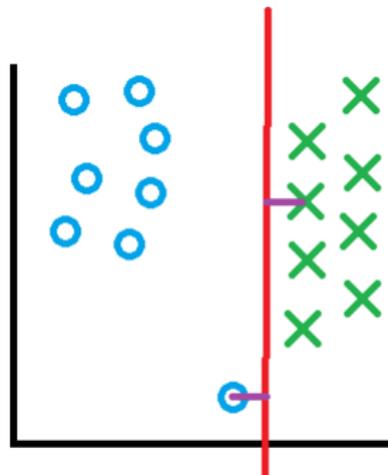
3. Descubrimiento de información

Clasificación lineal: clasificador de vectores soporte (SVC)

- ▶ Hiperparámetro C
 - ▶ Valor típicamente configurado vía validación cruzada
 - ▶ Controlar overfitting



low c



large c

3. Descubrimiento de información

Clasificación binaria de texto: aprendizaje y predicción

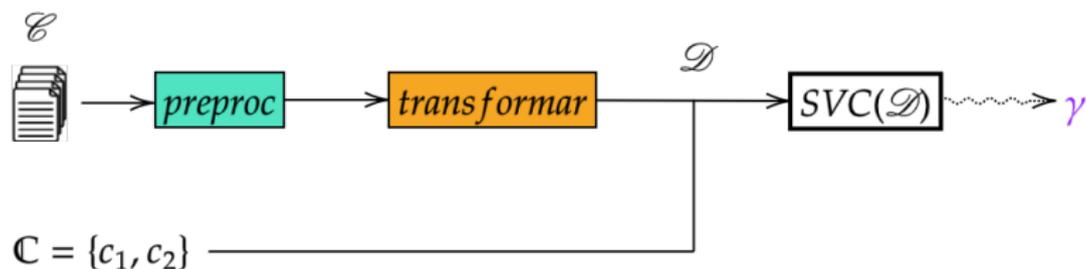


Figura: Aprendizaje de modelo clasificador binario de texto γ .

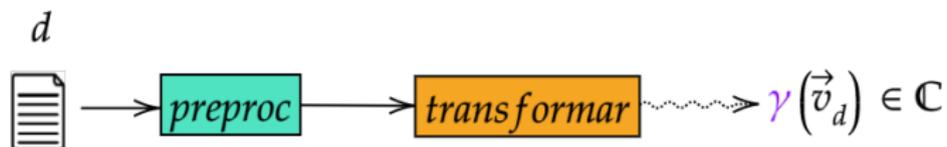


Figura: Predicción de clase del documento d .

3. Descubrimiento de información

Clasificación de texto multi-valor

- ▶ Por lo general, $|\mathbb{C}| > 2$ y sus clases no son mutuamente excluyentes

Problema de **clasificación multi-valor**:

- ▶ d puede pertenecer a varias clases, a una sola, o a ninguna

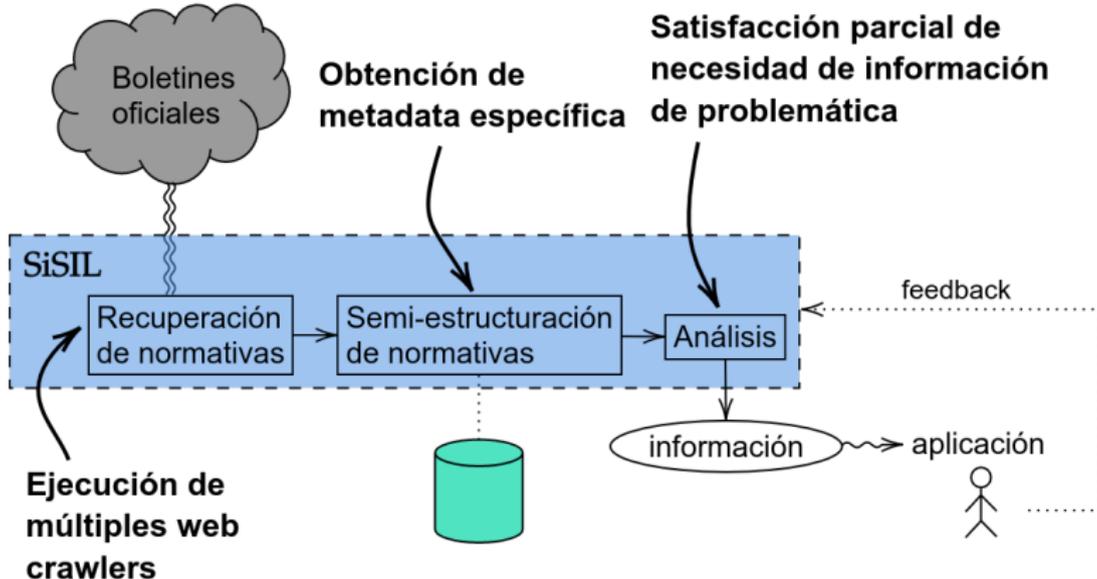
Solución:

- ▶ **Aprender** $J = |\mathbb{C}|$ **clasificadores binarios de texto**
 $\gamma_j \mid \gamma_j(d) \in \{c_j, \bar{c}_j\}$
- ▶ Dado un documento de prueba, se **aplica** cada γ_j de **forma separada**

3. Propuesta de asistente a la matricería legal

Sistema de soporte a la ingeniería legal (SiSIL)

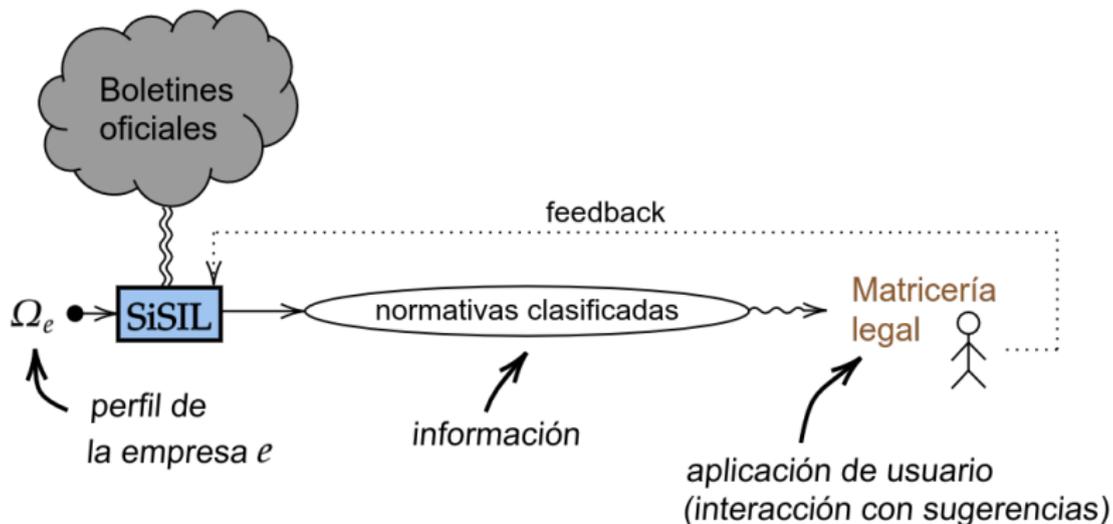
Arquitectura de asistente en línea para tareas de la ingeniería legal



Problemática: matricería legal en una empresa e

- ▶ I = «Obtener, de determinados boletines, normativas relevantes a la **actividad productiva** de e»

Tarea de SiSIL: satisfacer parcialmente I mediante la sugerencia de normativas



Actividad productiva

Recopilación de información: exploración del Derecho

- ▶ **Dialogo** entre el sistema y el usuario experto
 - ▶ El sistema asiste al usuario mediante una fuente de conocimiento externa: el **Tesoro del Derecho Argentino (TDA)**

Ramas del Derecho

<input checked="" type="checkbox"/>	Administrativo
<input type="checkbox"/>	Aeronáutico
<input checked="" type="checkbox"/>	Ambiental
⋮	
<input type="checkbox"/>	Humanos

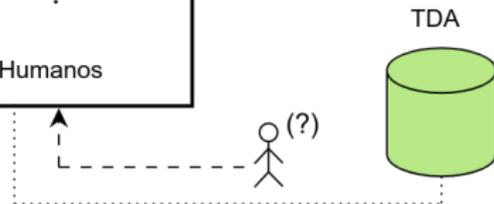


Figura: 1er. etapa del dialogo.
Selección de principales ramas.

Actividad productiva

Recopilación de información: exploración del Derecho

- ▶ El experto escoge **tópicos** más **específicos**

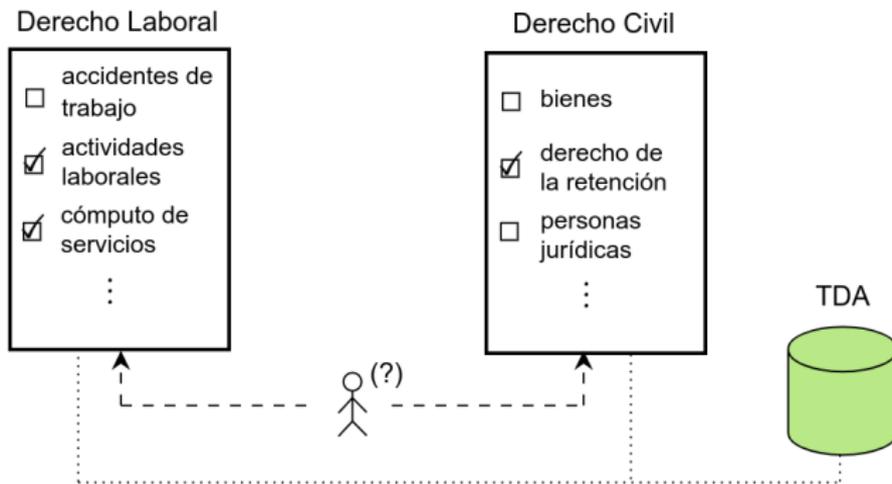
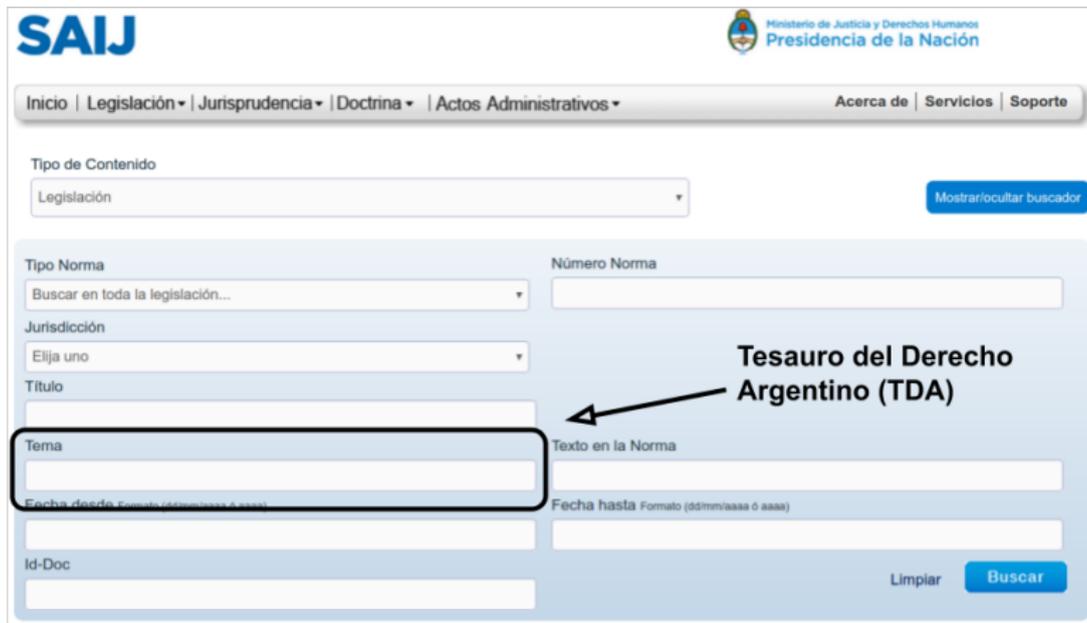


Figura: 2da. etapa del diálogo. Exploración de ramas seleccionadas.

Antes de continuar: ¡necesitamos normativas etiquetadas!



SAIJ Ministerio de Justicia y Derechos Humanos
Presidencia de la Nación

Inicio | Legislación ▾ | Jurisprudencia ▾ | Doctrina ▾ | Actos Administrativos ▾ Acerca de | Servicios | Soporte

Tipo de Contenido
Legislación Mostrariocultar buscador

Tipo Norma: Buscar en toda la legislación...
Número Norma:

Jurisdicción: Elija uno ▾

Título:

Tema: ← Tesoro del Derecho Argentino (TDA)

Texto en la Norma:

Fecha desde: Consulte (dd/mm/aaaa ó yyyy) Fecha hasta: Formato (dd/mm/aaaa ó yyyy)

Id-Doc: Limpiar

Figura: Sistema Argentino de Información Jurídica (SAIJ).

Antes de continuar: ¡necesitamos normativas etiquetadas!

Poder Ejecutivo Provincial: Ordenamiento de Plantas de Incineración de Residuos Peligrosos.

LEY 1.005

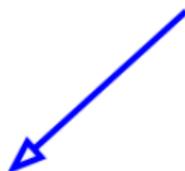
USHUAIA, 21 de Agosto de 2014

Boletín Oficial, 22 de Diciembre de 2014

Vigente, de alcance general

Id SAIJ: LPV0001501

“Residuos peligrosos”:
tópico específico del
Derecho Ambiental

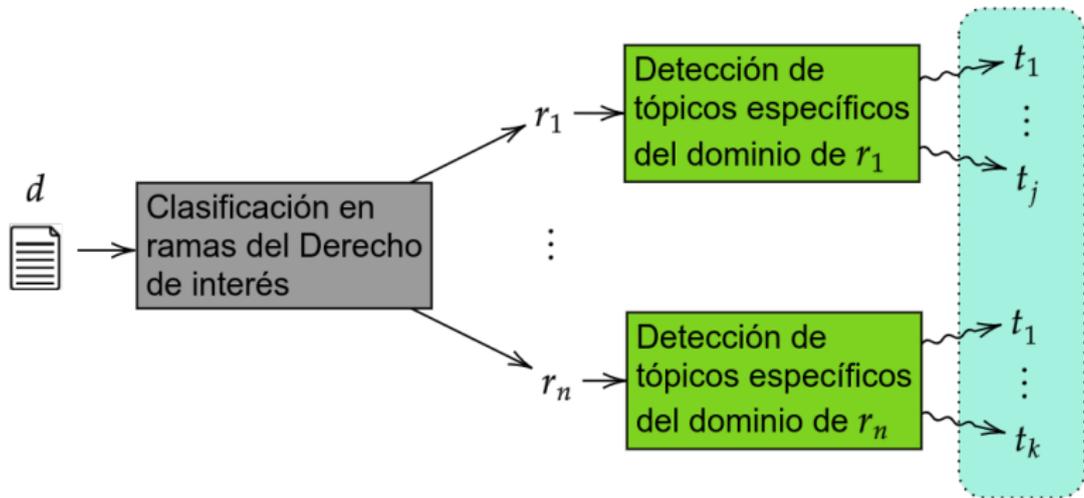


SUMARIO

Poder Ejecutivo Provincial, **residuos peligrosos** ley provincial, Secretaría de Desarrollo Sustentable y Política Ambiental, **Derecho constitucional** **Derecho ambiental**

Análisis de SiSIL: clasificación de normativas

Flujo propuesto de clasificación: dos etapas

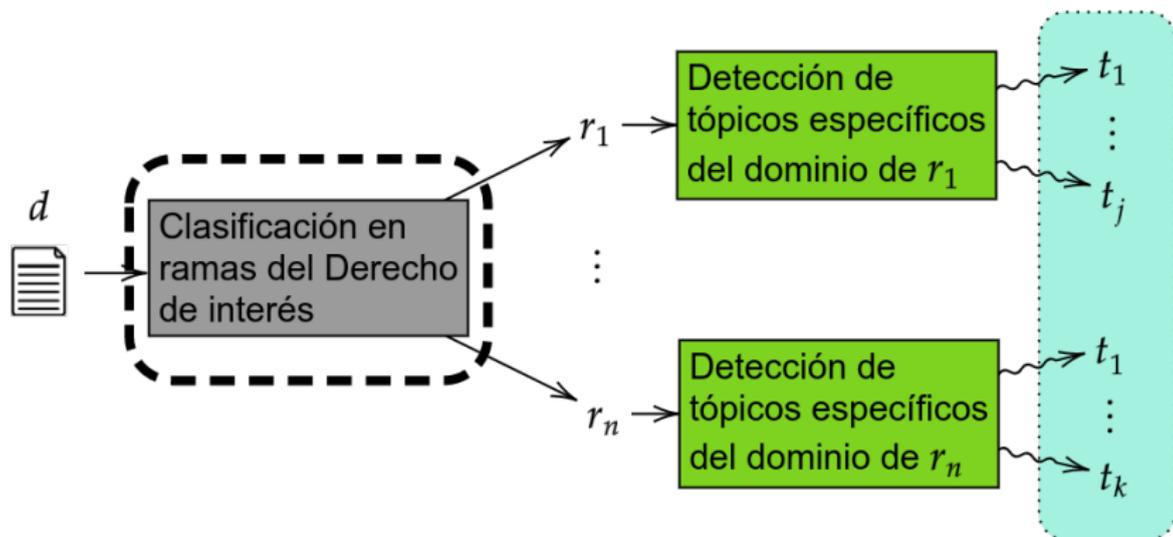


d es considerada **potencialmente relevante** si:

- ▶ es clasificada como **perteneciente a alguna rama de interés r** y además
 - ▶ trata sobre algún **tópico específico del dominio** de r

Análisis de SiSIL: clasificación de normativas

- ▶ **1er. etapa:** clasificación en ramas del Derecho



Análisis de SiSIL: clasificación de normativas

1er. etapa: clasificación en ramas del Derecho

- ▶ \mathbb{D} : conj. de las **principales ramas del Derecho** (TDA: $|\mathbb{D}| = 16$)
- ▶ Clasificación **multi-valor**
- ▶ Problema de **clases desbalanceadas**
 - ▶ **Submuestreo aleatorio**
 - ▶ Construcción de conj. de obs. aproximadamente balanceado

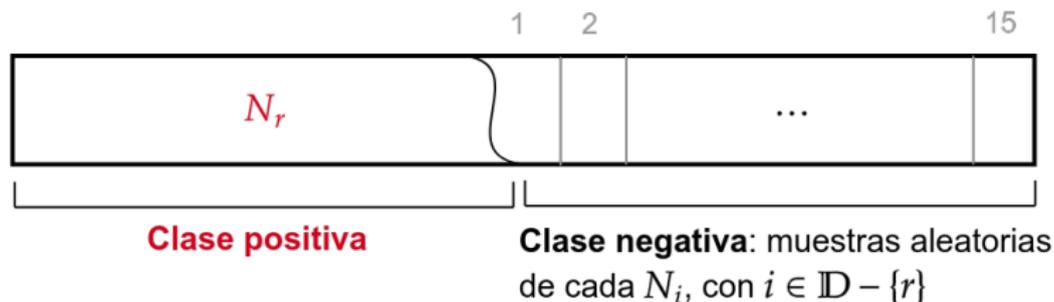
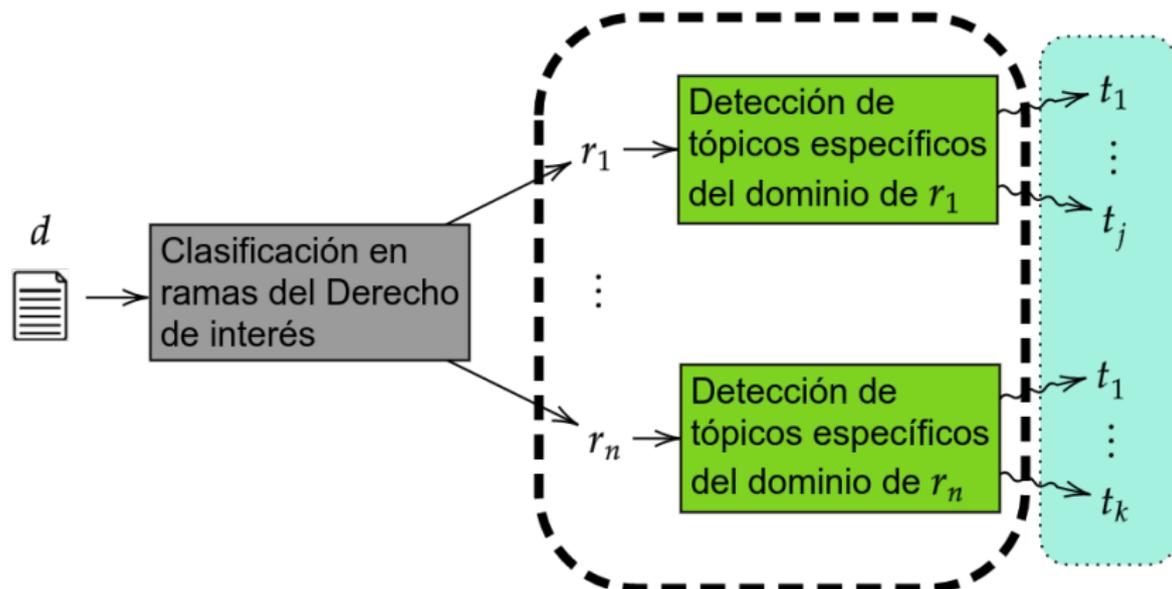


Figura: Construcción de conj. de observaciones. N_r : conj. de las normativas recuperadas con clase $r \in \mathbb{D}$.

Análisis de SiSIL: clasificación de normativas

- **2da. etapa:** detección de tópicos específicos de ramas predichas



Análisis de SiSIL: clasificación de normativas

2da. etapa: detección de tópicos específicos

- ▶ **Evaluación** de contenido textual **refinada**
 - ▶ Se aprende un **clasificador** binario de texto **por cada tópico**
- ▶ Problema de **clases desbalanceadas**
 - ▶ **Submuestreo aleatorio**
 - ▶ Construcción de conjunto de obs. balanceado

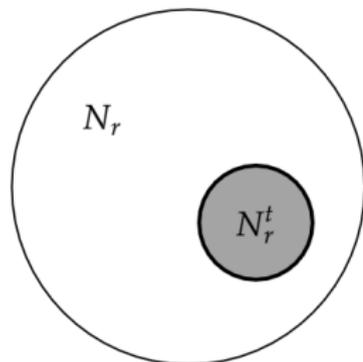


Figura: N_r^t : conj. de normativas con clase $r \in \mathbb{D}$ etiquetadas con algún $t \in \text{dom}(r)$.

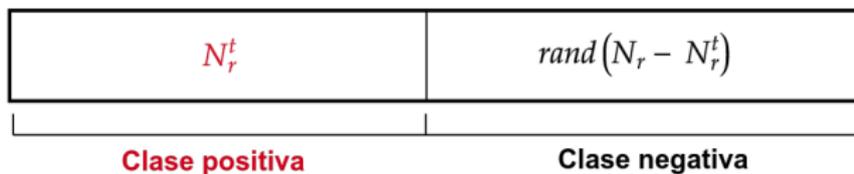


Figura: Construcción de conj. de obs.

Aplicación de usuario

- ▶ SiSIL **sugiere** al experto aquellas **normativas** recuperadas y clasificadas como **potencialmente relevantes** con respecto a la **actividad productiva** de e

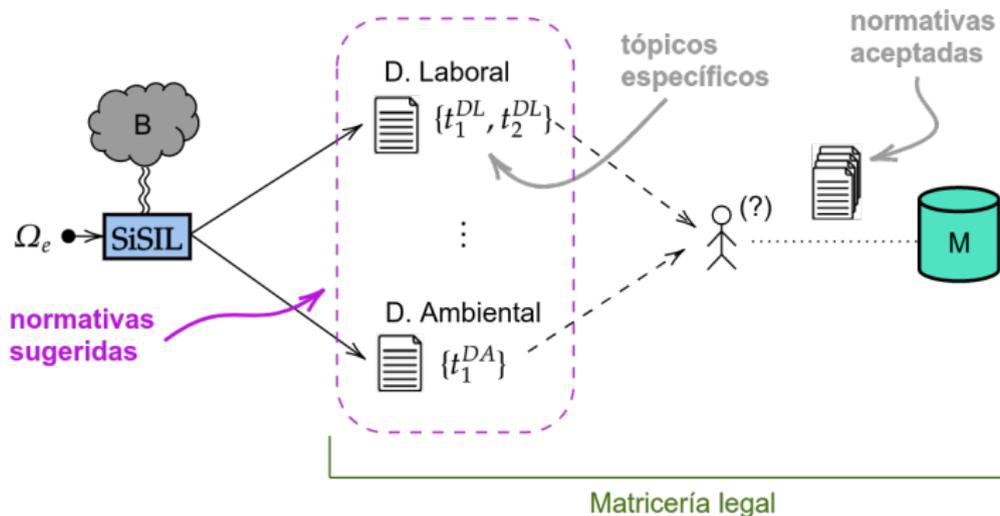


Figura: Interacción de usuario experto con normativas sugeridas. Experto descarta o acepta.

4. Experimentación

Experimentación

Caso de estudio

Industria aceitera localizada en la ciudad de San Lorenzo, Santa Fe

- ▶ **Ramas del Derecho de interés:**
 - ▶ Derecho Ambiental → desechos peligrosos
 - ▶ Derecho Laboral → accidentes de trabajo
- ▶ **Boletines oficiales** (implementación de **web crawlers**):
 - ▶ Boletín Oficial de la República Argentina
 - ▶ Boletín Oficial de la Provincia de Santa Fe
 - ▶ Boletín Oficial de la Ciudad de San Lorenzo, Santa Fe

Normativas recuperadas del SAIJ

Recuperación mediante crawler

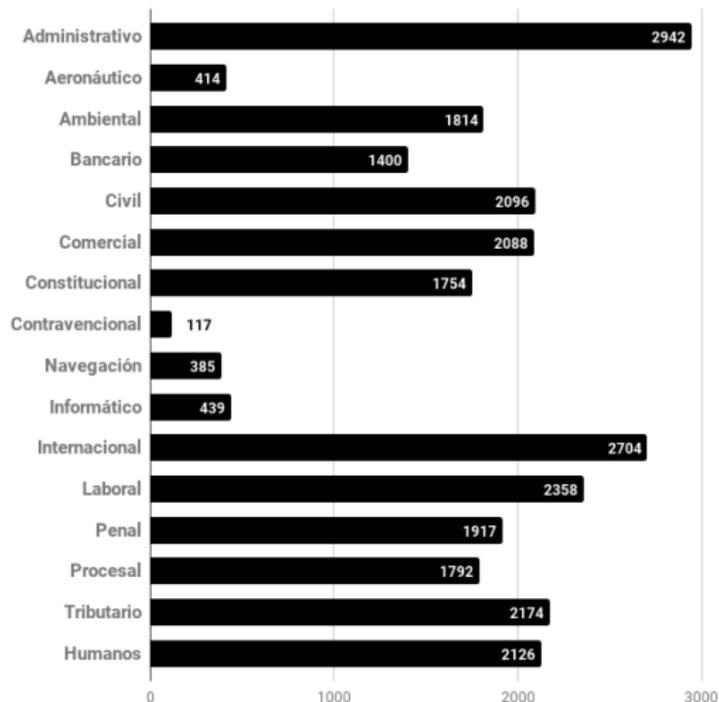


Figura: Conjuntos de normativas recuperadas. 26520 normativas totales.

Experimentación

Aprendizaje y validación de clasificadores

Metodología aplicada:

1. Construcción de un conjunto total de observaciones \mathcal{D}
2. Partición aleatoria y estratificada de \mathcal{D} en conjuntos de observaciones de entrenamiento (\mathcal{T}) y validación (\mathcal{V})
3. Ajuste del hiperparámetro C aplicando validación cruzada en el conjunto \mathcal{T}
4. Aprendizaje de modelo clasificador binario de texto:
 $SVC(\mathcal{T}) = \gamma$
5. Validación de γ mediante las observaciones del conjunto \mathcal{V}

Dominio: Derecho Ambiental

Clasificador del **Derecho Ambiental (DA)**

	# pos.	# neg.	# total
\mathcal{D}	1814	1797	3611

- ▶ $|\mathcal{V}| \sim 30\%$ de obs. totales
- ▶ Se estima **ACCURACY = 0.91**

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

Figura: Matriz de confusión.

Dominio: Derecho Ambiental

Clasificador de **desechos peligrosos (dp)**

	# pos.	# neg.	# total
\mathcal{D}	241	241	482

- ▶ $|\mathcal{Y}| \sim 25\%$ de obs. totales
- ▶ Se estima **ACCURACY = 0.87**

		Predicho		total
		\bar{dp}	dp	
Real	\bar{dp}	54	6	60
	dp	10	50	60
total		64	56	

Figura: Matriz de confusión.

Dominio: Derecho Laboral

Clasificador del **Derecho Laboral (DL)**

	# pos.	# neg.	# total
\mathcal{D}	2358	2315	4673

- ▶ $|\mathcal{V}| \sim 30\%$ de obs. totales
- ▶ Se estima **ACCURACY = 0.88**

		Predicho		total
		\overline{DL}	DL	
Real	\overline{DL}	619	76	695
	DL	94	613	707
total		713	689	

Figura: Matriz de confusión.

Dominio: Derecho Laboral

Clasificador de **accidentes de trabajo** (*at*)

	# pos.	# neg.	# total
\mathcal{D}	228	228	456

- ▶ $|\mathcal{Y}| = 25\%$ de obs. totales
- ▶ Se estima **ACCURACY** = **0.88**

		Predicho		total
		\bar{at}	<i>at</i>	
Real	\bar{at}	48	9	57
	<i>at</i>	5	52	57
total		53	61	

Figura: Matriz de confusión.

5. Conclusiones y trabajo futuro

Conclusiones

- ▶ SiSIL + matricería legal --> **aproximación de semi-automatización del monitoreo regulatorio** en empresas
 - ▶ ¡**Resultados** preliminares **alentadores!**
- ▶ SiSIL: **arquitectura conceptual de soporte** para tareas de la ingeniería legal
 - ▶ Posibilidad de incluir **distintas formas de análisis** de normativas \implies nuevas aplicaciones $\square \implies$ **Justicia más abierta, inclusiva y moderna**
 - ▶ ¡Tecnólogos y profesionales de la Ley invitados! 

Trabajo futuro

Otras técnicas a aplicar

- ▶ Clasificación de texto **basado en reglas**
 - ▶ Construcción de conjunto de reglas
$$\mathcal{R}_i : (t_a \in d) \wedge (t_b \in d) \wedge (\dots) \implies c$$
- ▶ Filtrado por **organismos del Estado** (Admin. Pública Nacional)
 - ▶ \sim 200 organismos
 - ▶ Detección de normativas emitidas por organismos relevantes
- ▶ **Aprendizaje en línea** (ejemplo: algoritmo de aprendizaje *Passive-Agressive*)

¡Gracias!