Un algoritmo para mejorar la defensa jurídica del Estado:

Tutelas contra la Agencia Nacional de Infraestructura con extracción de texto.

1 ABSTRACT

En medio de la "Cuarta Revolución Industrial", donde la tecnología está transformando los "trabajos mecánicos" y también los que se consideraban calificados, surge la pregunta de ¿Cómo es que este cambio tecnológico impacta el ejercicio del derecho?. En este trabajo se presenta un ejemplo de cómo la implementación de los algoritmos más recientes de aprendizaje de máquinas para el procesamiento de texto puede hacer más eficiente la labor de defensa jurídica del Estado. Estas herramientas reemplazan horas de trabajo humano en tareas como la realización de resúmenes de los textos jurídicos[1]y la extracción de temas y otros elementos básicos como fecha, demandantes y términos, entre otros. A partir de las tutelas interpuestas contra la Agencia Nacional de Infraestructura (ANI) entre el 2019 y el 2021 se realizan dos tipos de análisis estadístico (frecuencial y bayesiano) para proponer elementos de una estrategia de defensa jurídica que haga más eficiente la labor de la entidad. En primer lugar, se presentan los resultados de un análisis de estadísticas descriptivas para focalizar y coordinar el trabajo de defensa jurídica en la Agencia Nacional de Infraestructura. En segundo lugar, se presentan los resultados de un algoritmo de LDA para clasificar los textos por sus temas y así coordinar mejor al equipo de defensa por sus especialidades. Este segundo resultado muestra inmediatamente qué temas pertenecen a cada tutela y se puede usar directamente como insumo para planear la defensa de la entidad en cada caso.

2 INTRODUCCIÓN:

La "Cuarta Revolución Industrial"[2] está cambiando radicalmente nuestra forma de vida. La adopción masiva de nuevas tecnologías va a transformar la forma en la que trabajamos (y en general la manera en la que nos relacionamos). Con cada revolución en los modos de producción hay unos empleos que se van reemplazando poco a poco por alternativas más eficientes. Hoy nos encontramos frente a un movimiento casi sin precedentes, porque esta nueva etapa llega para automatizar no solo los trabajos "no calificados" sino también los que se consideraban especializados e irremplazables por las máquinas. En este caso me refiero a las decisiones de inversión para personas

[1] Por ejemplo, se puede automatizar la realización de fichas jurisprudenciales en sus elementos más básicos.
[2] Este concepto lo han definido varios autores, se recomiendan Perasso (2016) y Schwab (2017).

naturales y para fondos de inversión[3], la escritura de artículos y novelas[4], incluso la lectura y escritura de textos jurídicos[5]. Esto cambia por completo las reglas de juego en el mercado laboral de profesiones como el derecho[6]. Desafortunadamente, parece que las Universidades no están reaccionando con suficiente rapidez a este cambio en la demanda[7]. Los estudiantes no se están preparando para trabajos mucho más tecnificados, donde su capacidad para usar las herramientas tecnológicas a su alcance son un requisito esencial para la integración en el mercado laboral.

Por mi trabajo en la Agencia Nacional de Infraestructura[8](ANI) escogí a esta entidad como unidad de análisis. La ANI aportó los textos de las tutelas que han recibido desde 2019. Sobre estos documentos se realizó un ejercicio de digitalización, extracción y procesamiento de texto para definir cuáles son los temas más importantes en cada tutela y conocer la incidencia de esos temas en el tiempo. De ese análisis se llega a unas conclusiones útiles para definir la estrategia de defensa jurídica de la entidad y para el ejercicio de la defensa en sí mismo.

Los resultados de este trabajo permiten diseñar estrategias a nivel macro y micro para la atención de tutelas en la entidad. Las estrategias en sentido macro son las que permiten a los directivos de la entidad tomar las decisiones que hacen más eficiente el trabajo del equipo. Por ejemplo, se observa un aumento significativo entre tutelas por consulta previa y mínimo vital cuando se inicia la construcción de una vía que tiene paso por zonas de poblaciones étnicas. Entonces, la dirección de defensa jurídica puede prepararse con más tiempo[9] para contestar las tutelas que llegan por este concepto cuando se sabe que está a punto de empezar esta fase (construcción) de alguno de los proyectos concesionados.

Las ventajas a nivel micro se ven en el día a día de los funcionarios directamente encargados de responder a las tutelas. En este caso, la implementación de LDA le ofrece al funcionario una herramienta útil para procesar de manera eficiente y clara las respuestas. Hoy en día, una buena parte de las tutelas y derechos de petición que reciben las entidades se contestan con formatos. El valor agregado del uso de este algoritmo está en que ayuda a construir mejores formatos y a escoger cuál utilizar de manera prácticamente instantánea.

[3] Los roboadvisors automatizan la toma de decisiones de inversión en el mercado bursátil, de futuros, de divisas y criptomonedas. Llegaron para quedarse. (Puschmann, 2017).

[4] Hay multiples softwares que cumplen este propósito, Sevilla, A. (2021) hace una recopilación de los más populares en 2021.

[5] Ashley (2017) lista este algoritmo como uno de los avances que eran impensables hace unos años y hoy son una realidad.

[6] Ashley (2017) introduce su libro hablando de cómo los robots están reemplazando a los paralegals en Estados Unidos.

[7] Nuevamente, Ashley (2017) profundiza sobre cómo las universidades se están rezagando frente a ese nuevo desarrollo en el mercado laboral.

[8] Trabajo como asesora de la presidencia desde diciembre de 2020 en la entidad.

[9]Los términos para contestar las tutelas son muy cortos por la naturaleza de esta acción. Por esa razón, una ganancia en tiempo es muy importante para mejorar la probabilidad de éxito (no desacato o triunfo en segunda instancia) de la entidad.

3 BASE DE DATOS: CONSTRUCCIÓN

La construcción de los datos tuvo 3 pasos: **1.** La extracción de texto a partir de reconocimiento de imagen. **2**. Limpieza del texto extraído. **3.** Análisis estadístico sobre esa información extraída. A continuación, se detalla cada paso de este proceso.

3.1 EXTRACCIÓN DE TEXTO A PARTIR DE RECONOCIMIENTO DE IMAGEN[10]

La construcción de los datos para este trabajo comienza con una muestra de 275 acciones de tutela escaneadas y recopiladas en formato PDF. La Agencia Nacional de Infraestructura (ANI) aportó los datos que hicieron posible el piloto de este algoritmo, sin embargo, este es un análisis que se puede realizar con las tutelas de cualquier entidad estatal. Si bien había documentos muy bien digitalizados, también había otros menos claros, lo que dificulta la extracción directa de los datos. La *ilustración 1* muestra un contraste entre dos documentos: Uno que está "bien escaneado" y otro que no. También es frecuente encontrar documentos manuscritos. La extracción de texto en estos casos no es trivial y es computacionalmente costosa.

Como el objetivo era conservar los 275 textos, fue necesario ir más allá de los códigos usuales de importación de texto. Para lograr la extracción del texto a partir de los PDFs planos[11] se usó tecnología de reconocimiento de imagen[12]. El software libre más popular para este propósito está optimizado para la extracción de textos en inglés, por esta razón hubo un paso extra[13] de entrenamiento del algoritmo para reconocimiento de tildes. Se probaron diferentes modelos para este propósito y se construyeron unas medidas de calidad para escoger el mejor. La opción escogida era la que más palabras podía reconocer de la imagen del pdf[14].

[10] En este punto se utilizaron diferentes formas de extracción de texto y se escogió el método de PDFminer.py (Shinyama, Yusuke et al, 2007). por ser el que extraía más palabras reconocibles en promedio.

[11] Es decir, los que no tienen las letras reconocidas en el formato. Son esos en los que no se pueden usar los buscadores directos de texto.

[12] El algoritmo formulado para este propósito está incluido en el anexo 3: Código de limpieza de texto.

[13] Entrenamiento de redes neuronales con tarjeta de gráficos discretos para reconocer los acentos del castellano.

[14] Los métodos de prueba de calidad están incluidos en el anexo 3: Código de limpieza de texto.

3.2 LIMPIEZA DEL TEXTO EXTRAÍDO:[15]

Para poder utilizar estos datos con los métodos de análisis de texto que se presentan en este trabajo, fue necesaria una limpieza final. Los textos que se pueden extraer de un PDF escaneado desde fotos tienden a tener niveles intolerables de ruido[16]. Esto hace imposible el análisis directo de esa información. Para solucionar ese problema de ruido en la muestra, se implementaron algoritmos de limpieza por colección de caracteres[17] y luego una partición de cada documento de tokens (palabras o grupos de palabras[18]). Es decir, que se utiliza un software de reconocimiento de imagen para poder identificar mejor cada letra del documento y después de eso, con las letras bien clasificadas y las palabras reconstruidas, se separan los párrafos resultantes en palabras.

[15] Para este propósito se utilizó software libre de Python y una modificación del paquete de vectorización de scikit para python. La modificación se realizó sobre el original de los autores para solucionar un problema de compilación en la matriz resultante del proceso de vectorización. Por la presencia de números complejos en la matriz, no se podía implementar el algoritmo de métodos numéricos de inversión que estaba utilizando el programa original.

[16] Por ejemplo, caracteres que no son palabras. El texto extraído antes del entrenamiento y antes de limpiar se ve así: "Freámbulc constitueiorral. articulo 2 De ta Garta Poiíticaüeereto 25Si de 1Sü15" PRUEBA§$".

17 Esto se refiere a borrar los caracteres y palabras que se consideran supérfluos en un análisis de texto.

[18] Donde cada palabra es una colección de caracteres.

3.3 Análisis estadístico sobre la información extraída[19]

Para el análisis estadístico se produjeron dos tipos de resultados. El primer resultado es una clasificación de los documentos en todos los temas de las tutelas (obtenido por LDA). El segundo tipo de resultado viene de filtrar esos temas para quedarnos solo con la información del derecho fundamental tutelado. En la siguiente sección se explicará con más detalle en qué consisten estos dos tipos de análisis y qué resultados útiles se pueden extraer para un análisis jurídico. En el procesamiento de datos se obtiene primero el LDA y después la información de frecuencia de los derechos fundamentales, pero el orden de la explicación se va a invertir en la siguiente sección del documento.

4 Metodología:

4.1 Análisis de frecuencia

Este es un análisis de la correlación entre la frecuencia[20] con la que se tutela cada derecho fundamental y unos hechos significativos a través del tiempo. Para lograr esto, se tomó directamente el tema "derecho fundamental" como resultado de análisis. Recordando que cada tema se compone de palabras, se correlacionaron unas palabras especiales con diferentes hechos en el tiempo. Esas palabras son las que se obtienen filtrando por el tema "derecho fundamental" e incluyen "debido proceso", "derecho de petición", "mínimo vital", "consulta previa" y "derecho a la vida" entre otros. Ese insumo se contrastó contra una matriz aportada por la ANI de los principales derechos fundamentales tutelados en cada documento. Habiendo revisado que se hubiera extraído bien esa categoría[21], se pueden empezar a buscar relaciones entre los derechos fundamentales de las tutelas y hechos en el tiempo.

Por ejemplo, se revisa la palabra "debido proceso" que pertenece al tema de derechos fundamentales para analizar en qué momentos se tutela más por eso. Esto permite analizar una correspondencia con los momentos electorales del país y con los procesos de contratación pública de la ANI. De ahí, se pueden sacar intuiciones que permiten entender, entre otros, 1. ¿Cuáles son los derechos más tutelados en torno a las adjudicaciones? 2. ¿Cómo se relacionan los tiempos electorales con las tutelas que se interponen contra la ANI?

4.2 Latent Dirichlet Allocation (LDA): El algoritmo de clasificación de textos por temas.

El segundo tipo de resultado (LDA) genera un reporte de cuáles son los temas más frecuentes en estos documentos, en particular, ¿qué temas corresponden a cada tutela? también permite evaluar

[19] El software para este análisis parte de la librería pyLDAvis (Sievert & Shirley, 2014) con modificaciones de la autora para mejorar la visualización.
[20] Cuántas veces ocurre ese evento.
[21] Es decir, que los derechos fundamentales extraídos del LDA coincidan con los que reportó a mano la entidad.

correlaciones entre los temas que componen a los documentos. Todo esto se hace a partir de una categoría de clasificación de texto llamada Latent Dirichlet Allocation (Blei & Jordan, 2003). El supuesto fundamental de este modelo es que los textos se pueden agrupar por temas. Es decir, que hay correlaciones entre los textos y estos no se han generado de manera completamente aleatoria[22]. Partiendo de ese supuesto, lo que el algoritmo hace es identificar esos temas no explícitos (no observados) que explican las relaciones entre los textos y declararlos. Si se considera que las tutelas que se interponen contra una misma entidad pueden tener temas comunes, este supuesto sobre la forma como se generan los textos es un supuesto razonable. Así, cada texto se formaría de una mezcla de distintos temas y lo que hace el algoritmo de LDA es recuperar esas categorías que no podemos observar a priori. Por ejemplo, una tutela que exige la garantía del derecho fundamental de petición sobre una solicitud referente a precios de peajes tendría por lo menos 3 temas: 1. el derecho fundamental tutelado que es el de petición, 2. El tema de peajes, que es frecuente en las tutelas a la entidad y 3. algún tema procesal (como los términos del derecho de petición) que no es fácil de identificar sin revisar más tutelas similares.

En esta estructura de datos los textos se componen de palabras. Entonces cada tema tiene unas palabras específicas que le corresponden. Hay palabras que pueden pertenecer a más de un tema, la partición no es estricta. El algoritmo de LDA lo que hace es atribuirle palabras a cada tema y suponiendo que un texto es una combinación de palabras, le asigna temas a cada documento.

[22] Un ejemplo de textos aleatorios puede ser extraer 1000 tweets de cuentas seleccionadas aleatoriamente alrededor del mundo. En este caso, puede ser que no haya coincidencia ni siquiera de idioma, mucho menos de tema.

5 RESULTADOS

5.1 ANÁLISIS FRECUENCIAL

Esta nube de palabras muestra los resultados más frecuentes de derechos fundamentales tutelados. Naturalmente, hay tutelas en las que se exige la garantía de más de un derecho fundamental a la vez. También hay derechos que casi siempre se exigen en conjunto con algún otro (e.g. estabilidad laboral y derecho al trabajo). Esa correlación entre los resultados se expresa a partir de la cercanía de las palabras en esta gráfica. Por ejemplo, el mínimo vital y vivienda digna (abajo, centrado) son derechos que con frecuencia se reclaman de manera conjunta. De este resultado se puede concluir que los derechos más tutelados son el de petición, igualdad, debido proceso y mínimo vital. En estos derechos debería enfocarse preliminarmente la defensa jurídica de la entidad por ser las tutelas de mayor volumen. Sin embargo, vale la pena revisar si hay algún fenómeno temporal que pueda ayudar a focalizar mejor ese esfuerzo. Las siguientes gráficas muestran la distribución de esos derechos en el tiempo.

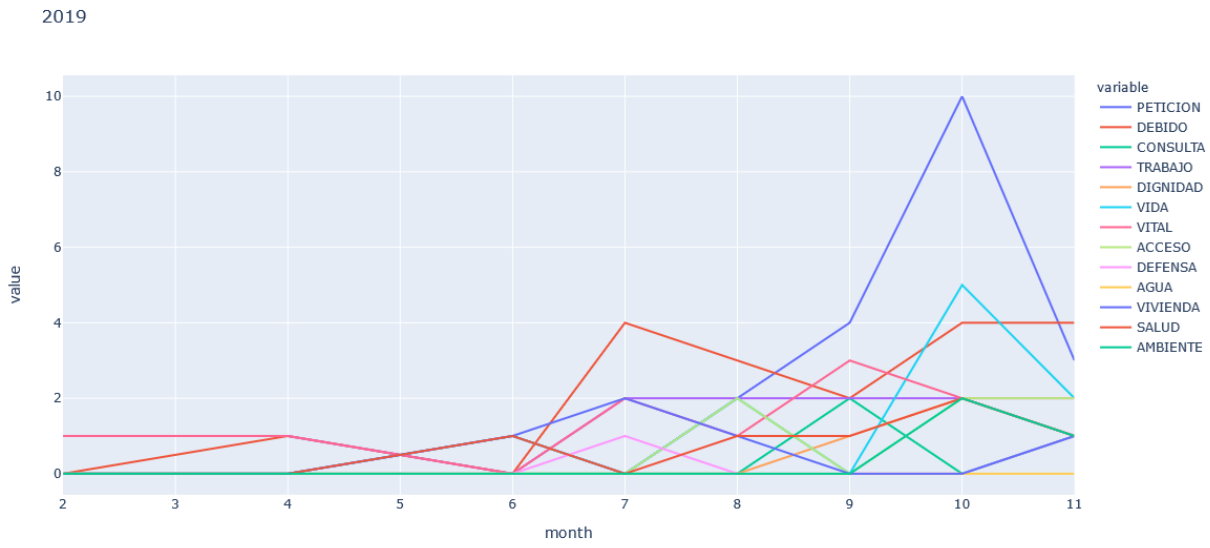

ILUSTRACIÓN 1 FRECUENCIA CON LA QUE SE RECLAMA CADA DERECHO (2019)

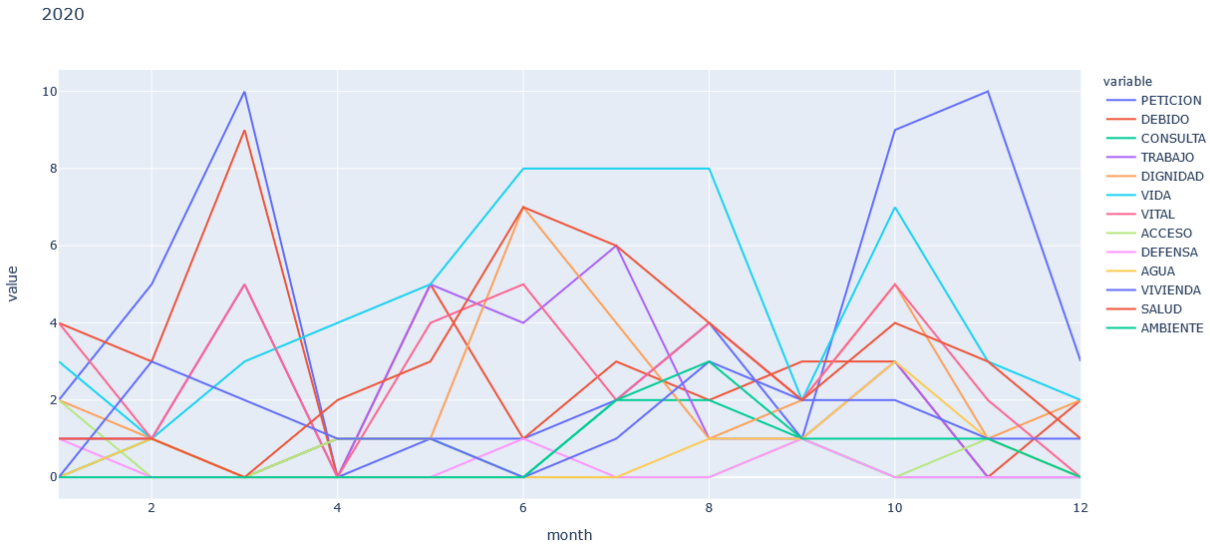

ILUSTRACIÓN 2 FRECUENCIA CON LA QUE SE RECLAMA CADA DERECHO (2020)

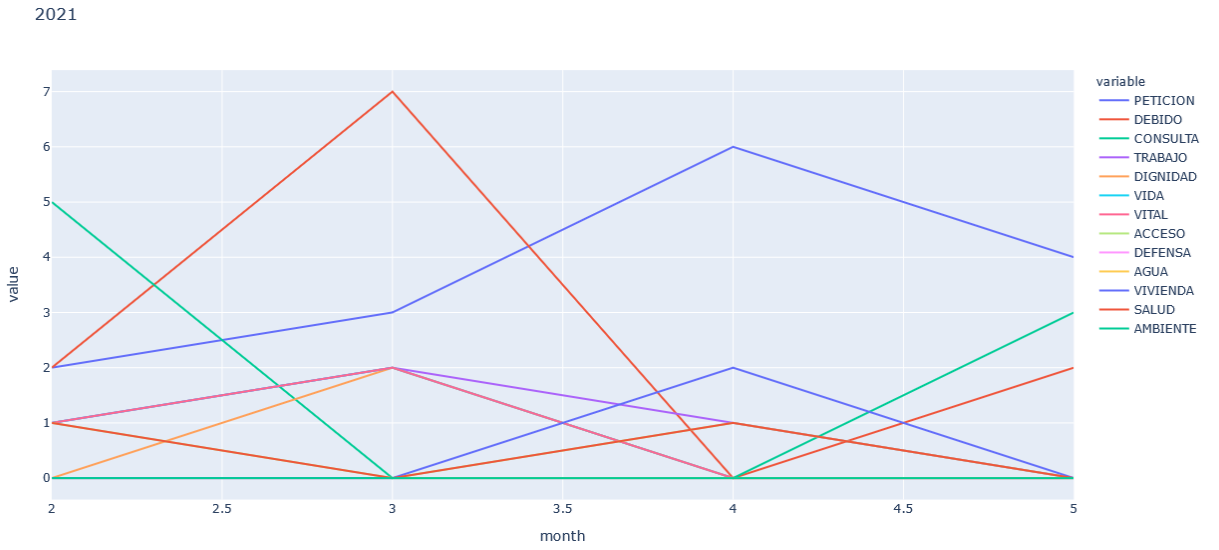

ILUSTRACIÓN 3 FRECUENCIA CON LA QUE SE RECLAMA CADA DERECHO (2021)

De estas gráficas de frecuencia en el tiempo se pueden sacar, por lo menos, 3 conclusiones: La primera, sobre el volumen de tutelas a lo largo del año; la segunda, sobre correlación entre las etapas de los proyectos en la ANI con los derechos más tutelados en esos momentos; y la tercera, sobre el derecho de petición.

Como se puede observar, en los primeros meses del año llegan pocas tutelas a la entidad. En esos meses, el trabajo de defensa jurídica se puede enfocar en otro tipo de acciones. En contraste, el volumen de trabajo aumenta significativamente en los últimos meses del año. En particular, se vuelve mucho más frecuente la llegada de derechos de petición.

El año 2019 y 2020 fueron años de frecuentes entregas parciales de los proyectos para inicio de operación. En los estados financieros de la entidad[23] para 2020 y 2021 se puede observar cómo aumentaron las entregas de unidades funcionales de los proyectos (lo que da derecho a retribución al concesionario) contabilizadas en los activos no corrientes como bienes de uso público. No es coincidencia que con estas entregas hayan aumentado las reclamaciones por derecho a la vida, al mínimo vital, a la consulta previa y al medio ambiente sano (entre otras). Esto dio origen a compensaciones y revisiones de las licencias ambientales de los proyectos. Tiene sentido que el número de reclamos aumentara en este momento clave en el que tanto ciudadanos como concesionarios[24] tienen el interés de hacer presión para que se solucionen las controversias.

Los derechos de petición se hacen especialmente frecuentes al final del año y los temas varían dependiendo del momento electoral del país. Del análisis de LDA se puede ver que los derechos de petición relacionados con información sobre peajes se han intensificado en el año electoral.

[23] Disponibles en https://www.ani.gov.co/rendicion-de-cuentas/informacion-contable-financiera
[24] Los concesionarios presionan porque este es un momento clave para ellos, en el que cada día cuenta para que les reconozcan el derecho a retribución económica por haber cumplido con la entrega de una unidad funcional.

Esto es porque en años electorales, el ejercicio de control político por parte de congresistas aumenta y eso aumenta las solicitudes por derecho de petición a las entidades[25].

Estas tres observaciones, que facilitan la labor de defensa jurídica, son muestras de por qué es útil estructurar y reportar con frecuencia la información a la que tiene acceso la entidad. Un buen uso de los datos vuelve más eficiente la gestión pública. Que los datos se presenten de forma accesible también facilita la veeduría ciudadana y es algo que conviene a todos los ciudadanos.

5.2 LDA: ANÁLISIS POR TEMAS:

https://lauragrandas.github.io/LDA-tutelas/

Los resultados del LDA se pueden visualizar de diferentes maneras. Aquí se propone una visualización por temas (y no por texto) para poder explorar las intuiciones generales para la entidad. En el link están esos resultados. Se pueden explorar las palabras (hacer click) que componen cada tema, la correlación entre temas (los círculos superpuestos, también click) y la importancia relativa de cada palabra en cada tema (λ[26]), entre otros. Un buen punto de partida para explorar la página web sería mover ese parámetro lambda entre 0, 0.6 y 1 para cada tema y después hacer click en las palabras de interés para ver cómo se mueve la distribución en los círculos.

A continuación, se muestran dos ejemplos de lo que se puede ver en esta representación del LDA. El tema 2 incluye términos como sinú, comunidades, consulta, indígenas, etc. Esto sugiere[27] que este tema es el de consultas previas. Este ha sido motivo constante de reclamo de las comunidades, como es usual en megaproyectos de infraestructura. El tema 4, no se refiere a un derecho fundamental específico (como el de consulta previa) pero sí captura una estrategia probatoria frecuente en algunas tutelas. Estas palabras y temas se le atribuyen a cada documento específico con el que se alimentó el algoritmo. Entonces, con la simple inclusión de una nueva tutela en la base de datos, se pueden conocer los temas en los que estaría preliminarmente clasificada. Esto le facilitaría al funcionario o a la entidad preparar una primera fase de defensa jurídica.

[25] Esto lo confirmé en conversaciones con el equipo de defensa jurídica de la ANI, que reporta que esto pasa con cada año electoral en el que los candidatos hacen campaña alrededor de las tarifas y el recaudo de los peajes.

[26] La métrica λ se refiere a la importancia del número de repeticiones de la palabra para determinar si pertenece o no a cada tema. Un λ=1 dice que incluyo todas las palabras más frecuentes en cada tema. Esto no es ideal porque el algoritmo explora otras relaciones entre palabras que van más allá de esa linealidad. Definir un punto óptimo para esa métrica es un proceso complejo que depende del resultado que se esté buscando en cada caso concreto, pero una heurística recomendada por los autores es λ=0.6.

[27] El computador no "conoce" el significado de las palabras, entonces no le da un nombre a cada tema. Solo los agrupa y en un análisis posterior se define a qué corresponde cada uno.

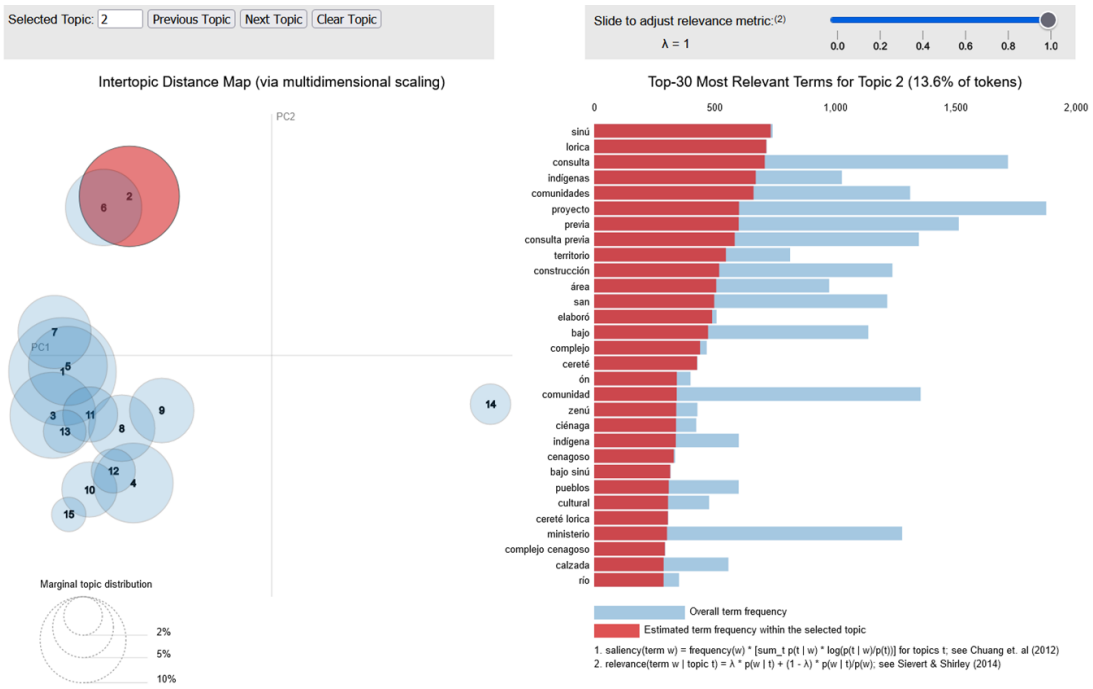

ILUSTRACIÓN 4 TEMA 2: CONSULTA PREVIA

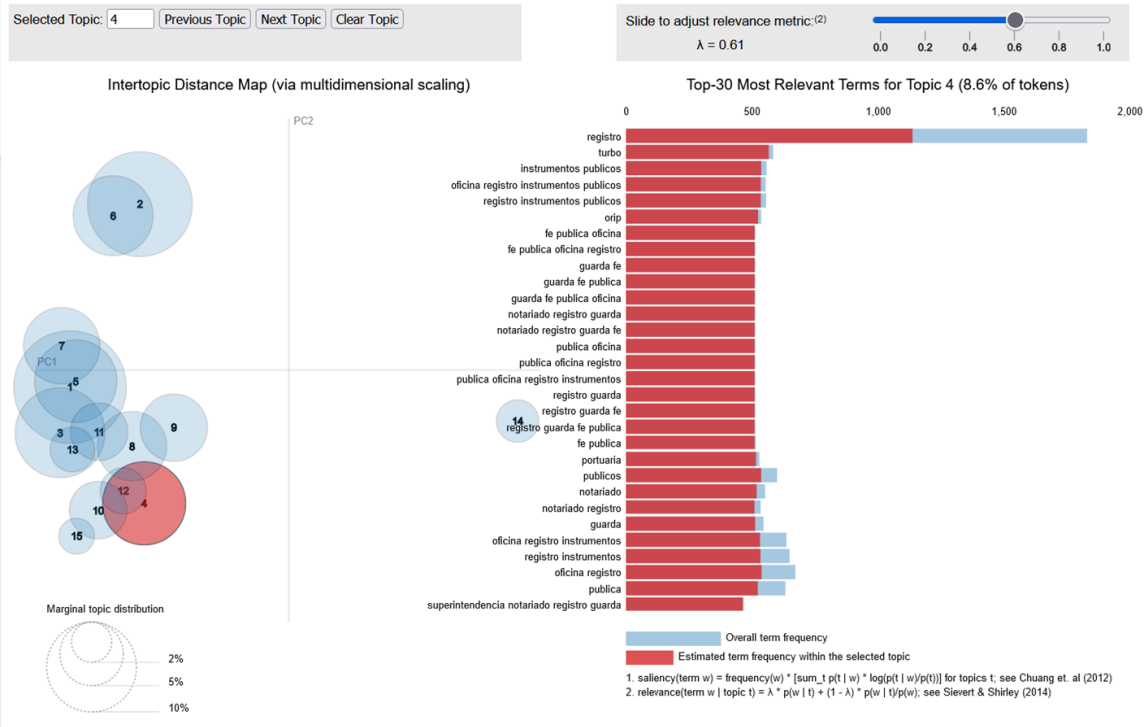

ILUSTRACIÓN 5 TEMA 4: PROCESAL

6 APORTES:

6.1 ANÁLISIS FRECUENCIAL

El análisis de la frecuencia con la que se tutela cada derecho fundamental permite identificar patrones de temporalidad que hacen más fácil la labor de defensa jurídica al focalizar los esfuerzos del equipo a temas previsibles. Por ejemplo, se puede observar que los derechos de petición son especialmente frecuentes en los últimos meses del año. También se puede ver que los derechos de petición que buscan la protección del debido proceso se concentran hacia el tercer trimestre. Estas observaciones facilitan la distribución del trabajo en los equipos de defensa jurídica de las entidades. En el caso de la ANI, el análisis de la frecuencia con la que se reciben tutelas permite distribuir labores en el equipo de manera que hay más personas contestando tutelas en los meses más populares del año y menos personas cuando el volumen baja.

Para el caso específico de la ANI, un análisis de correlación temporal con momentos hito de las adjudicaciones y ejecución de proyectos de concesión permite ver que las tutelas por consulta previa, mínimo vital y debido proceso se hacen más frecuentes cuando está a punto de cerrarse la adjudicación de un proyecto. Considerando que los términos para contestar las tutelas constituyen una restricción activa[28] y fuerte para los trabajadores, la preparación (antes de que llegue la tutela) es muy importante para dar respuesta a aquellas que constituyen casos complejos para la entidad.

6.2 LDA

Esta metodología es especialmente útil para hacer más eficientes las tareas de los funcionarios que se dedican a la defensa jurídica de las entidades públicas. La posibilidad de conocer los temas que tiene la tutela antes de leerla permite una distribución más fácil de la carga de trabajo. La acción de tutela tiene unos términos muy exigentes. Por esta razón, es común que las respuestas a las mismas se construyan con formatos preestablecidos. Conocer cuáles son los principales temas tutelados permite construir mejores formatos y saber cuál utilizar con la llegada de cada tutela. En este caso, el algoritmo se entrenó con 275 textos. Cuando llega el documento no. 276 la delimitación de los temas que lo componen son inmediatos.

También facilita la identificación de problemas estructurales que esté teniendo la entidad y que lleven a los ciudadanos a interponer estas tutelas en primer lugar. Un ejemplo de esto puede ser que en cada momento clave de la ejecución de alguna concesión se estén tutelando los derechos de una misma comunidad o se esté relacionando al mismo derecho fundamental (e.g. el agua). Otro ejemplo es el tema de los derechos de petición. Que una entidad tenga este volumen de derechos de petición que terminan en tutelas[29] es un indicio de que hay un problema por el que no se está dando respuesta a tiempo a esas peticiones de los ciudadanos. Este tipo de análisis también facilita

[28] Es decir que el límite de tiempo que se impone para responder la tutela sí importa. En ausencia de ese límite el tiempo natural que se tomaría en responder cada acción sería más largo.

[29] No todos los derechos de petición terminan en tutelas, si la entidad respondiera oportunamente, los ciudadanos no se verían en la necesidad de acudir a esta acción.

la veeduría ciudadana a la gestión de las entidades, pues hace explícitos esos mismos problemas estructurales a los que hace referencia el punto anterior.

7 EXPANSIONES Y FUTUROS DESARROLLOS POSIBLES

Este trabajo tiene múltiples expansiones posibles. En primer lugar, porque el software que se diseñó para la digitalización de textos crudos y su procesamiento permite llevar este análisis a cualquier entidad pública. Así, cualquier entidad que reciba tutelas de manera "masiva" puede aprovechar estas herramientas para mejorar su estrategia de defensa y hacer más eficiente el trabajo de los funcionarios que responden a estas tutelas. La segunda expansión es a otras acciones, este análisis no se limita a la tutela. Cualquier acción que tenga una estructura de texto similar[30] a la de las tutelas puede ser objeto de esta revisión. La tercera expansión posible ya entra en el campo de nuevos algoritmos de análisis de texto. En este caso se usó una lógica de estadística bayesiana[31] clasificar los textos por temas, pero hay muchas más opciones teniendo los textos digitalizados. Una de las opciones más atractivas es la generación automática de resúmenes con algoritmos de procesamiento de lenguaje natural[32].

Este trabajo es un ejemplo de por qué la adopción de estas nuevas tecnologías es vital para hacer más eficiente el ejercicio del derecho. La adopción de software como el que se propone aquí ha tenido resistencia de la comunidad jurídica por múltiples razones. Hay una preocupación latente de que los robots terminen por reemplazar a los humanos en labores que a juicio de algunos deberían ser irremplazables. Sin embargo, la realidad es que el aprendizaje de máquinas y la inteligencia artificial son más una herramienta que un sustituto para los abogados. Negarse a aceptar este cambio le niega oportunidades de progreso al derecho. En el litigio, la automatización de labores manuales que antes costaban horas de trabajo "calificado" permite ahorros muy importantes que se pueden traducir en mayor acceso a asesoría jurídica para la población tradicionalmente excluida. En el sector público, conocer los beneficios que puede traer la implementación de estas herramientas motiva a las entidades a hacer mejor gestión de su conocimiento, los impulsa a la adecuada digitalización. Esto es porque entre menos estructurados estén los datos, más costoso[33] es su procesamiento. Las nuevas tecnologías traen más oportunidades que peligros y tienen un potencial sin precedentes para democratizar[34] el derecho y mejorar nuestros procesos.

[30] Es decir, que tenga al menos 200 palabras y que se recoja en un texto escaneable. Los requisitos son muy laxos.
[31] El LDA se basa en estadística bayesiana.
[32] Merchant y Pande (2018). Construyeron un algoritmo que hace esto con casos civiles y penales de distintas cortes en Estados Unidos.
[33] Más costoso computacionalmente, pero también en términos de capital humano. Si los datos tienen ruido como en el caso del "mal ejemplo" de la ilustración 1, el software para procesar los datos necesita pasos y entrenamiento. Esto invita a las entidades públicas a recoger mejor información, que requiera de menos pre-procesamiento.
[34] Botero (2019) en su artículo "La robótica y la inteligencia artificial podrían reemplazar a los abogados (al menos a algunos)" es un buen ejemplo local de esta postura -aunque no tan pesimista como algunos.

Este ejercicio es un aporte a una discusión más grande que es la resistencia al cambio de los abogados para incluir la tecnología en su profesión. Por ejemplo, es frecuente encontrarse con la percepción de que los procesos no digitalizados/físicos son más seguros y se niega el uso de tecnologías que facilitarían el acceso a la justicia con medios remotos de conexión. Esta postura lleva a que no haya respaldos digitales de los expedientes y no haya posibilidad de recuperación cuando hay un daño físico a los mismos. Otra arista del problema es un miedo latente a los sesgos que los clasificadores estadísticos (una forma de algoritmo) pueden inducir en los procesos. Esto no es una preocupación infundada[35], pero por problemas de algunos algoritmos no se puede condenar a todos los demás. Incluso se ha demostrado que, en los casos más famosos, que son los de sesgos discriminatorios (raciales) en los algoritmos de predicción de reincidencia, el sesgo del algoritmo solo reproduce el sesgo que ya existe en la realidad[36].

El mecanismo de clasificación (LDA) presentado en este trabajo se escogió con la intención de mostrar que hay implementaciones tecnológicas que mitigan esos problemas de sesgos estadísticos en la clasificación de casos jurídicos. En el caso de las tutelas siempre habrá un funcionario que tenga que leer cada documento. Una clasificación previa solo facilita y acelera esa labor, pero, si un tema importante no fuera detectado, el funcionario encargado lo descubriría al leerla. Esto es una prueba de las oportunidades de progreso en el ejercicio del derecho que se pierden por la resistencia al cambio.

8 REFERENCIAS:

Perasso, V. (2016). Qué es la cuarta revolución industrial (y por qué debería preocuparnos*). BBC Mundo*, *12*.

Schwab, K. (2017). The fourth industrial revolution. *Currency*.

Puschmann, T. (2017). Fintech. *Business & Information Systems Engineering*, 59(1), 69-76.

Sevilla, A. (2021) Best AI writer of 2021. *Tech Radar* on March 09, 2021. Disponible en https://www.techradar.com/best/ai-writer

[35] Tal es el caso de Compas v. Propublica donde se disputa un posible sesgo racial en el algoritmo de predicción de reincidencia. El sistema no es transparente y cuando se hace un análisis cuidadoso del mismo se encuentra una relevancia indeseable de la variable de raza, por la que se termina por calcular una probabilidad de reincidencia mucho más alta para una persona negra[35] que para una persona blanca de las mismas características (Rizer y Watney, 2018).

[36] La base de datos con la que se entrenaron los algoritmos tenía muchas más personas de raza negra que personas blancas por una tendencia de la policía a perseguir a estas personas. Entonces, el algoritmo lo único que hace es reproducir y mostrarnos los sesgos con los que ya vivimos.

Ashley, K. D. (2017). *Artificial intelligence and legal analytics: new tools for law practice in the digital age*. Cambridge University Press.

Shinyama, Yusuke. (2007). PDFMiner - Python PDF Parser.

Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, *Journal of Machine Learning Research* 12, pp. 2825-2830, 2011.

Sievert, C., & Shirley, K. (2014, June). LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the workshop on interactive language learning, visualization, and interfaces* (pp. 63-70).

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, 3, 993-1022.

Merchant, K., & Pande, Y. (2018, September). Nlp based latent semantic analysis for legal text summarization. In *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)* (pp. 1803-1807). IEEE.

Botero, A. (2019). La robótica y la inteligencia artificial podrían reemplazar a los abogados (al menos a algunos). *Revista Semana*. Disponible en https://www.semana.com/impresa/internet/articulo/la-robotica-y-la-inteligencia-artificial-podrian-reemplazar-a-los-abogados-al-menos-a-algunos/78240/

Brown, T. (2008). Design thinking. *Harvard business review*, *86*(6), 84.

Rowe, P. G. (1987). *Design thinking*. MIT press.

ANEXO:

Repositorio de los códigos con los que se construyó este documento: https://github.com/LauraGrandas/text-analysis-Tutelas-Colombia-

- Criterio	Calificación	Comentarios
Identificación de un problema legal real		
Exploración del reto e ideación y desarrollo de un prototipo.		
¿El prototipo resuelve el problema legal identificado?		
Criterios de forma: claridad, sintaxis y orden del texto.		
¿El prototipo incorporó la experiencia de usuario de las personas involucradas en el proceso explorado?		