Propuesta recibida en junio 2022 y aceptada para publicación en diciembre 2022.

ARAPP: Análisis y Resumen Automático de Políticas de Privacidad

Analysis and Automatic Summary of Privacy Policies

Rodrigo Alfaro
Pontificia Universidad Católica de Valparaíso

René Venegas
Pontificia Universidad Católica de Valparaíso

Alan Bronfman
Pontificia Universidad Católica de Valparaíso

Miguel Valenzuela
Pontificia Universidad Católica de Valparaíso

Stephanie Riff
Pontificia Universidad Católica de Valparaíso

Enrique Sologuren
Universidad del Desarrollo

Resumen

Un derecho fundamental de los usuarios de aplicaciones informáticas es que puedan conocer las políticas de privacidad (PP) que tales aplicaciones establecen, en particular es relevante que conozcan acerca del tratamiento que aceptan sobre el uso de sus datos. No obstante, estas PP son muy extensas y escritas en un lenguaje administrativo-jurídico y comercial, lo que dificulta su lectura y comprensión. El objetivo de este artículo es resumir automatizadamente las PP de cinco aplicaciones de redes sociales (Facebook, Twitter, TikTok, Snapchat e Instagram) en español, a través de técnicas extractivas y abstractivas. Para ello se utilizan tres aproximaciones de representación desde el Procesamiento de Lenguaje Natural, estas son: Teoría de Grafos, TF-IDF y Gensim. A partir de ellas, se generan automáticamente 15 resúmenes, los que son evaluados por un experto en derecho, para medir la legibilidad y relevancia en base a 20 preguntas confeccionadas por un estudio de la Universidad de Austin, Texas (Zaeem et al., 2018). Por último, a partir de una clasificación de cada política de privacidad, según distintos factores de riesgos, se comprueba que el método Gensim es el más adecuado para la representación y resumen. Además se identifica a Snapchat como la aplicación que mejor cumple dichos factores.

Palabras clave

resumen automático; políticas de privacidad; factores de riesgo; textos jurídicos; Gensim; redes sociales

Abstract

A fundamental right of the users of computer applications is that they can know the privacy policies (PP) that such applications establish. It is particularly relevant that they know about the treatment that they accept regarding the use of their data. However, these PP are very extensive and written in administrative-legal and commercial language, which makes them difficult to read and understand. The aim of this paper is to automatically summarize the PPs of five social network applications (Facebook, Twitter, TikTok, Snapchat and Instagram) in spanish, through extractive and abstractive techniques. For this purpose, three representation approaches from Natural Language Processing are used, these are: Graph Analysis, TF-IDF and Gensim. Fifteen summaries were automatically generated and evaluated in order to measure the readability and relevance, by an expert in law, based on 20 questions prepared by a study of the University of Austin, Texas (Zaeem et al., 2018). Finally, based on a classification of each privacy policy according to different risk factors, the Gensim method is found to be the most suitable for the representation and summarization of the PP's. The PP of Snapchat is also identified as the application that best meets these risk factors.

Keywords

automatic summarization; private policies; risk factors; legal texts; Gensim; social networks

1. Introducción

La privacidad es entendida como todo ámbito de la vida privada que se tiene derecho a proteger de cualquier intromisión (RAE, 2021). Hacia fines del siglo XIX el derecho a la privacidad fue entendido como *"the right to be let alone"*, el derecho a ser dejado solo (Warren & Brandeis, 1890). Este derecho protege la vida privada como un ámbito o espacio libre de intrusión o invasión por parte de un tercero, sin perjuicio de intromisiones justificadas por una necesidad manifiesta de una comunidad que vive en un estado de derecho (Corral, 2000). En este sentido, el reconocimiento y garantía de la esfera privada no solo atiende a una dimensión individual o subjetiva, sino que tam-

DOI: 10.21814/lm.14.2.375

bién considera la defensa de la privacidad desde una dimensión colectiva y social, que coadyuva al mantenimiento y avance del sistema democrático (Saldaña, 2012).

Un elemento distintivo de aquello que forma parte de la vida privada es el control sobre el acceso, el cual pertenece a su titular. La pérdida o menoscabo de dicho control constituye una pérdida o amenaza sobre el derecho a la vida privada.

El desarrollo de la informática y el tratamiento de datos en la segunda mitad del siglo pasado puso de manifiesto su incidencia en ámbitos que forman parte de la vida privada, lo que impulsó la construcción del concepto de datos personales y el establecimiento de medios para su protección. Hace más de cuarenta años, leyes tales como la *Datenshutz* (Alemania, 1970), *Data Lag* (Suecia, 1973) o la *Privacy Act* (Estados Unidos, 1974), entre otras, reconocieron la necesidad de proteger los datos personales mediante la regulación de su uso, almacenamiento y transmisión (de la Cueva & Fernández-Miranda, 1990; Pérez Luño, 1984). Así, por ejemplo, en 1983 un fallo del Tribunal Constitucional alemán precisó esta tutela con la idea de *autodeterminación informativa*, la que constituye un derecho que permite al individuo decidir por sí mismo cuándo y dentro de qué límites procede revelar situaciones referidas a su propia vida (Schwabe, 2009; TCCh, 1989).

El derecho a la vida privada, entonces, reconoce a cada persona el poder de decidir acerca del uso, almacenamiento y transmisión de sus datos personales. Esta facultad de decidir requiere conocer qué datos personales serán registrados en las bases de datos de cada aplicación y cómo serán tratados y utilizados. Las políticas de privacidad (PP) contienen una declaración pública, de base jurídica y estándar, de los propietarios o controladores de una aplicación dirigida a sus potenciales usuarios respecto a los datos que ella utilizará y su tratamiento. Dichas PP han sido redactadas de modo unilateral por los propietarios o controladores de la aplicación de acuerdo con su conveniencia e interés y, por lo mismo, no contienen necesariamente la información pertinente y relevante acerca de la afectación de la privacidad derivada del uso de la aplicación. Es claro que el registro y tratamiento de datos por parte de estas aplicaciones o plataformas genera instrumentos precisos de focalización publicitaria que son difíciles de compatibilizar con la tutela jurídica de la vida privada. En las dimensiones globales y potencial económico del mercado creado por la red internet, la opacidad de las PP ofrece un claro beneficio para los propietarios o controladores de las aplicaciones.

Ahora bien, las PP son una herramienta esencial para la toma de decisiones del potencial usuario en lo que a su derecho a la vida privada se refiere (Anastasopoulou et al., 2017). Muchas veces son extensas, confusas y utilizan de modo reiterado términos técnicos, lo que dificulta su lectura. Con frecuencia son aceptadas sin ser comprendidas por los usuarios que desean o necesitan acceder a los servicios de la aplicación ofrecida. De acuerdo con los estudios que informan acerca de los niveles de lecturabilidad, la comprensión de este tipo de documento requiere de personas que al menos hayan completado dos años de estudios universitarios (Zaeem et al., 2018). La complejidad de estos textos genera costos de oportunidad del orden de U$D 781 mil millones de dólares (McDonald & Cranor, 2009), si se considera el tiempo de lectura que requieren, los cuales ascienden de forma individual a 201 horas por año. Lo que excede el porcentaje de tiempo dedicado en promedio a la navegación en la web (Meier et al., 2020). Este costo podría reducirse si es posible comunicar al usuario de modo más breve y rápido aquello que es esencial para decidir sobre la tutela de su vida privada.

En este marco, adquieren especial importancia las técnicas para el análisis de los textos que utilizan lenguaje jurídico o legal y que exhiben una alta dificultad para su comprensión en las personas debido a su opacidad discursiva (Meza et al., 2021; Montolío & López Samaniego, 2008). Dentro de las técnicas disponibles para el estudio de estas formas textuales, en este ejercicio, utilizamos aquellas dirigidas a evaluar la posibilidad de generar información sintética y comprensible para el usuario de aplicaciones tecnológicas como insumo para el ejercicio de su derecho a la privacidad (Montolío & Tascón, 2020). Ello porque solo la comprensión cabal del uso, almacenamiento y transmisión de los datos personales asociado a la utilización de cada aplicación permite, desde el punto de vista jurídico, otorgar un consentimiento válido para el acceso al ámbito protegido por el derecho a la vida privada.

El enfoque de la comunicación clara agrega otro paradigma aplicable al análisis de las PP. La comunicación clara precave litigios y conflictos entre empresas y los usuarios de sus servicios, en este caso, gracias a la oportuna entrega de información completa e inteligible sobre la afectación de la privacidad ocasionada por el uso de una aplicación. La comunicación clara enfatiza un cambio de cultura comunicativa en las empresas y organizaciones, promoviendo que las cláusulas de los contratos sean claras, precisas y con una estructura lo más transparente y visual

posible. Todo ello con el fin de mejorar la experiencia del usuario así como de que cualquier cliente pueda entenderlas y saber a qué está adhiriendo (da Cunha & Escobar, 2021; Montolío & López Samaniego, 2008).

Los estudios en el ámbito del lenguaje jurídico claro son de reciente data en el idioma español. Si bien existen propuestas de modernización, clarificación y normalización del discurso jurídico en España (García Calderón, 2019; Montolío & López Samaniego, 2008) y Latinoamérica (Cucatto, 2011; Meza et al., 2021; Poblete & Fuenzalida González, 2018), los intentos por solucionar estos problemas con ayuda de herramientas automatizadas son aún escasos Ruidías et al. (2018). En este sentido, el trabajo presentado por Valenzuela et al. (2020) y el realizado en este artículo busca llenar este vacío, en especial considerando la potencial utilidad de las PP en la protección efectiva de la privacidad de millones de usuarios de las aplicaciones más populares. Según el informe *Global Digital Overview* 2022 (DataReportal, 2022), elaborado por las empresas *We are Social* y *Hootsuite*, en el mes de abril, el 63 % de las personas del planeta usaron internet y, de ellas, más de 326 millones utilizaron las redes sociales. En el caso de facebook, por ejemplo, el número mensual de usuarios activos de en el año 2021 fue de 2,9 billones (DataReportal, 2022).

En este estudio se somete a las PP en español de Twitter, Instagram, Facebook, Snapchat y TikTok a métodos automatizados de resumen con el propósito de mejorar el conocimiento y comprensión de los usuarios sobre el tratamiento que recibirán sus datos personales. Para esta labor se intenta identificar los factores de riesgo más importantes de las políticas de privacidad declaradas y, sobre esta base, seleccionar el resumen automático de mayor calidad.

En la identificación de los factores de riesgo se utiliza el estudio de Zaeem et al. (2020), que fundamenta la herramienta que se ofrece con el nombre de *PrivacyCheck*. *PrivacyCheck* es una aplicación innovadora que analiza las PP mediante la extracción automática de textos en línea, utilizando modelos de minería de textos. Si bien esta herramienta presenta una propuesta para evaluar PP en idioma inglés, no se presenta un resumen para que usuarios comprendan qué están aprobando cuando aceptan una PP.

Como se indicó arriba, el *input* del proceso son las PP de Twitter, Instagram, Facebook, Snapchat y TikTok que se utilizaron en Chile durante el período octubre 2020–enero 2021. En el proceso se aplican dos métodos de técnica extractiva (Análisis de Grafos y TF-IDF) y uno de técnica abstractiva (Gensim), para obtener como resultado quince resúmenes automáticos distintos. Las dos primeras técnicas tienen en común la extracción y priorización de oraciones de los textos originales, en tanto que en la tercera se hace uso de técnicas de similitud semántica, para generar oraciones parafraseadas, y luego se priorizan las más relevantes (ver 3.3). Las variables analizadas en los textos son la cantidad de palabras y el tiempo de lectura; así como la legibilidad y la relevancia (van de Luijtgaarden, 2019). La primera se define como la fluidez, gramaticalidad y coherencia, en tanto la segunda como presencia de información importante en el resumen. Además se complementa el análisis con las veinte preguntas proporcionadas por el estudio de *PrivacyCheck* para medir el riesgo en los resúmenes de las PP.

Cabe señalar que el desarrollo de esta investigación enfrentó tres dificultades. Por una parte, existen pocos estudios relacionados al análisis y resumen automático de las políticas de privacidad (ARAPP) en español, por lo que no se cuenta con un *corpus* público de políticas de privacidad, especialmente de las redes sociales y, menos aún, con un algoritmo eficiente para su desarrollo dentro de América Latina. Por otra parte, no se cuenta con resúmenes de referencia elaborados por humanos que sirvan para construir una evaluación computacional de los resúmenes automáticos. Finalmente, tampoco existe una regulación de protección de datos que sea imperativo cumplir por las aplicaciones analizadas (sin perjuicio de la existencia de normas y estándares nacionales e internacionales que no son vinculantes para ellas) y que nos permita determinar el correcto contenido de una PP.

La exploración interdisciplinar que proponemos es el inicio de futuras investigaciones y podría generar beneficios para los usuarios de aplicaciones informáticas como también para las empresas que buscan avanzar hacia una relación comunicativa más transparente, leal y efectiva con sus usuarios.

2. Marco Teórico

El Procesamiento de Lenguaje Natural (PLN) es una subdisciplina aplicada que se centra en investigar y formular soluciones computacionales que faciliten la interrelación hombre-máquina mediante la automatización de procesos relacionados con la comunicación humana (Zhang & Lu, 2021). Por lo tanto, el PLN es la manipulación informática del lenguaje natural que integra la lingüística y la matemática (Rodrigo & Allende, 2020) y que puede aplicarse con diversos matices

(Bird et al., 2009). Entre sus potenciales aplicaciones se encuentran la recuperación de información, la clasificación automatizada, la generación automática de resúmenes, entre otras.

La generación automática de resúmenes nace con Luhn (1958) y es un área del PLN que según Maybury (1995), puede entenderse como un conjunto de procesos que "destila la información más importante de una fuente (o fuentes) para producir una versión abreviada de la información original para un usuario o tareas en particular" (p. 735). Existen dos tipos de técnicas de resumen automático: extractivas y abstractivas. Rane & Govilkar (2019) explican que las técnicas extractivas generan resúmenes mediante la frecuencia de las frases del texto original para ser clasificadas por prioridad, la cual está basada en características lingüísticas y estadísticas, para luego combinarlas en un solo texto de salida. Según Hernández (2017), su implementación es más fácil que su contraparte abstractiva y garantiza la coherencia mínima de la información generada. En cuanto a la técnica abstractiva, las oraciones generadas son parafraseadas del texto de entrada, esto es, se incluyen palabras que no aparecen en el texto original (Widyassari et al., 2022), obteniéndose un resultado similar a un resumen creado por una persona. Esta técnica es más compleja, ya que necesita procesamiento de lenguaje natural y recursos de lenguaje para obtener palabras similares provenientes de otros textos (Gambhir & Gupta, 2016).

Entre las técnicas extractivas se cuenta *Term frequency - Inverse document frecuency* (TF-IDF) Christian et al. (2016), que calcula la estadística de una palabra dentro de un documento y un *corpus* de documentos (Salton & Buckley, 1988), y representa aquellas palabras más relevantes del texto y subrepresenta aquellas que se repiten en todos los textos del corpus. El término de frecuencia (TF) significa la frecuencia bruta de un término en un documento y el término de la frecuencia inversa en el corpus (IDF) es una medida que indica si la palabra es común o rara en todos los documentos analizados.

Otra técnica es el método basado en grafos, que fue planteado para resúmenes automáticos por Erkan & Radev (2004). La técnica se denomina *LexRank* y define las aristas como relaciones de co-ocurrencia de términos entre oraciones. Ella ha sido perfeccionada en el tiempo, evolucionando desde los modelos bipartitos (Wan et al., 2021) hasta los modelos de hipergrafo. El método grafo-analítico se concibe como un modelo más preciso, en contraposición al método por frecuencias. Como es posible notar, el método grafo-analítico requiere más tiempo de procesamiento a medida que los textos de entrada se incrementan. Aun así, es capaz de mostrar información relevante (Karmaker & Hossen, 2019) y, por lo mismo, más preciso que el método por frecuencias. Lierde & Chow (2019) presentan el método grafo en el que los nodos son oraciones y las aristas el número de palabras en común entre las oraciones. Ellos proponen un nuevo modelo de hipergrafo para capturar la relación semántica entre oraciones. En él se utilza un enfoque de selección de oraciones, basado en la maximización de la relevancia de oraciones individuales y la cobertura temática, y un algoritmo polinomial basado en la teoría de funciones submodulares para resolver problemas de optimización.

Una técnica abstractiva actual es la proporcionada por Řehůřek (2021), denominada Gensim, la cual está documentada en lenguaje de programación para su utilización mediante una librería informática. Esta es una librería abierta (*open-source*), desarrollada en Python para representar documentos en vectores semánticos de manera eficiente computacionalmente. Está diseñada para ser de fácil utilización. Gensim procesa textos planos con algoritmos de aprendizaje de máquina no supervisados, entre ellos, Word2Vec, FastText, Latent Semantic Indexing (LSI, LSA, LsiModel), Latent Dirichlet Allocation (LDA, LdaModel), etc. Estos permiten descubrir automáticamente la estructura semántica de los textos utilizados para el entrenamiento de los algoritmos, a través del análisis de los patrones estadísticos de co-ocurrencia de la información lingüística del texto. Estos algoritmos no requieren información aportada por el analista humano y solo necesitan el texto plano como entrada del modelo. Así se evita el sesgo humano en el análisis. Una vez identificados los patrones estadísticos de palabras, frases u oraciones, estos pueden ser representados semánticamente y ser utilizados para la comparación de documentos, a través de las técnicas de similitud de tópicos. Para el caso del resumen automatizado, se ofrece un módulo en el que dado un texto se extraen de este una o más oraciones relevantes, así como las palabras clave y se entrega un resumen. Se utiliza para ello una variación del algoritmo TextRank (Barrios et al., 2016).

En relación con la calidad de los resúmenes que puedan ser producidos automáticamente, van de Luijtgaarden (2019) propone una metodología que identifica la legibilidad y la relevancia en una escala del 1 al 10, de acuerdo con evaluadores humanos, en nuestro caso un abogado.

Un segundo estándar de calidad podría construirse desde la regulación. En Europa en ma-

yo de 2018 entró en vigor el *Reglamento General de Protección de Datos* (RGPD) (2016/679), que contiene un conjunto de principios que las empresas deben observar, tales como el de responsabilidad, la transparencia y la protección de datos personales. Este reglamento europeo dispone de medidas técnicas y organizativas para garantizar que los datos sean únicamente objeto de tratamiento para los fines específicos que sean entregados, reduciendo la extensión del tratamiento, limitando el plazo de conservación y su accesibilidad. En el estado de California en Estados Unidos rige la Ley de privacidad del consumidor (CCPA) (Becerra, 2020), la que estipula que las empresas reguladas deberán proteger la información personal de los consumidores en California y en Chile existe la (Ley Chile, 2014) 19.628 para la protección de datos de carácter personal. También existen directrices de la OCDE sobre protección de la privacidad y los flujos transfronterizos (OCDE, 2002). El estándar de calidad que podría nacer de la identificación en los resúmenes de referencias relevantes a las materias objeto de reglas y principios podría adolecer de problemas de generalidad (proveniente de las propias normas y la técnica legislativa) y territorialidad, toda vez que dicho estándar solo tendría valor para el territorio estatal donde la correspondiente regulación tiene vigencia y aplicación.

Por último, Zaeem et al. (2018) proponen *Privacy Check*, una herramienta que aborda el análisis de las políticas de privacidad mediante la extracción automática de resúmenes de las políticas de privacidad en línea, utilizando modelos de minería de textos que responden a diez preguntas dirigidas a evaluar la privacidad y seguridad de los datos del usuario. Las preguntas (traducidas al español) son:

1. ¿Qué tan bien protege este sitio web su dirección de correo electrónico?
2. ¿Qué tan bien protege este sitio web la información y la dirección de su tarjeta de crédito?
3. ¿Qué tan bien maneja este sitio web su número de seguro social?
4. ¿Este sitio web utiliza o comparte su información de identificación personal con fines de marketing?
5. ¿Este sitio web rastrea o comparte su ubicación?
6. ¿Este sitio web recopila información de identificación personal de niños menores de 13 años?
7. ¿Este sitio web comparte su información con las fuerzas del orden?
8. ¿Este sitio web notifica o le permite darse de baja después de cambiar su política de privacidad?
9. ¿Este sitio web le permite editar o eliminar su información de sus registros?
10. ¿Este sitio web recopila o comparte datos agregados relacionados con su identidad o comportamiento?

Este modelo se entrenó con cuatrocientas empresas de diversas industrias. Su implementación utiliza una extensión en el navegador Chrome, gratuita y disponible para todo público, aplicable a cualquier política de privacidad en línea. Los resultados demostraron que los resúmenes de *PrivacyCheck* son precisos entre el 40 % y el 73 %. Finalmente, *PrivacyCheck* clasifica el riesgo de las Políticas de Privacidad (PP) en tres niveles (verde, amarillo o rojo) para cada pregunta (factor de riesgo) en más de cuatrocientas compañías independientes.

En 2020 se presentó *PrivacyCheck v2*, que agregó diez nuevos factores de riesgo a partir del Reglamento General de Protección de Datos (RGPD), arriba mencionado, e incorporó una herramienta de análisis de las PP. Las 10 preguntas que se incorporaron son:

1. ¿Comparte la información del usuario con otros sitios web solo con el consentimiento del usuario?
2. ¿Revela dónde se encuentra la empresa / dónde se procesará y transferirá la información de identificación personal del usuario?
3. ¿Apoya el derecho al olvido?
4. Si retienen información de identificación personal con fines legales después de la solicitud del usuario de ser olvidado, ¿informarán al usuario?
5. ¿Permite al usuario rechazar el uso de la información de identificación personal del usuario?
6. ¿Restringe el uso de información de identificación personal de niños menores de 16 años?
7. ¿Advierte al usuario que sus datos están encriptados incluso en reposo?
8. ¿Solicita el consentimiento informado del usuario para realizar el procesamiento de datos?
9. ¿Implementa todos los principios de protección de datos por diseño y por defecto?
10. ¿Notifica al usuario las violaciones de seguridad sin demoras indebidas?

Si bien, la herramienta *PrivacyCheck* permite evaluar riesgos de las PP en idioma inglés, no se aplica al idioma español, ni tampoco permite a los usuarios comprender, mediante un texto resumido, qué están aprobando cuando aceptan una PP.

3. Metodología

En primer lugar, se recopilaron los documentos a ser resumidos por medio de la identificación de las principales aplicaciones informáticas, con base en un criterio intencionado por los investigadores. La recopilación, en este sentido, fue de carácter manual. Luego, se aplicaron las técnicas de elaboración automatizada de resumen consideradas en esta investigación. Finalmente, se comparó la calidad de los resúmenes, utilizando criterio de experto.

3.1. TF-IDF

Se utilizan programas computacionales implementados en las siguientes librerías: numpy, pandas, nltk, heapq y re. Se inició con la tokenización del texto original por palabras. Se procede con la limpieza de los datos, se remueven los caracteres no alfabéticos y se reemplazan las mayúsculas por minúsculas, para ser almacenados en variables tipo texto. Luego de generar las condiciones de trabajo, se obtiene el puntaje TF-IDF, que consiste en encontrar la frecuencia de las palabras dentro del texto, para calcular la frecuencia ponderada de cada término y en consecuencia calcular puntuaciones de las oraciones a partir de la frecuencia de aparición de palabras. Finalmente, se genera el resumen a partir de las diecisiete oraciones con mejor puntaje (Christian et al., 2016).

3.2. Análisis de Grafos

Se utiliza las librerías numpy, pandas, nltk y re para resumir el texto a través de *text rank*. Se inicia con la limpieza de datos, removiendo las *stopwords*, corchetes y espacios extras, caracteres que no son letras y cambios de mayúsculas a minúsculas. Luego, se procede con la representación del vector de oración, el cual se crea utilizando el modelo *word2vec* previamente entrenado de Gensim, es decir, se componen las *word embeddings* (representación de palabras como vectores de números reales). Se conforman los grafos a partir de las oraciones, para ello se consideran las oraciones como los nodos y las aristas se ponderan a partir de la similitud del coseno entre las oraciones. Finalmente, se producen las oraciones utilizando el algoritmo *PageRank*, imprimiendo las cinco oraciones mejor clasificadas.

3.3. Gensim

Como ya se mencionó, este es un método de representación de textos, el que, a partir de la co-ocurrencia de oraciones, frases y palabras permite realizar resúmenes automáticos de manera no supervisada. Para su implementación se utiliza las librerías gensim, logging, numpy, pandas y re. Su implementación es más simple que los dos anteriores, los datos se leen en la fase inicial, se ingresan al método de resumen y se obtienen resultados variando los valores del parámetro *split*. Los parámetros de este método son:

- Text (str): Texto de entrada.
- Ratio (flotante, opcional): Número entre 0 y 1 que determina la proporción del número de frases del texto original que se elegirán para el resumen.
- Word_count (int o None, opcional): determina cuántas palabras contendrá la salida. Si se proporcionan ambos parámetros, la relación se ignorará.
- Split (*bool*, opcional): si es *True*, se devolverá la lista de oraciones. De lo contrario, las cadenas unidas se devolverán.

La salida de este algoritmo es un resumen que tiene como parámetro el porcentaje de palabras a incorporar en él, en nuestro caso utilizamos el 10 %.

3.4. Evaluación de los Resúmenes

Se utiliza una metodología de evaluación y recolección de datos sobre la base de dos formularios, utilizando la plataforma *Google Forms*. El primero se compone de tres secciones. En la primera se identifica el documento evaluado según el Cuatro 1 y se señala el tiempo de lectura.

TF-IDF	Gensim	Grafos
001-Twitter	011-Twitter	101-Twitter
002-Facebook	012-Facebook	102-Facebook
003-Instagram	013-Instagram	103-Instagram
004-Snapchat	014-Snapchat	104-Snapchat
005-TikTok	015-TikTok	105-TikTok

Cuadro 1: Codificación de los resúmenes automáticos.

En la siguiente sección se evalúan los resúmenes en una escala del 0 al 10 de legibilidad, la que se define como la fluidez, gramaticalidad y coherencia del resumen. Luego se evalúa la relevancia, es decir, si contiene información importante y destacada de la política de privacidad, siendo correcto al evitar información contradictoria, no relacionada y evita información repetida o redundante. Por último, se finaliza con una descripción de las fortalezas, debilidades, oportunidades y amenazas, incluyendo apreciaciones generales respecto de los resúmenes. La segunda encuesta, comienza de la misma manera y considera una sección en la que se identifica el contenido de las PP ideales, a partir de los factores de la Tabla 4, clasificándolos del 1 al 10 (no excluyentes), además de tener la posibilidad de incluir otros. Es importante destacar que el evaluador, experto en derecho, se enfrenta a los resúmenes sin saber qué método fue el utilizado, con el fin de disminuir el sesgo de la persona.

Los datos se analizan utilizando el Principio de Pareto, también conocido como la regla del 80–20, que plantea que, en diferentes fenómenos, una pequeña proporción de los elementos es la que explica la mayor parte del efecto. Lo anterior es complementado con el análisis FODA por método y por política. Bajo este análisis se busca identificar las fortalezas y debilidades de los métodos, así como las dificultades y oportunidades para el trabajo con textos de PP. A partir de lo anterior, se logra la sinergia entre los análisis cuantitativo y cualitativo. Finalmente, con el fin de encontrar relaciones entre las variable de relevancia, cantidad de palabras y tiempo de lectura, se utiliza la correlación de Pearson.

Se utilizó la herramienta *Google Colaboratory* y el lenguaje de programación *Python*, dada su fácil implementación desde cualquier ordenador con conexión a internet, mediante la plataforma de *Google Drive*. Los códigos computacionales de las librerías para la implementación de las técnicas se obtuvieron desde el repositorio *GitHub* [1].

4. Resultados

Se generó un corpus de cinco políticas de privacidad de las redes sociales en cuestión, que en promedio tienen 5.042,8 de palabras, tal como se presenta en la Figura 1, los que para ser procesados deben tener un formato `*.txt`.

Además, se implementan tres métodos de resumen automático: basado en Análisis de Grafos, TF-IDF y Gensim, lo que permite tener quince

[1] https://github.com/ShristiK/Text-Summarization

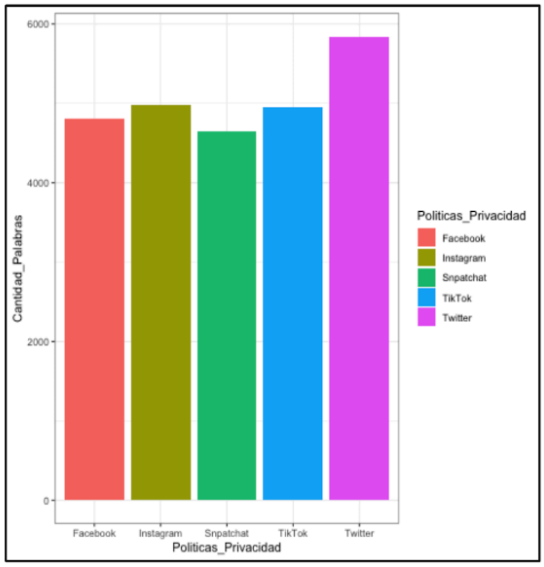

Figura 1: Cantidad de palabras de las 5 políticas de privacidad.

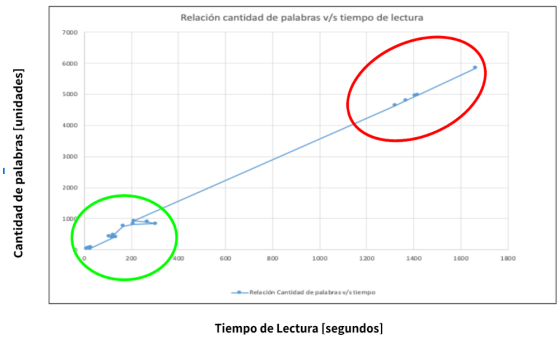

Figura 2: Relación entre cantidad de palabras y tiempo de lectura.

resúmenes distintos. Para el primero obtenemos en promedio una cantidad de 51,4 palabras y un tiempo de lectura de 21,4 segundos, para el segundo 420,6 palabras y 121,6 segundos respectivamente, y, por último, 845,6 palabras y 232 segundos, los que fueron evaluados de forma cualitativa con su complemento cuantitativo.

A partir del formulario respondido por el experto (abogado evaluador), se obtuvo una legibilidad de un 100 % para los quince resúmenes. Esto quiere decir, que en la lectura de los resúmenes no tuvo problemas respecto de la fluidez, gramaticalidad y coherencia de los mismos.

La Figura 2 presenta la relación entre la cantidad de palabras y el tiempo de lectura. Se observa una relación significativa alta, según el coeficiente de correlación de Pearson de $r = 0,955$ ($p < 0,001$), considerando las políticas de privacidad completa. Similar resultado se obtiene para la relación entre relevancia y la cantidad de palabras, $r = 0,878$ ($p < 0,001$), tal como se presenta en las Figuras 3 y 4 respectivamente.

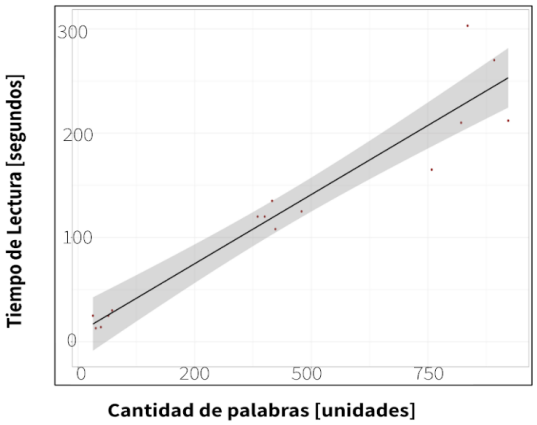

Figura 3: Correlación de Pearson: Cantidad de palabras (unidades) vs Tiempo de lectura (segundos) (promedios).

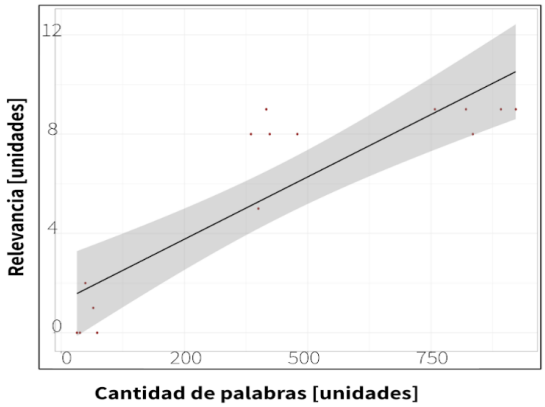

Figura 4: Correlación de Pearson: Cantidad de palabras (unidades) vs Relevancia (unidades) (promedios).

A continuación, se presentan las Figuras 5, 6, 7 y 8. En la primera se indica que las preguntas 4, 10 y 7 fueron las mejores evaluadas y, contrariamente, las preguntas 12, 14, 16, 17, 19 y 20 como las peor evaluadas.

La Figura 6 representa los mejores resúmenes, sin considerar el valor de la relevancia. Así el 015, 001 y el 011 se identifican como los mejor evaluados, por contraparte, los resúmenes 102, 104 y el 103 se identifican como los peor evaluados.

La Figura 7 nos indica qué resúmenes obtuvieron mejores y peores resultados solamente en términos de la relevancia, siendo los primeros el 003, 013, 014, 015, 011 y, por el otro lado, el 103, 105 y el 101.

Finalmente, la Figura 8 es la combinación de las Figuras 6 y 7, es decir, consideran las veinte preguntas con más relevancia en la que los con mejor calificación son el 015, 011 y 103. Para todos los gráficos el color verde representa el mejor 80 % y el color rojo el 20 % restante, según el análisis de Pareto.

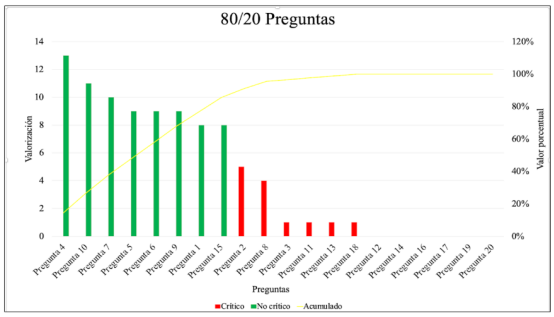

Figura 5: Análisis de Pareto de las 20 preguntas a partir de los resúmenes automáticos.

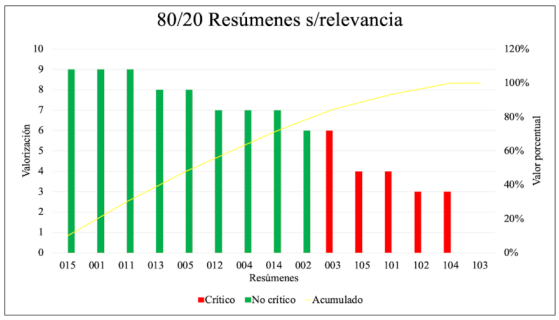

Figura 6: Análisis de Pareto de la Relevancia de las 20 preguntas, a partir de los resúmenes automáticos, sin considerar Relevancia.

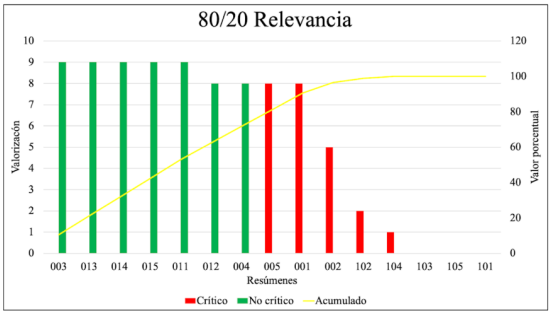

Figura 7: Análisis de Pareto de la Relevancia por resúmenes automáticos.

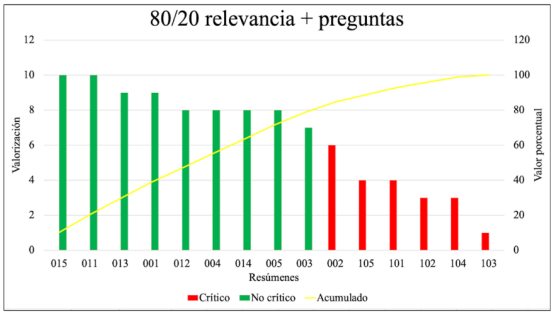

Figura 8: Análisis de Pareto de los resúmenes automáticos a partir de las 20 preguntas, más la Relevancia.

TF-IDF	Gensim	Grafos
001-Twitter	011-Twitter	101-Twitter
002-Facebook	012-Facebook	102-Facebook
003-Instagram	013-Instagram	103-Instagram
004-Snapchat	014-Snapchat	104-Snapchat
005-TikTok	015-TikTok	105-TikTok

Cuadro 2: Codificación de los resúmenes automáticos por métodos, categorizados a partir del Análisis de Pareto de la Figura 8.

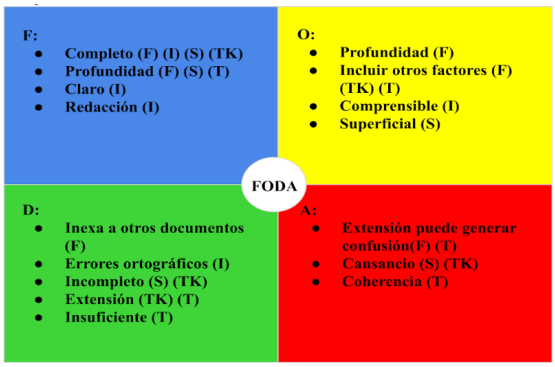

Figura 9: Análisis FODA del método Gensim de los resúmenes automáticos.

Todo lo anterior es resumido en el Cuadro 2, manteniendo la lógica de colores verde y rojo.

Tomando en cuenta el análisis FODA que desarrollamos para cada método y política de privacidad, los resultados son presentados en las Figuras 9, 10 y 11. Estos resulados están ordenados por método y cualidad, considerando cada política de privacidad, especificada entre paréntesis, a saber, **F**acebook, **T**witter, **T**ik**T**ok, **S**napchat e **I**nstagram. Para Gensim obtenemos que, de modo general, logra presentar información amplia de las veinte preguntas, es decir, es completo, abordando los temas con una profundidad tal que es identificada como aceptable para un resumen, además de ser claro y con buena redacción. No obstante, presenta algunas debilidades compartidas por las políticas de privacidad estudiadas, esto es, información insuficiente en el caso de Twitter o errores ortográficos. Además, se tiene en consideración que por su extensión, podrían requerir mayor esfuerzo cognitivo y sobrecarga de información al lector (Moy et al., 2018).

El método TF-IDF presenta características igual de aceptables, aunque muy segmentadas por tipo de empresa. Presenta menor extensión de los resúmenes y una redacción fácil de entender. Por lo mismo, no genera fatiga al lector, aunque si carecen de criterios o factores importantes. Asimismo, como principal amenaza no entrega información relevante para el lector meta.

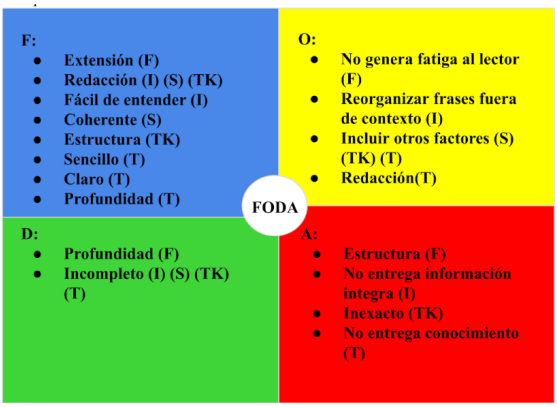

Figura 10: Análisis FODA del método TF-IDF de los resúmenes automáticos.

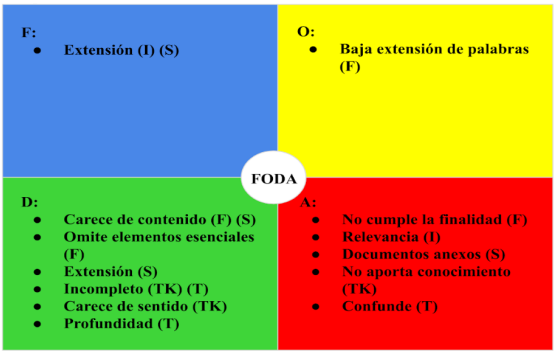

Figura 11: Análisis FODA del método basado en Grafos de los resúmenes automáticos.

Finalmente, el método grafo-analítico, tiene como principal fortaleza la extensión, pero no cumple la finalidad de informar o entregar conocimiento, carece de sentido y omite elementos esenciales. Cabe destacar, que a pesar de la poca cantidad de oraciones aún así permite generar resúmenes, aunque de menor calidad que con Gensim.

Luego de combinar el análisis Pareto con el FODA, obtenemos el Cuadro 3, el cual se presenta nuevamente con colores. Esta tabla agrega el amarillo a los posibles candidatos a ser un buen resumen, manteniendo el rojo para los peores evaluados y el verde para los mejores, que en este caso fueron el 014 y el 015, seguido levemente por el 012. A modo general, en verde encontramos resúmenes del método Gensim, en amarillo del método TF-IDF y en rojo el método grafo-analítico.

Respecto a los factores de riesgo de las PP, se obtiene un alto grado para todos los factores nombrados el Cuadro 4.

Como resultado concreto, a una velocidad de lectura de 3,5 palabras/segundo se logra reducir en un 91,29 % el tiempo de lectura general, un 83,84 % con el mejor modelo y se disminuyé la

TF-IDF	Gensim	Grafos
001-Twitter	011-Twitter	101-Twitter
002-Facebook	012-Facebook	102-Facebook
003-Instagram	013-Instagram	103-Instagram
004-Snapchat	014-Snapchat	104-Snapchat
005-TikTok	015-TikTok	105-TikTok

Cuadro 3: Codificación de los resúmenes automáticos por métodos, categorizados a partir del Análisis de Pareto de la Figura 8, sumado al Análisis FODA de los distintos métodos.

Datos bancarios	Finalidad/ Propósito para la recopilación de datos	Derecho al olvido
Datos sensibles	Seguro social	Editar o eliminar información personal del servicio
Información personal	Principios de protección de datos	Dirección de correo electrónico
Empresa proveedora del Servicio	Tratamiento de datos de menores de edad	Recopilación de datos
Consentimiento informado	Cambios en las Políticas de Privacidad	Violaciones de seguridad
Terceros involucrados		

Cuadro 4: Jerarquía de los factores de riesgo de las Políticas de Privacidad de aplicaciones infomáticas, identificando la mejor Política de Privacidad para Snapchat y Gensim.

cantidad de palabras en un 91,29 %, un 83,23 % con el mejor modelo, tal como muestra la Figura 12. A partir de lo anterior es posible proyectar que la utilización de resúmenes automáticos permitiría a los usuarios comprender mejor las PP que están aceptando, utilizando menos tiempo para su lectura.

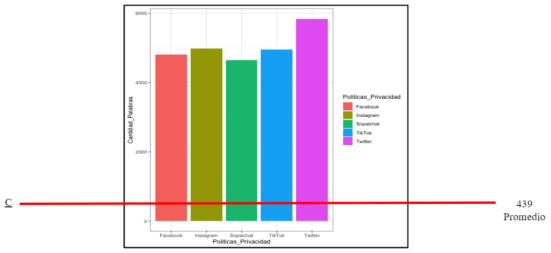

Figura 12: Cantidad de palabras de 5 Políticas de Privacidad: Facebook, Instagram, Snapchat, Twitter y TikTok, ya la curva $\bar{C}$ de cantidad de palabras promedio de los resúmenes automáticos.

5. Conclusiones y trabajo futuro

Como se aprecia en los resultados, los métodos evaluados tienen buenos índices en la construcción de los resúmenes, aunque TF-IDF, presenta una baja coherencia, dado su funcionamiento de repetición de palabras de contenido. Contrariamente, el método Gensim logra generar resúmenes más coherentes y con contenido más adecuado. Ello debido al procesamiento de textos con *Deep Learning*, lo que permite otros tipos de operaciones como el cálculo de similitud entre textos o palabras con mayor frecuencia.

Entre los resúmenes generados, el mejor evaluado es el 014 de Gensim, presentado en el Cuadro 5, relativo a las políticas de privacidad de Snapchat. No obstante, en el Cuadro 6 vemos que aún hay información que no es captada por los métodos de resumen. Este es el caso de la información relativa a los usuarios menores de 16 años, factor que se considera importante dentro de una PP. Además, se observa una correlación positiva alta entre la cantidad de palabras de un texto y su relevancia, así también entre la cantidad de palabras y el tiempo de lectura.

PREGUNTA 7: Snapchat

Podremos compartir la siguiente información con todos los Snapchatters, nuestros socios comerciales y el público en general: información pública, como tu nombre, nombre de usuario, Snapcódigo e imágenes de perfil; los envíos de Historia que se configuren para que todo el mundo pueda verlos y todo el contenido que envíes a un servicio inherentemente público tal como Nuestra Historia, y otros servicios de múltiples fuentes.

Cuadro 5: Extracto del resumen automático 014 a partir del método Gensim.

PREGUNTA 16: Snapchat

Niños

Nuestros servicios no están destinados a menores de 13 años y no los dirigimos a ellos. Ese es el motivo por el que no recabamos conscientemente información personal de ningún menor de 13 años. Por otra parte, podemos limitar el modo en que recabamos, usamos y guardamos parte de la información de los usuarios de la UE de entre 13 y 16 años. En algunos casos, esto significa que no podremos facilitar determinadas funciones a estos usuarios. Si tuviéramos que recurrir a un consentimiento como base jurídica para tratar esta información y tu país exige el consentimiento de uno de los padres, tal vez necesitemos el consentimiento de tu padre para recabar y utilizar dicha información.

Cuadro 6: Extracto de las políticas de privacidad de Snapchat.

El método TF-IDF sigue en la jerarquía de resultados, al ser un método de nivel intermedio, mientras que el de tipo grafo-analítico se queda en último lugar. No obstante, para trabajos futuros se planea potenciar el algoritmo de la Análisis de Grafos para ampliar y mejorar los contenidos de sus resúmenes.

Un aspecto interesante, en relación con los criterios de evaluación de la calidad de los resúmenes, es que no toda la información debe estar literalmente presentada en la política ni en el resumen. Así, por ejemplo, al evaluar Gensim (Cuadro 5) con la pregunta 7, se observa que esta no se responde literalmente y aun así se posiciona entre las tres mejor evaluadas. No obstante, de acuerdo con los expertos, esto tampoco es necesario. Ello porque al compartirse información con el "público en general" se hace evidente, utilizando la premisa del máximo-mínimo, que puede hacerse también con un destinatario específico como "la policía".

En suma, entre los beneficios de resumir las políticas de privacidad de las distintas aplicaciones informáticas con el método no supervisado Gensim, destacamos el sustancial ahorro de tiempo de lectura y una mejor comprensión por parte de los usuarios de los aspectos más relevantes declarados en las políticas de privacidad acerca del uso de sus datos privados.

En trabajos futuros se proyecta utilizar otros métodos de evaluación de resúmenes, como por ejemplo el paquete de medidas propuestas por (Lin, 2004), denominada ROUGE: Recall-Oriented Understudy for Gisting Evaluation, conformado por ROUGE-N, ROUGE-L, ROUGE-W y ROUGE-S, para evaluar la precisión (*precision*) y sensibilidad (*recall*) de los resúmenes automáticos respecto a resúmenes de referencia creados por humanos. No obstante, dado que no contamos con resúmenes de referencia creados por humanos o bien proporcionados por trabajos previos, el uso de esta métrica debe ser abordada independientemente con el fin de evaluar la calidad en términos computacionales. Así también, se proyecta explorar el uso de algoritmos de evaluación automatizada de resúmenes en los que no se requiere modelamiento por parte de escritores humanos (Louis & Nenkova, 2009; Saggion et al., 2010; Torres-Moreno et al., 2010).

Agradecimientos

Nuestros agradecimientos a la Pontificia Universidad Católica de Valparaíso por el financiamiento de esta investigación a través de los fondos Incentiva en el Aula (039.330/2020), 039.406/2021 y 039.344/2022. Así como al Núcleo de Procesamiento del Lenguaje Natural Aplicado (NIPLNA, https://www.niplna.com)

Referencias

Anastasopoulou, Kallia, Spyros Kokolakis & Panagiotis Andriotis. 2017. Privacy decision-making in the digital era: A game theoretic review. En *Human Aspects of Information Security, Privacy and Trust*, 589–603. Springer International Publishing. doi 10.1007/978-3-319-58460-7_41.

Barrios, Federico, Federico López, Luis Argerich & Rosa Wachenchauzer. 2016. Variations of the similarity function of textrank for automated summarization. ArXiv:1602.03606 [cs.CL].

Becerra, Xavier. 2020. California consumer privacy act (ccpa) fact sheet. https://oag.ca.gov/system/files/attachments/press_releases/CCPAFactSheet(00000002).pdf.

Bird, Steven, Ewan Klein & Edward Loper. 2009. *Natural language processing with Python*. O'Reilly.

Christian, Hans, Mikhael Pramodana Agus & Derwin Suhartono. 2016. Single document automatic text summarization using

term frequency-inverse document frequency (TF-IDF). *ComTech: Computer, Mathematics and Engineering Applications* 7(4). 285. 10.21512/comtech.v7i4.3746.

Corral, Hernán F. 2000. Configuración jurídica del derecho a la privacidad II: Concepto y delimitación. *Revista Chilena de Derecho* 27(2). 331–355.

Cucatto, Mariana. 2011. Algunas reflexiones sobre lenguaje jurídico como lenguaje de especialidad: más expresión que verdadera comunicación. *Revista virtual intercambios* 15. 1–19.

de la Cueva, Pablo Lucas M. & Alfonso Fernández-Miranda. 1990. *El derecho a la autodeterminación informativa: la protección de los datos personales frente al uso de la informática*. Tecnos.

da Cunha, Iria & M. Ángeles Escobar. 2021. Recomendaciones sobre lenguaje claro en español en el ámbito jurídico-administrativo: análisis y clasificación. *Pragmalingüística* 29. 129–148. 10.25267/Pragmalinguistica.2021.i29.07.

DataReportal. 2022. Digital 2022 global digital overview. https://datareportal.com/reports/digital-2022-global-overview-report.

Erkan, Günes & Dragomir R. Radev. 2004. LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research* 22(1). 457–479.

Gambhir, Mahak & Vishal Gupta. 2016. Recent automatic text summarization techniques: a survey. *Artificial Intelligence Review* 47(1). 1–66. 10.1007/s10462-016-9475-9.

García Calderón, Jesús. 2019. Una retórica para la igualdad. *Revista del Ministerio Fiscal* 8. 41–57.

Hernández, Ángel Javier Alonso. 2017. *Deep learning aplicado al resumen de textos*: Universidad Complutense de Madrid. Trabajo de Fin de Máster.

Karmaker, Pradip Chandra & Md. Sharif Hossen. 2019. Performance analysis of frequency and graph theoretic based text summarization. En *International Conference on Electrical, Computer and Communication Engineering (ECCE)*, 10.1109/ecace.2019.8679452.

Ley Chile. 2014. Ley núm. 19.628, sobre protección de la vida privada (Chile). Ministerio Secretaría General de la Presidencia, http://www.leychile.cl/Navegar?idNorma=141599.

Lierde, Hadrien Van & Tommy W. S. Chow. 2019. Learning with fuzzy hypergraphs: A topical approach to query-oriented text summarization. *Information Sciences* 496. 212–224. 10.1016/j.ins.2019.05.020.

Lin, Chin-Yew. 2004. ROUGE: A package for automatic evaluation of summaries. En *Text Summarization Branches Out*, 74–81.

Louis, A. & A. Nenkova. 2009. Automatically evaluating content selection in summarization without human models. *Conference on Empirical Methods in Natural Language Processing* 306–314.

Luhn, Hans Peter. 1958. The automatic creation of literature abstracts. *IBM Journal of Research and Development* 2(2). 159–165. 10.1147/rd.22.0159.

van de Luijtgaarden, Nick. 2019. *Automatic summarization of legal text*: Utrecht University. Trabajo de Fin de Máster.

Maybury, Mark T. 1995. Generating summaries from event data. *Information Processing & Management* 31(5). 735–751. 10.1016/0306-4573(95)00025-c.

McDonald, Aleecia M. & Lorrie Faith Cranor. 2009. The cost of reading privacy policies. *I/S: A Journal of Law and Policy for the Information Society* 4(3). 543–568.

Meier, Yannic, Johanna. Schäwel & Nicole C. Krämer. 2020. The shorter the better? effects of privacy policy length on online privacy decision-making. *Media and Communication* 8(2). 291–301.

Meza, Paulina, & Felipe González Catalán and. 2021. Un instrumento para evaluar la calidad lingüístico-discursiva de textos disciplinares producidos por estudiantes de derecho. *Onomázein Revista de lingüística filología y traducción* 51. 164–184. 10.7764/onomazein.51.08.

Montolío, Estrella & Anna López Samaniego. 2008. La escritura en el quehacer judicial: Estado de la cuestión y presentación de la propuesta aplicada en la escuela judicial de españa. *Revista Signos* 41(66). 33–64.

Montolío, Estrella & Mario Tascón. 2020. *El derecho a entender: la comunicación clara, la mejor defensa de la ciudadanía*. Catarata.

Moy, Naomi, Ho Fai Chan & Benno Torgler. 2018. How much is too much? the effects of information quantity on crowdfunding performance. *PLOS ONE* 13(3). e0192012. 10.1371/journal.pone.0192012.

OCDE. 2002. Resumen directrices de la OCDE sobre protección de la privacidad y flujos transfronterizos de datos personales. overview. https://www.oecd.org/sti/ieconomy/15590267.pdf.

Poblete, Claudia Andrea & Pablo Fuenzalida González. 2018. Una mirada al uso de lenguaje claro en el ámbito judicial latinoamericano. *Revista de Llengua i Dret* 69. 119–138. 10.2436/rld.i69.2018.3051.

Pérez Luño, Antonio-Enrique. 1984. *Derechos humanos, estado de derecho y constitución*. Tecnos.

RAE, Real Academia Española. 2021. Diccionario de la lengua española: privacidad. https://dle.rae.es/privacidad?m=form.

Rane, Neha & Sharvari Govilkar. 2019. Recent trends in deep learning based abstractive text summarization. *International Journal of Recent Technology and Engineering (IJRTE)* 8(3). 3108–3115. 10.35940/ijrte.c4996.098319.

Řehůřek, Radim. 2021. Gensim: Topic modelling for humans. https://radimrehurek.com/gensim/.

Rodrigo, Alfaro. & Héctor Allende. 2020. Clasificación de textos multi-etiquetados con modelo bernoulli multi-variado y representación dependiente de la etiqueta. *Revista signos* 53(104). 549–567. 10.4067/s0718-09342020000300549.

Ruidías, Héctor J., Karina Eckert, Juan M. Lezcano & Carolina V. Rosas. 2018. Tecnologías de la web semántica aplicadas al tratamiento de documentos jurídicos electrónicos. En *XX Workshop de Investigadores en Ciencias de la Computación (WICC)*, 290–294.

Saggion, Horacio, Juan-Manuel Torres-Moreno, Iria da Cunha, Eric SanJuan & Patricia Velázquez-Morales. 2010. Multilingual summarization evaluation without human models. En *23rd International Conference on Computational Linguistics (COLING)*, 1059–1067.

Saldaña, Maria Nieves. 2012. «the right to privacy»: la génesis de la protección de la privacidad en el sistema constitucional norteamericano, el centenario legado de warren y brandeis. *Revista de Derecho Político* 85. 195–239. 10.5944/rdp.85.2012.10723.

Salton, Gerard & Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing & Management* 24(5). 513–523. 10.1016/0306-4573(88)90021-0.

Schwabe, Jürgen. 2009. *Jurisprudencia del tribunal constitucional federal alemán*. Konrad Adenauer Stiftung.

TCCh. 1989. Sentencia nº rol 65 de tribunal constitucional, 16 de enero de 1989. Tribunal Constitucional de Chile, https://vlex.cl/vid/-58943056.

Torres-Moreno, Juan-Manuel, Horacio Saggion, Iria da Cunha, Eric SanJuan & Patricia Velazquez-Morales. 2010. Summary evaluation with and without references. *Polibits* 42. 13–19.

Valenzuela, Miguel, Stephanie Riff, Rodrigo Alfaro, Alan Bronfman & René Venegas. 2020. Análisis y resumen automático de políticas de privacidad. En *III Congreso Internacional de Lingüística Computacional y de Corpus (CILCC) y V Workshop en Procesamiento Automatizado de Textos y Corpus (WoPATeC)*, 286.

Wan, Changlin, Muhan Zhang, Wei Hao, Sha Cao, Pan Li & Chi Zhang. 2021. Principled hyperedge prediction with structural spectral features and neural networks. ArXiv:2106.04292 [cs.SI].

Warren, Samuel D. & Louis Brandeis. 1890. The right to privacy. *Harvard Law Review* 4(5). 193–219.

Widyassari, Adhika Pramita, Supriadi Rustad, Guruh Fajar Shidik, Edi Noersasongko, Abdul Syukur, Affandy Affandy & De Rosal Ignatius Moses Setiadi. 2022. Review of automatic text summarization techniques & methods. *Journal of King Saud University - Computer and Information Sciences* 34(4). 1029–1046. 10.1016/j.jksuci.2020.05.006.

Zaeem, Razieh Nokhbeh, Safa Anya, Alex Issa, Jake Nimergood, Isabelle Rogers, Vinay Shah, Ayush Srivastava & K. Suzanne Barber. 2020. PrivacyCheck v2: A tool that recaps privacy policies for you. En *29th ACM International Conference on Information & Knowledge Management*, 10.1145/3340531.3417469.

Zaeem, Razieh Nokhbeh, Rachel L. German & K. Suzanne Barber. 2018. PrivacyCheck: Automatic summarization of privacy policies using data mining. *ACM Transactions on Internet Technology* 18(4). 1–18. 10.1145/3127519.

Zhang, Caiming & Yang Lu. 2021. Study on artificial intelligence: The state of the art and future prospects. *Journal of Industrial Information Integration* 23. 100224. 10.1016/j.jii.2021.100224.