ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE INGENIERÍA ELECTRÍCA Y ELECTRÓNICA

RECOLECCIÓN AUTOMÁTICA DE POLÍTICAS DE PRIVACIDAD EN ECUADOR

TRABAJO DE INTEGRACIÓN CURRICULAR PRESENTADO COMO REQUISITO PARA LA OBTENCIÓN DEL TÍTULO DE INGENIERO EN TECNOLOGÍAS DE LA INFORMACIÓN

JOSÉ ANDRÉS CAÑAR ROMERO

jose.canar@epn.edu.ec

DIRECTOR: ANA FERNANDA RODRIGUEZ HOYOS

ana.rodriguez@epn.edu.ec

DMQ, abril 2023

CERTIFICACIONES

Yo, JOSÉ ANDRÉS CAÑAR ROMERO declaro que el trabajo de integración curricular aquí descrito es de mi autoría; que no ha sido previamente presentado para ningún grado o calificación profesional; y, que he consultado las referencias bibliográficas que se incluyen en este documento.

JOSÉ ANDRÉS CAÑAR ROMERO

Certifico que el presente trabajo de integración curricular fue desarrollado por JOSÉ ANDRÉS CAÑAR ROMERO, bajo mi supervisión.

ANA FERNANDA RODRIGUEZ HOYOS
DIRECTORA

DECLARACIÓN DE AUTORÍA

A través de la presente declaración, afirmamos que el trabajo de integración curricular aquí descrito, así como el (los) producto(s) resultante(s) del mismo, son públicos y estarán a disposición de la comunidad a través del repositorio institucional de la Escuela Politécnica Nacional; sin embargo, la titularidad de los derechos patrimoniales nos corresponde a los autores que hemos contribuido en el desarrollo del presente trabajo; observando para el efecto las disposiciones establecidas por el órgano competente en propiedad intelectual, la normativa interna y demás normas.

JOSÉ ANDRÉS CAÑAR ROMERO

ANA FERNANDA RODRIGUEZ HOYOS

DEDICATORIA

A Dayana Maribel, mi enamorada, por ser una persona sumamente valiosa en mi vida durante estos últimos años. Su apoyo incondicional ha sido constante y ha sido mi fiel compañera en los momentos difíciles. Le doy las gracias por estar a mi lado siempre, brindándome sus palabras de aliento y confiando en mí incluso en las situaciones más complicados. Espero que sigamos caminando juntos hacia el futuro, apoyándonos mutuamente en cada paso que demos para cumplir nuestros sueños y objetivos.

Su amor y cariño son un regalo que valoro enormemente, y me siento profundamente agradecido por tenerla en mi vida.

AGRADECIMIENTO

Quiero expresar mi agradecimiento a Dios por su amor, guía y protección constante en mi vida.

Asimismo, deseo agradecer de todo corazón a mis padres, José Luis Cañar Farinango y Juana Mercedes Romero Suarez, por su apoyo incondicional a lo largo de mi formación personal y académica. Les agradezco por su sacrificio, dedicación y esfuerzo. Agradezco que siempre se preocupen por mí y por jamás dejarme solo. Este logro no habría sido posible sin su ayuda y siempre estaré agradecido por haberme brindado la mejor herencia que un hijo puede tener: la educación. Este logro no es solo mío, sino de ellos también y quiero decirles con orgullo: ¡Papi, mami, lo logramos!

También quiero expresar mi agradecimiento a mis hermanos Erick, Micaela y Daniel, quienes me brindaron motivación constante en cada paso de este camino. Ellos me ayudaron a no rendirme y a demostrarles que, con esfuerzo y con la guía de nuestros padres, todo es posible.

A mis profesores, por compartir sus conocimientos y experiencias valiosas en mi educación. Les agradezco por su dedicación, paciencia y motivación constante.

A mis amigos en general por su apoyo en cada semestre.

A Hernán Mejía por brindarme la oportunidad de comenzar mi vida profesional en Secuoia IT. Gracias a su apoyo, pude aplicar y desarrollar los conocimientos adquiridos durante mi formación académica. Estoy agradecido por haber tenido mi primer trabajo profesionalmente en esta empresa y por todo lo que he aprendido durante mi tiempo aquí.

Por último, pero no menos importante, quiero expresar mi agradecimiento al PhD. José Estrada Jiménez y a la PhD. Ana Fernanda Rodriguez, quienes me brindaron su orientación y conocimientos en el desarrollo de este trabajo. Estoy muy agradecido por el tiempo, dedicación y esfuerzo que dedicaron a mi proyecto.

ÍNDICE DE CONTENIDO

RESUMEN

La privacidad se estudia mucho últimamente debido a la gran cantidad de información que los usuarios comparten en línea y a los graves escándalos sobre el mal uso de datos personales. Medirla es necesario para poder aplicar correctivos, y estudiar las políticas de privacidad es una forma de hacerlo.

Este trabajo se orienta a la recolección automática de políticas de privacidad, que facilite hacerlo fácilmente y de forma periódica. Para ello, se recolectó y analizó manualmente las políticas de sitios ecuatorianos para determinar criterios que permitan automatizar la identificación y descarga de estos documentos.

Se construyó un script sencillo que, usando esos criterios y basado en las palabras clave incluidas en los enlaces a las políticas, identifica, descarga, y obtiene automáticamente el texto de dichas políticas. Para identificar los falsos positivos, se construyó un modelo predictivo para clasificar un documento como política o no política.

Los resultados muestran que el script ofrece una precisión interesante (80%) al identificar políticas de privacidad automáticamente, incluso cuando se evalúa en políticas de privacidad de México. El modelo predictivo igualmente permite identificar políticas de privacidad en Ecuador con una accuracy de 80%, lo que permite descartar una buena porción de los falsos positivos.

PALABRAS CLAVE: privacidad, políticas, datos personales, protección de datos, seguridad, Ecuador.

ABSTRACT

Privacy is being studied a lot lately due to the large amount of information that users share online and serious scandals about the misuse of personal data. Measuring it is necessary to be able to apply corrective measures, and studying privacy policies is one way to do it.

This work is oriented towards the automatic collection of privacy policies, which facilitates doing so easily and periodically. For this, the policies of Ecuadorian sites were manually collected and analyzed to determine criteria that allow the identification and download of these documents to be automated.

A simple script was built that, using these criteria, and based on the keywords included in the links to the policies, automatically identifies, downloads, and obtains the text of said policies. To identify false positives, a predictive model was built to classify a document as political or non-political.

The results show that the script offers interesting accuracy (80%) in automatically identifying privacy policies, even when evaluated against Mexico privacy policies. The predictive model also makes it possible to identify privacy policies in Ecuador with an accuracy of 80%, which allows us to rule out a good portion of the false positives.

KEYWORDS: privacy, policies, personal data, data protection, security, Ecuador.

1. INTRODUCCIÓN

Actualmente, la interacción de las personas desde sus dispositivos móviles, con las redes sociales y, en general, con Internet, produce una gran cantidad datos, muchos de ellos son datos personales, que son entregados a terceros [1] [2]. En Ecuador, por ejemplo, gracias a la reducción de costos del servicio de Internet [3], existen más de 13 millones de dispositivos móviles conectados a la red, asociados a más de 10 millones de usuarios, que utilizan 14 millones de cuentas en redes sociales [2].

La vasta información que generan estas interacciones es usualmente recolectada y procesada por distintas entidades. Aunque la disponibilidad de tanta información tiene muchos beneficios [2] [4], esto también genera preocupaciones relacionadas con la privacidad de los sujetos de dicha información. Particularmente, las entidades de terceros que tienen acceso a estos datos adquieren mucho poder, tanto que podría utilizarse para manipular o, incluso, atacar a los usuarios involucrados [4].

En esa medida, la privacidad tiene una relevancia tal que ha sido reconocida por las Naciones Unidas como un derecho fundamental. De hecho, se plantea a la privacidad como un pilar fundamental para el desarrollo de la libertad de los individuos [5].

Para hacer efectivo este derecho, recientemente, en la legislación local ecuatoriana, se ha promulgado una regulación que garantiza a los usuarios derechos relacionados con su privacidad y la protección de sus datos. Esta ley es la Ley Orgánica de Protección de Datos Personales (LOPDP) [6]

Uno de los derechos que garantiza esta ley a los usuarios es el de estar informados con respecto al uso que se les da a sus datos. Una forma de garantizar esto es mediante una política de privacidad, que publican ciertos sitios web. desde hace tiempo, como mecanismo informativo.

Así, este documento puede ser un insumo muy útil para comprender cómo las entidades a cargo de datos de usuario interactúan con esta información. Un análisis a mayor escala de este recurso podría incluso ayudarnos a tener un diagnóstico de la situación de la protección de datos personales en un escenario regional o incluso mundial. Sin embargo, como estos documentos por lo general están distribuidos en varios sitios, encontrar y descargar estas políticas de privacidad de forma manual resultaría un reto muy grande [7], más aún si se desea realizar un análisis escalable en el país o en la región. Así, un mecanismo que realice este trabajo de recolección automáticamente sería de mucha utilidad.

A partir de esta problemática, se propone implementar un mecanismo de recolección automática de políticas de privacidad en Ecuador que permita la obtención escalable en el tiempo de dichas políticas. Este mecanismo se construirá y evaluará a partir de un conjunto de sitios web populares en el país, buscando obtener un modelo de aprendizaje automático para clasificar una política de privacidad. Este mecanismo podría ayudar a que estas políticas de privacidad sean accesibles masivamente para investigadores o periodistas, que busquen entender, por ejemplo, el impacto de la regulación en el estado de la privacidad en el país [7].

1.1 Objetivo general

Recolectar automáticamente políticas de privacidad en Ecuador.

1.2 Objetivos específicos

1. Recolectar información de sitios web ecuatorianos.
2. Detectar los patrones de publicación de políticas de privacidad en línea en Ecuador.
3. Desarrollar scripts para identificar y descargar una política de privacidad en Ecuador.
4. Evaluar el proceso de identificación de políticas de privacidad.

1.3 Alcance

El trabajo de titulación propuesto se enfoca en la recolección de políticas de privacidad en Ecuador. Con ese fin, se seleccionarán inicialmente 150 sitios web populares o que por su naturaleza podrían albergar información personal sensible. Para obtener los sitios populares en el país, se utilizará el servicio Alexa Top Sites de Amazon.

Luego, se explorará manualmente cada uno de estos sitios, recolectando la información de al menos 150 sitios web ecuatorianos, incluyendo, si fuese el caso, la política de privacidad publicada. En el proceso de exploración, se identificarán patrones de publicación de políticas, que podrían incluir: visibilidad de la política en la página principal, formato de publicación, errores de despliegue de la política, etc., que luego puedan servir para la recolección automatizada.

Con base en la información recolectada previamente, se programará un script básico en Python para la recolección automatizada de estas políticas de privacidad, teniendo como entrada las URLs de los sitios web, y se identificarán los falsos positivos en el caso de que existiesen.

Para la implementación de la navegación web automática se utilizará el *framework OpenWPM*, y para el procesamiento de la información se utilizarán las librerías disponibles de Python para ello. Para la detección automatizada de los documentos descargados que no sean políticas de privacidad (falsos positivos), se utilizará el conjunto de datos recolectado (documentos de políticas y no políticas) para construir un modelo de aprendizaje automático que permita clasificar un documento como política de privacidad o no política de privacidad, utilizado procesamiento de lenguaje natural.

Finalmente, se evaluará la eficacia del script de recolección de políticas y la exactitud (*accuracy*) del modelo obtenido.

Este trabajo no tendrá un producto final demostrable.

1.4 Marco teórico

En esta sección se empezará por explicar algunas definiciones relevantes para comprender el trabajo que se ha desarrollado. A partir de estas definiciones, se examinará muy brevemente el marco legal que sustenta el derecho a la privacidad en el país. Después se definirá la política de privacidad, lo que permitirá tener una visión más completa de su relevancia e importancia. Por último, se llevará a cabo una revisión teórica del procesamiento de datos, en particular de las técnicas y herramientas que se han utilizado en el trabajo actual.

1.4.1 Definiciones y marco legal

En este trabajo se presentan conceptos importantes que se necesitan para entender el contenido.

1.4.1.1 Definiciones

La privacidad según la RAE [8] se refiere a la capacidad de una persona para proteger su información privada y evitar que se divulgue sin su consentimiento.

Los datos personales según la Comisión Europea [9], son cualquier información que pertenezca a una persona y que permita su identificación total o parcial. Algunos ejemplos de estos datos son nombres, apellidos, domicilio, número de identificación nacional, dirección IP, entre otros.

La política de privacidad es un documento legal que establece las reglas y medidas que una empresa utiliza para manejar y proteger la información proporcionada por sus clientes o usuarios [10].

1.4.1.2 Derecho a la privacidad

La privacidad es un derecho humano fundamental que se refiere al control que las personas tienen sobre su información personal. En la era digital, la privacidad se ha vuelto cada vez más importante, ya que la información personal puede ser fácilmente recopilada, almacenada y compartida en línea. En este sentido, es crucial analizar las garantías de privacidad en el Ecuador, para entender cómo se protege este derecho y qué medidas se están implementando para asegurar la privacidad de los ciudadanos.

El artículo 12 de la Carta de los Derechos Humanos asegura que nadie debe sufrir injerencias arbitrarias en su vida privada, familiar, hogar o correspondencia, ni ser objeto de ataques a su reputación o honor [11]. Es un derecho fundamental que protege a las personas de cualquier amenaza o interferencia injustificada.

La privacidad es un derecho fundamental que debe ser protegido y preservado en todas las áreas de la vida, incluyendo el manejo de datos personales [12]. La Ley Orgánica de Protección de Datos Personales (LOPDP) reconoce la importancia de la privacidad y establece principios para su protección en diversas áreas, como la libertad de expresión, la seguridad nacional y la defensa del Estado. Sin embargo, también reconoce que, en ciertas situaciones, como las solicitudes y órdenes emitidas por autoridades administrativas o judiciales, la protección de la privacidad puede estar sujeta a otras normativas específicas. En cualquier caso, la protección de la privacidad debe ser un objetivo primordial en todas las situaciones donde se manejen datos personales [13].

El artículo 12 del capítulo III de la Ley de Protección de Datos Personales establece que, en base a los principios de lealtad y transparencia, cualquier persona que posea datos personales tiene derecho a conocer diversos aspectos relacionados con el tratamiento de sus datos [13].

Estos aspectos incluyen los fines, la base legal y los tipos de tratamiento, el tiempo de conservación, la existencia de una base de datos que contenga sus datos personales, el origen de los datos personales, la identidad y los datos del responsable del tratamiento de los datos personales proporcionados.

Además, la persona también tiene derecho a conocer las transferencias nacionales e internacionales de datos personales, los destinatarios y la garantía de protección de estos.

La publicación de la política de privacidad como una herramienta para hacer efectivo ese derecho es fundamental para que la persona pueda ejercer sus derechos de acceso,

eliminación, rectificación y actualización, oposición, anulación, limitación del tratamiento y a no ser objeto de una decisión basada únicamente en valoraciones automatizadas.

Además, la política de privacidad debe contemplar la posibilidad de solicitar la portabilidad de los datos y presentar reclamaciones ante el responsable y la autoridad competente. Por otro lado, también es importante que la política de privacidad informe sobre la existencia de decisiones automatizadas, incluyendo la creación de perfiles, para que la persona pueda tomar decisiones informadas sobre su uso de los servicios.

1.4.2 Medición de la privacidad

La creciente aplicación de sanciones por infracciones de privacidad está motivando el desarrollo de un enfoque para valorar tanto la relevancia de la información como la salvaguarda de los datos antes de su publicación [14]. Por esta razón, en esta sección se abordarán temas relacionados con la política de privacidad como un medio de evaluación.

1.4.2.1 Importancia de la medición de la privacidad

En la actualidad, es crucial abordar la medición de la privacidad, sobre todo en un mundo en el que la tecnología se encuentra en constante cambio y se han desarrollado múltiples herramientas para recolectar y procesar información personal [14]. Sin una métrica, es imposible evaluar el estado actual de la privacidad, o el impacto de su vulneración, y más difícil aún es proponer mecanismos de mejora.

1.4.2.2 La política de privacidad como instrumento de medición

En la sección anterior se discutió la importancia de medir la privacidad. En esta sección, se explicará por qué la política de privacidad se ha convertido en un instrumento clave para lograr esta medición.

La política de privacidad establece las reglas y condiciones para la recopilación, uso y protección de la información personal [13]. Por lo tanto, su existencia y calidad pueden ser indicadores de cumplimiento de las garantías de los usuarios según la legislación vigente. Además, el monitoreo de estos indicadores a lo largo del tiempo puede servir para medir la evolución de la protección de privacidad de acuerdo con ciertos factores, por ejemplo, del escenario nacional.

Además, una política de privacidad de calidad contiene mucha información sobre el tratamiento de datos personales, cuyo análisis en una muestra amplia de sitios web permitiría obtener conclusiones sobre el nivel de cumplimiento de varias garantías relacionadas con la privacidad de los ciudadanos.

La política de privacidad también muestra el compromiso de una empresa u organización con la protección de los datos personales y es una forma de medir su transparencia y confianza.

La política de privacidad se ha convertido en un elemento esencial en la protección de la privacidad en línea. Esta política establece las normas para el tratamiento de la información personal y es importante para la protección de los derechos de privacidad de los individuos [15].

1.4.3 Recolección de información de la Web

Se refiere a un método que recopila información y datos de una página web, ya sea manual o automáticamente (*web scraping*), usando diversas fuentes como motores de búsqueda, redes sociales, sitios gubernamentales y de noticias [16]. En esta sección se explicarán los principios teóricos del procesamiento de datos utilizados en este trabajo para recolectar políticas de privacidad.

1.4.3.1 Web Scraping

El término *web scraping* se refiere a la automatización de la extracción y almacenamiento de datos de páginas web, correos electrónicos, enlaces, bases de datos y documentos adjuntos, entre otros. Este proceso se lleva a cabo mediante el uso de software que analizan el código HTML de las páginas web y extraen la información necesaria.

La actividad de *web scraping* implica la determinación de qué información se debe almacenar y la navegación a través de diferentes enlaces incrustados en el sitio web para encontrar la información deseada. En el presente trabajo, se utilizó este método para obtener las políticas de privacidad de sitios web en Ecuador de manera automatizada.

1.4.3.2 OpenWPM para la recolección de información

OpenWPM es una herramienta de código abierto que automatiza la recolección de información y la evaluación de la privacidad en la web. Para recopilar información detallada sobre el comportamiento de los sitios web, utiliza un navegador web automatizado y un motor de JavaScript para interactuar con ellos [17].

Sin embargo, es importante destacar que la configuración de los navegadores utilizados es específica para cada uno de ellos y puede ser personalizada por los usuarios. Para facilitar la lectura de la configuración por defecto [18].

OpenWPM es una herramienta que ofrece diversos módulos y clases para definir secuencias de comandos, obtener el código fuente de una página web y almacenar datos en una base de datos. En particular, para la configuración, se utiliza el módulo

`openwpm.config` y las clases `BrowserParams` y `ManagerParams` que permiten configurar el motor de navegación web y el administrador de tareas de OpenWPM, respectivamente.

Por otro lado, para definir secuencias de comandos que se ejecutarán en un motor de navegación web con *OpenWPM* y obtener el código fuente de una página web, se emplea el módulo `openwpm.command_sequence` y las clases `CommandSequence` y `DumpPageSourceCommand`.

Asimismo, el módulo `openwpm.commands.browser_commands` y la clase `GetCommand` permiten obtener el contenido de una URL en un motor de navegación web con *OpenWPM*. En conjunto, estos módulos y clases hacen posible la automatización de tareas de navegación web y el análisis de datos de sitios web. [18]

1.4.4 Procesamiento de texto

El procesamiento de texto es una técnica que se utiliza para manipular, analizar y transformar datos escritos en texto [19], lo que puede ayudar a automatizar tareas de procesamiento de lenguaje natural y a comprender mejor el significado y las características del texto [20].

El procesamiento de texto se compone de varias etapas que se describen a continuación.

Limpieza de datos. Una parte fundamental del procesamiento de texto es la limpieza de datos, que consiste en el preprocesamiento de los textos para eliminar información innecesaria o redundante, corregir errores ortográficos y gramaticales, y normalizar el formato de los textos. Esto implica la eliminación de caracteres no deseados, como signos de puntuación o caracteres especiales, así como la eliminación de espacios en blanco adicionales. También puede incluir la conversión de letras mayúsculas a minúsculas. La limpieza de datos es esencial para mejorar la calidad de los datos y evitar problemas en el análisis posterior, ya que los datos no limpios pueden llevar a conclusiones incorrectas o sesgadas.

Tokenización. Se refiere al proceso de dividir el texto en unidades (palabras o frases) más pequeñas, llamadas tokens [21]. Es posible también etiquetar partes del texto para determinar la influencia de cada palabra en una oración.

Se suelen aplicar también otras técnicas como la **eliminación de *stop words***, que se refiere a la eliminación de palabras comunes que no aportan mucho valor semántico al texto, como "el", "la", "de", etc. [21]; o la **fusión de las palabras en un solo texto** con el objetivo de obtener un texto limpio y bien estructurado que esté listo para su análisis posterior.

Creación de modelos de aprendizaje. El procesamiento de texto también puede implicar la creación de modelos de lenguaje, como los modelos de traducción automática o los modelos de reconocimiento de texto, o los de clasificación (como los de detección de spam). Estos modelos utilizan algoritmos y técnicas de aprendizaje automático para analizar el texto y realizar tareas específicas, como la traducción de un idioma a otro o la identificación de temas y patrones en el texto.

A continuación, en la tabla 1, se describen muy brevemente algunas de las herramientas relacionadas con el procesamiento de texto, especialmente algunas librerías de Python [22].

Tabla 1. *Herramientas relacionadas*

Nombre de la herramienta	Descripción
Librería `re`	librería que trabaja con expresiones regulares.
Librería `tempfile`	permite crear y manejar archivos temporales.
Librería `pathlib` y clase `Path`	proporciona una forma fácil de trabajar con rutas de archivos y directorios.
Modulo `bs4` y clase `BeautifulSoup`	permite analizar y extraer datos de HTML.
Librería `pandas`	proporciona herramientas para manipular y analizar datos.
Librería `glob`	utilizada para buscar rutas de archivo que coincidan con un patrón especificado.
Librería `sqlite3`	proporciona una API de base de datos relacional de alto nivel.
Librería `PyPDF2`	se utiliza para trabajar con archivos PDF. Permite leer, escribir y modificar archivos PDF.
Módulo `sklearn.feature_extraction.text` y clase `CountVectorizer`	se utiliza para convertir un conjunto de documentos de texto en una matriz de características numéricas.
Módulo `sklearn.metrics` y clase `accuracy_score`	se utiliza para calcular la precisión de un modelo de aprendizaje automático.
Módulo `sklearn.naive_bayes` y clase `MultinomialNB`	implementa el clasificador Naive Bayes para datos multinomiales.

Módulo `sklearn.model_selection` y clase `train_test_split`	se utiliza para dividir un conjunto de datos en conjuntos de entrenamiento y prueba para el aprendizaje automático.
Librería `nltk`	se utiliza para procesar texto en lenguaje natural.
Módulo `nltk.corpus` y clase `stopwords`	se utilizan para filtrar el texto durante el procesamiento del lenguaje

2. METODOLOGÍA

En esta sección se expondrá el desarrollo de *scripts* utilizados para recolectar y analizar las políticas de privacidad en Ecuador. Estos *scripts* se basaron en el uso de herramientas automatizadas que permitieron extraer y procesar la información contenida en los textos de las políticas de privacidad de diversas entidades.

Con el objetivo de llevar a cabo esta tarea, se realizaron diversas pruebas y experimentos para evaluar la efectividad de las herramientas de extracción de información

2.1 Recolección manual de políticas de privacidad

A continuación, se detalla el proceso de recolección manual de políticas de privacidad realizado para este trabajo.

2.1.1 Criterios de selección de sitios web

Para escoger los sitios web donde luego se buscaron las correspondientes políticas de privacidad, se usaron varios criterios. En general, se incluyeron sitios web de gran alcance (populares), pero también que por su naturaleza podrían recolectar información sensible. Es decir, aquellos sitios en los que el impacto de su manejo de la privacidad de los usuarios puede ser importante. Más específicamente, los criterios de selección de sitios web fueron los siguientes:

- que sean sitios web ecuatorianos;
- que sean populares, o que pudiesen recolectar información sensible (salud, con aspectos financieros, o asociada a menores de edad); o,
- que sean sitios web relevantes para el país (estatales, gubernamentales).

2.1.2 Proceso de selección de los sitios web

Para seleccionar los sitios web ecuatorianos más populares se usó el servicio Alexa Top Sites de Amazon [23]. La lista de Alexa también incluye sitios populares en Ecuador que son internacionales, que no fueron considerados en nuestro estudio. Entre los sitios locales, se incluyeron sitios de deportes, de periódicos, de entretenimiento, etc.

Una vez obtenidos solo los sitios web ecuatorianos más populares, se agregaron a la lista sitios web que pudieran recolectar información sensible. Por ejemplo, se incluyeron sitios de hospitales, bancos, universidades, etc.

Entre los sitios relevantes para el contexto nacional se incluyeron sitios de Ministerios, de GADs, empresas públicas, etc.

Cabe destacar que la información recolectada en el trabajo de análisis automatizado de políticas de privacidad de [15] sirvió como base para este proceso de recolección. El análisis en el presente trabajo se realizó con más del doble de sitios web.

2.1.3 Proceso de identificación de políticas de privacidad

Una vez seleccionados los sitios web, se identificó de forma manual, en cada uno de ellos, la política de privacidad de la entidad. Para ello, se visitó cada uno de esos sitios usando un navegador web.

Durante la visita manual, se buscó un enlace a la política de privacidad en la página principal del sitio. Si ésta no era visible, se buscó esa información en el código fuente de dicha página [24]. Si tampoco se encontraba el enlace, se buscó la política en el resto del sitio, usando Google, y particularmente el *dork* [25]

```
site:[URL del sitio] política de privacidad
```

Para identificar el enlace a la política de privacidad, se buscaron aquellos que incluyesen las siguientes palabras clave: `privacidad`, `política`, `términos`, `datos`, `información`, `personales`, `ética`, `legal`, `transparencia`, `policies`, `privacy` y `policy`. Cuando se encontraba varios resultados (URLs) para un mismo sitio, se seleccionó el más relacionada con la temática de privacidad.

Si al final no se encontraba ningún documento de política de privacidad, se asumió que el sitio no tenía política.

Durante el proceso de análisis, se registraron tanto los sitios que tenían política de privacidad como aquellos que no la tenían.

2.1.4 Proceso de registro de datos de los documentos

Con la finalidad de recopilar y organizar la información de los sitios web seleccionados, se creó una tabla en una hoja de cálculo de Excel. A continuación, se describe cada uno de los campos de esta tabla:

- `Entity`*:* nombre de la entidad a la que pertenece el sitio web.
- `Acronym`*:* siglas o abreviaturas utilizadas para referirse a la entidad.
- `Entity URL Full`*:* URL completa del sitio web.
- `Entity URL`*:* nombre de host del sitio web.
- `Policy URL`*:* URL donde se ubica la política de privacidad de la entidad.

- `Observations`*:* cualquier observación o comentario relevante sobre el sitio web relacionada con el campo Policy URL.
- `Category`*:* categoría a la que pertenece la entidad (p. ej., banca, seguros, educación, entre otras.)
- `Visible in main page`*:* identifica si la política de privacidad es visible en la página principal del sitio web.
- `Visible in source code`*:* identifica si la política de privacidad es visible en el código fuente de la página principal.
- `Has policy`*:* identifica si tiene o no una política de privacidad.

2.2 Análisis manual de la información recolectada

En esta subsección se caracteriza la información obtenida a partir de la exploración de los sitios web, al buscar una política de privacidad.

2.2.1 Análisis descriptivo

Se obtuvo una serie de estadísticas que permitieron ilustrar el escenario de información recolectada. Entre ellas se incluyen datos de las categorías de sitios web involucrados o del nivel de cumplimiento de publicación de política de privacidad.

Asimismo, se identificaron problemas de visibilidad y acceso a la política de privacidad de varios sitios analizados.

2.2.2 Identificación de criterios para lectura de políticas de privacidad

Al analizar manualmente las políticas de privacidad, se identificaron detalles que luego se consideraron para su recolección automatizada. Entre esos detalles se encuentra el formato del documento de política (HTML, PDF, etc.).

A partir de la exploración manual, se detectaron también problemas de visibilidad de la política que pudiesen complicar una exploración automática.

También se consideró la existencia de falsos positivos al identificar la política de privacidad al usar las palabras clave indicadas en la sección 2.1.3. Ya que las palabras clave que se buscan son limitadas y podrían encontrarse en otras secciones de los sitios web.

Para la extracción del texto de la política de privacidad, se inspeccionó el código fuente de donde se publican para ubicar elementos (por ejemplo, de HTML) que comúnmente contienen ese texto.

Durante la inspección del código fuente, se encontró una variedad de etiquetas HTML para identificar el texto de las políticas de privacidad. Algunas de las etiquetas más comunes utilizadas fueron `<div>`, `<tr>` y `<p>`. Estas etiquetas permiten delimitar secciones específicas del contenido de la página web y facilitan la recolección del texto deseado. Asimismo, la identificación de estas etiquetas permitió posteriormente extraer el texto de la política de privacidad.

Durante el proceso de análisis, se caracterizaron las URLs de las políticas de privacidad, y particularmente la sección final de la URL. De esta caracterización se identificaron las palabras más comunes ahí incluidas. Entre estas palabras se seleccionaron posteriormente las más populares como palabras clave para la identificación de enlaces a políticas de privacidad.

2.3 Automatización de descarga de políticas de privacidad

En la presente sección, se expondrá la descripción del proceso de automatización de descarga de la política de privacidad de un sitio web a partir de su URL. Este procedimiento se llevó a cabo mediante la implementación de ciertos pasos específicos, que se detallan en la Fig. 1.

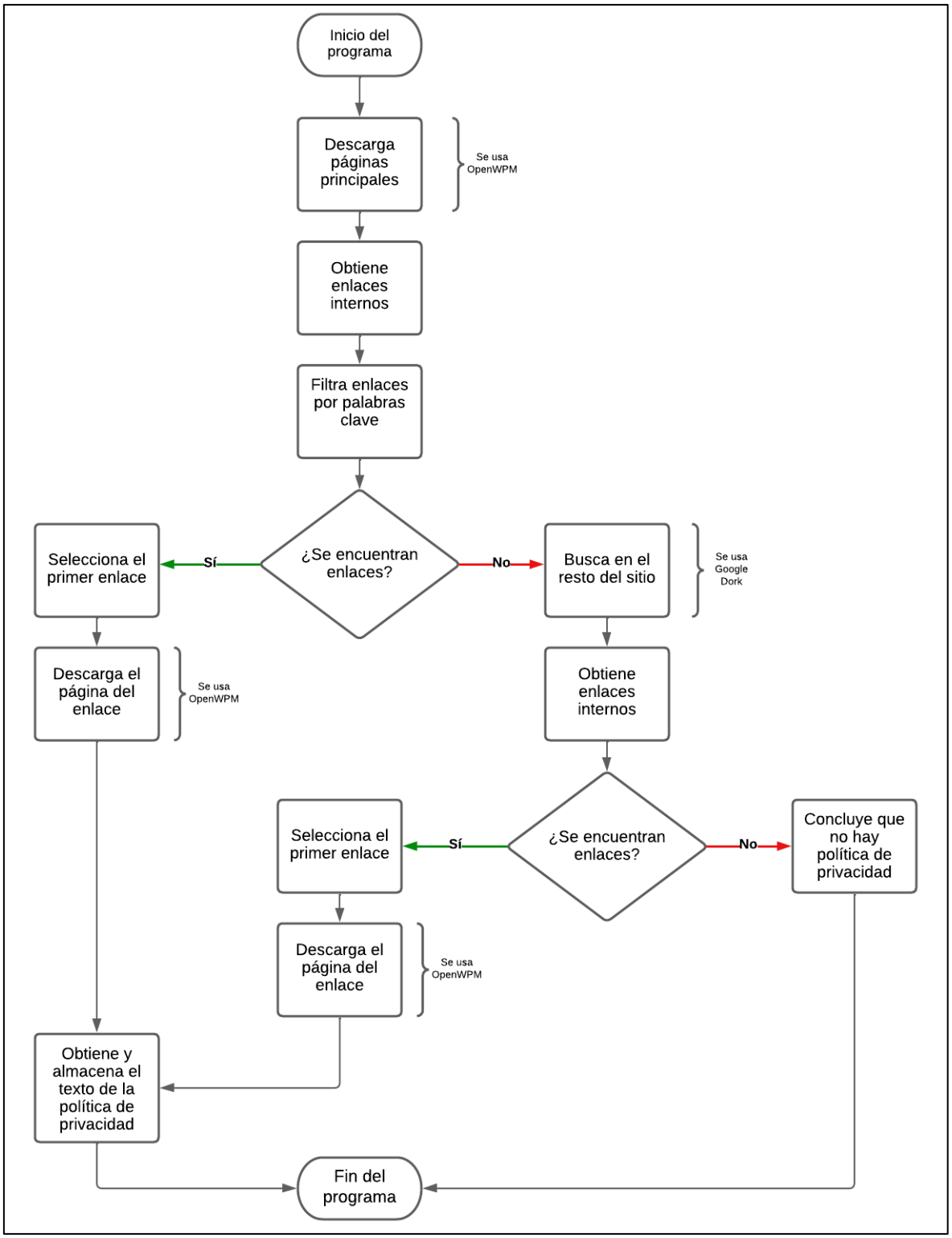

Figura 1. *Diagrama de flujo de automatización de descarga de políticas de privacidad.*

2.3.1 Descarga del código fuente de páginas principales

Para llevar a cabo la descarga de páginas principales, se utilizó la herramienta *OpenWPM* y su librería `DumpPageSourceCommand`.

La configuración de la herramienta empieza por definir los parámetros mediante la instancia de la clase `ManagerParams`. Dentro de los parámetros, se especificó la cantidad de navegadores que se desean utilizar (`NUM_BROWSERS`) para la exploración web, y los del navegador, mediante la instancia de la clase `BrowserParams`, la cual permite establecer el modo de visualización de la página web.

A continuación, se muestra el Código 1, lo cual es un fragmento de la configuración utilizada.

```
Manager_params_OpenWPM = ManagerParams(num_browsers=NUM_BROWSERS_OpenWPM)
browser_params_OpenWPM = [BrowserParams(display_mode=”headless”) for _ in
range(NUM_BROWSERS_OpenWPM)]
manager_params.data_directory= Path(“/home/jandres/Escritorio/Tesis/PaginasWeb”)
manager_params.log_path = Path(“/home/jandres/Escritorio/Tesis/PaginasWeb/openwpm.log”)
```

Código 1. *Configuración implementada para el uso de OpenWPM.*

En el código previo, se utilizó la opción `display_mode="headless"`, para la ejecución de los navegadores en segundo plano [26], mejorando así la eficiencia en la ejecución de la tarea, al no requerir que se ejecute la interfaz gráfica del navegador. Además, se estableció la ubicación de los archivos a descargar y del registro de logs que se utilizaron.

Asimismo, se utilizó el objeto `TaskManager`. Este objeto es responsable de recorrer una lista de sitios web y definir objetos `CommandSequence` para cada uno de ellos. Estos objetos establecen los comandos necesarios para obtener el código fuente y analizar el contenido de la página.

Uno de los comandos utilizados en el objeto `CommandSequence` es el `DumpPageSourceCommand` que permite descargar el código fuente de la página explorada. Para especificar el nombre del archivo de salida, se usó la opción `suffix`, para guardar el código descargado en un archivo con nombre igual al del dominio del sitio (`Entity URL` en la tabla de sitios web). El comando `DumpPageSourceCommand` también cuenta con la opción `timeout`, que establece el tiempo máximo de espera en segundos por cada descarga.

2.3.2 Identificación y descarga de URLs de políticas de privacidad

Una vez obtenido el código fuente de las páginas principales de los sitios, se procedió a obtener los enlaces internos de las páginas principales de los sitios web. Para lograr este objetivo, se generó una lista que contenía los archivos que se deseaban procesar, y cada archivo fue iterado para su procesamiento.

Durante el proceso de iteración, se utilizó la biblioteca `BeautifulSoup` para garantizar una lectura adecuada del archivo HTML. Esta herramienta permitió identificar todas las etiquetas `<a href>` presentes en el archivo para obtener los enlaces de cada sitio web.

Durante la implementación del proceso, se detectaron situaciones específicas, como la utilización de la etiqueta `<data-href>` por parte de ciertas empresas. Se tomó la medida correspondiente para la correcta configuración de esta etiqueta, lo que permitió la obtención de los enlaces internos de manera efectiva.

A continuación, se muestra un fragmento del Código 2, utilizado para la obtención de los enlaces de las páginas principales.

```
for n in etiquetas:
    with open("/home/jandres/Escritorio/Tesis/salida_enlaces.csv", "a+") as enlaces_politicas:
        parametro_enlace = n.attrs.get('href')
        parametro_enlace_elcomercio = n.attrs.get('data-href')
enlaces_politicas.write('\n % s' % parametro_enlace)
        enlaces_politicas.write('\n % s' % parametro_enlace_elcomercio)
```

Código 2. *Implementación para la obtención de enlaces en cada sitio web.*

Para la extracción de los enlaces relacionados con políticas de privacidad, se filtró los enlaces que contuvieran ciertas palabras clave en su contenido, tales como `privac`, `personales`, `politica` y `terminos.` Este filtrado se realizó mediante una lista de consultas SQLite a la tabla donde estaban almacenados los enlaces. A continuación, se muestra esta lista de consultas.

Para especificar estas palabras clave, se utilizó la lógica mencionada en la sección 2.1.3. Se ofrecen detalles sobre los resultados en la sección 3.1.6.

Luego, se extrajo la sección final de los enlaces resultantes (por ejemplo, de `https://example.com/politica_privacidad.html`, sería `politica_privacidad.html`), que se volvió a filtrar, para cada sitio web, con base en las palabras clave previamente mencionadas. Antes de este filtrado, se limpió la lista de URLs, eliminando aquellas que apuntaban a sitios web externos. Así también, se descartaron URLs demasiado extensas.

A continuación, se muestra el Código 3, utilizado para el proceso de limpieza.

```
cur.executescript("""
CREATE TABLE tabla_enlaces_finales_politicas AS SELECT DISTINCT(TRIM(ENLACES)) AS 'Policy_URL'
FROM tabla_enlaces_finales;
ALTER TABLE tabla_enlaces_finales_politicas ADD COLUMN ROW_ID;
UPDATE tabla_enlaces_finales_politicas SET ROW_ID = ROWID;
DELETE FROM tabla_enlaces_finales_politicas WHERE Policy_URL LIKE '%accounts.google%';
DELETE FROM tabla_enlaces_finales_politicas WHERE Policy_URL LIKE '%/search%';
DELETE FROM tabla_enlaces_finales_politicas WHERE Policy_URL LIKE '%/advanced%';
DELETE FROM tabla_enlaces_finales_politicas WHERE Policy_URL LIKE '%google%';
```

```
DELETE FROM tabla_enlaces_finales_politicas WHERE length(Policy_URL) > 1000;
ALTER TABLE tabla_enlaces_finales_politicas ADD COLUMN URL_PARTE_FINAL;
ALTER TABLE tabla_enlaces_finales_politicas ADD COLUMN CARACTER;
UPDATE tabla_enlaces_finales_politicas SET CARACTER = substr(Policy_URL,-1);
UPDATE tabla_enlaces_finales_politicas SET Policy_URL =
trim(substr(Policy_URL,1,length(Policy_URL)-1)) WHERE CARACTER = '/';
UPDATE tabla_enlaces_finales_politicas SET URL_PARTE_FINAL = replace(Policy_URL,
rtrim(Policy_URL, replace(Policy_URL,'/','')),'');
""")
```

Código 3. *Implementación del proceso de limpieza de URLs.*

A continuación, se muestra el Código 4, utilizado para el proceso de filtrado con base en las palabras clave.

```
consultas_2 = [
" SELECT trim(DISTINCT(Policy_URL)) FROM tabla_enlaces_finales_politicas WHERE Policy_URL LIKE
'%" + nombre_archivo + "%' AND Url_Parte_Final LIKE '%privac%' ",
" SELECT trim(DISTINCT(Policy_URL)) FROM tabla_enlaces_finales_politicas WHERE Policy_URL LIKE
'%" + nombre_archivo + "%' AND Url_Parte_Final LIKE '% personales %' ",
" SELECT trim(DISTINCT(Policy_URL)) FROM tabla_enlaces_finales_politicas WHERE Policy_URL LIKE
'%" + nombre_archivo + "%' AND Url_Parte_Final LIKE '%politica %' ",
" SELECT trim(DISTINCT(Policy_URL)) FROM tabla_enlaces_finales_politicas WHERE Policy_URL LIKE
'%" + nombre_archivo + "%' AND Url_Parte_Final LIKE '% terminos %' "]
```

Código 4. *Implementación del proceso de filtrado.*

En caso de no hallar un enlace con las palabras clave en la página principal, se buscó la política de privacidad en el resto del sitio web, usando Google, tal como se lo hiciera para la recolección manual. Para ello, se empleó el operador `site` (*Google dork*), junto con la URL del sitio, y las palabras clave `politica` y `privacidad`.

Se utilizó la herramienta *OpenWPM* para llevar a cabo esta tarea de manera automatizada. En el Código 5, se especifican los parámetros que se pasaron a la herramienta para su ejecución.

```
pagina_no_index = cur.execute(
    "SELECT [Entity URL Full] FROM REPORTE_FINAL WHERE [Entity URL Full] LIKE '%" +
nombre_archivo + "%'").fetchone()

pagina_no_index = "https://www.google.com/search?channel=fs&client=ubuntu&q=site%3A" +
pagina_no_index[
    0] + "+politicas+de+privacidad"
```

Código 5. *Parámetros enviados a OpenWPM para la descarga del sitio web.*

Los enlaces obtenidos mediante el proceso de búsqueda usando Google, fueron procesados de la misma forma que los que se obtuvieron al extraerlos de la página principal. Una vez identificado el posible documento de política, se procedió a su descarga mediante la herramienta *OpenWPM*.

Finalmente, en caso de que la mencionada lógica no permitiera encontrar el enlace deseado, se asumió que dicho sitio web no disponía de una política de privacidad.

2.3.3 Obtención del texto de la política de privacidad

Una vez descargado el documento identificado como política de privacidad, se procesó cada uno, abriéndolo y recuperando su contenido. Se utilizó la biblioteca `BeautifulSoup` para extraer la información contenida en las etiquetas HTML y mencionadas en la sección 2.2.2, y se guardó el texto recuperado, usando el método `get_text,` en un archivo de texto aplicando el método `writelines`.

Para documentos en formato PDF, se descargó el archivo y se extrajo su contenido con la biblioteca `PyPDF2`. Luego se guardó esa información en un archivo de texto, en una ruta específica tanto para los casos en los que se encontraron archivos HTML como PDF. El Código 6, muestra esta implementación.

```
with open(path_politicas + lista_paginas_politicas[y], "r") as archivo_final:
    contenido_pagina = archivo_final.read()
    soup_end = BeautifulSoup(contenido_pagina, 'html.parser')
pdf_url = re.search(r'http.*\.pdf', contenido_pagina)
if pdf_url:
    pdf_url = pdf_url.group(0)
    response = requests.get(pdf_url)

    with tempfile.NamedTemporaryFile(delete=True) as temp:
        temp.write(response.content)
        temp.seek(0)
        pdf_file = PyPDF2.PdfFileReader(temp)
        text = ''          for page in pdf_file.pages:
            text += page.extract_text()

    with open('/home/jandres/Escritorio/Tesis/Politicas/Archivos finales/' +
              nombre_lista_paginas_final_politicas_priv[
                  y] + '.txt', "w") as f:
        f.write(text)
else:
    with open('/home/jandres/Escritorio/Tesis/Politicas/Archivos finales/' +
nombre_lista_paginas_final_politicas_priv[
    y] + '.txt', "w") as f:
        for datos_txt in soup_end.find_all(['div', 'tr', 'p']):
            suma_txt = datos_txt.get_text()
            f.writelines(suma_txt)
```

Código 6. *Implementación de la obtención del texto de la política de privacidad tanto para documentos HTML y PDF.*

2.3.4 Medición de la eficacia en la identificación de políticas de privacidad

Se realizaron distintas evaluaciones para analizar la efectividad de la identificación de políticas de privacidad. En primer lugar, se realizó el conteo de los documentos que fueron correctamente identificados como políticas de privacidad. Asimismo, se llevó a cabo el conteo de aquellos documentos que fueron incorrectamente identificados como políticas de privacidad.

Por otro lado, también se registró el número de documentos de políticas que no fueron descargados. A partir de estos datos, se procedió al cálculo de la precisión en la identificación de políticas de privacidad. De esta manera, se pudo obtener una evaluación cuantitativa de la capacidad de identificación de políticas (y no políticas) de privacidad; esto con relación al número de políticas y no políticas identificadas de forma manual.

2.3.5 Clasificación automática de políticas de privacidad

Considerando la posibilidad de que la implementación de identificación y descarga de políticas obtenga documentos que no sean políticas, se procedió a construir un modelo predictivo, basado en aprendizaje automático, capaz de clasificar un documento como política o no política, que se aplicaría luego de que nuestro script descargue los documentos.

Para esto, se etiquetó manualmente (como políticas y no políticas – datos de salida) los documentos descargados por el script, y se construyó el modelo, usando el texto de los documentos como datos de entrada para el algoritmo de aprendizaje automático.

Antes de enviar los datos al algoritmo, se realizó una limpieza de los datos, eliminando cualquier información redundante o innecesaria (saltos de línea). Después, se llevó a cabo la tokenización de los textos, dividiendo las palabras en unidades con sentido para que puedan ser procesadas. Además, se eliminaron las *stop words* del texto. El Código 7 la implementación para el preprocesamiento del texto de las políticas.

```
def preprocess_texto(texto):
    texto = texto.replace('\n', ' ')
    palabras = nltk.palabra_tokenize(texto.lower())
    palabras = [palabra for palabra in palabras if re.match(r'^[a-záéíóúñ]+$', palabra)]
    stop_palabras = set(stopwords.words('spanish'))
    palabras = [palabra for palabra in palabras if palabra not in stop_palabras]
    clean_texto = ' '.join(palabras)
    return clean_texto
```

Código 7. *Implementación de preprocesamiento de texto de las políticas de privacidad.*

Luego se formó el conjunto de datos a usarse, con dos campos: el texto procesado de los documentos, como datos de entrada; y la etiqueta de política o no política, a como datos de salida.

Este conjunto de datos se dividió en subconjuntos de training (80% de los datos) y test (20% de los datos) y se pasó como entrada al algoritmo de aprendizaje automático para la creación del modelo correspondiente y su evaluación. En el Código 8 se muestra el código utilizado para la estructuración del conjunto de datos, la división en subconjuntos, la construcción del modelo, y la evaluación, todo esto utilizando las librerías `pandas` y `scikitlearn`.

El algoritmo de aprendizaje automático utilizado fue `Naive Bayes`, pues es el que usualmente se utiliza para tareas de clasificación de texto como en la detección de spam.

```
for filename in os.listdir(valid_path):
    if filename.endswith(".txt"):
        with open(os.path.join(valid_path, filename), 'r') as f:
            contenido = f.read()
            clean_texto = preprocess_texto(contenido)
            entity_name = os.path.splitext(filename)[0]                df =
pd.DataFrame({"ENTITY": [entity_name], "contenido": [clean_text], "CLASS": [1]})
            df = df[["ENTITY", "contenido", "CLASS"]]
            frames.append(df)


for filename in os.listdir(invalid_path):
    if filename.endswith(".txt"):
        with open(os.path.join(invalid_path, filename), 'r') as f:
            contenido = f.read()
            clean_texto = preprocess_texto(contenido)
            entity_name = os.path.splitext(filename)[0]                df =
pd.DataFrame({"ENTITY": [entity_name], "contenido": [clean_text], "CLASS": [0]})
            df = df[["ENTITY", "contenido", "CLASS"]]
            frames.append(df)

resultado = pd.concat(frames, ignore_index=True)


df_data = resultado[["contenido", "CLASS"]]
df_x = df_data['contenido']
df_y = df_data['CLASS']
corpus = df_x cv = CountVectorizer()
X = cv.fit_transform(corpus)

X_train, X_test, y_train, y_test = train_test_split(X, df_y, test_size=0.20, random_state=50)
clf = MultinomialNB()
clf.fit(X_train,y_train)
print("Accuracy of Model",clf.score(X_test,y_test)*100,"%")
```

Código 8. *Implementación del algoritmo de aprendizaje automático utilizado para la clasificación de política o no política de privacidad.*

Finalmente, se evaluó el modelo mediante el cálculo de la exactitud o *accuracy* del mismo.

2.4 Obtención de información de políticas de privacidad de México

En esta sección, se llevó a cabo el proceso previamente descrito en las secciones 2.1, 2.2 y 2.3 para sitios web ecuatorianos. Sin embargo, en este caso, el análisis se enfocó en sitios web de México.

3. RESULTADOS, CONCLUSIONES Y RECOMENDACIONES

Luego de la recopilación y procesamiento de datos, es esencial realizar una interpretación rigurosa de los mismos para extraer conclusiones válidas y relevantes. Adicionalmente, estas conclusiones y resultados pueden utilizarse para generar recomendaciones, tanto para la toma de decisiones como para investigaciones futuras.

3.1 Resultados

En el presente trabajo, se muestran las interpretaciones obtenidas a través de un análisis sistemático de los datos recolectados. Estas interpretaciones permitieron identificar los patrones y relaciones relevantes en los datos.

3.1.1 Número de sitios web incluidos en el conjunto de datos

Se identificaron 211 URLs de sitios web ecuatorianos, a partir de los criterios descritos en la sección 2.1.2, que luego exploraron para la obtención de la política de privacidad de cada uno. Las categorías de estos sitios incluyeron `educación`, `empleo`, `entretenimiento`, `finanzas`, `gobierno`, `salud`, `noticias`, `seguros` y `compras en línea`. En la Fig. 2, se ilustra la cantidad de sitios web identificados en función de su categoría.

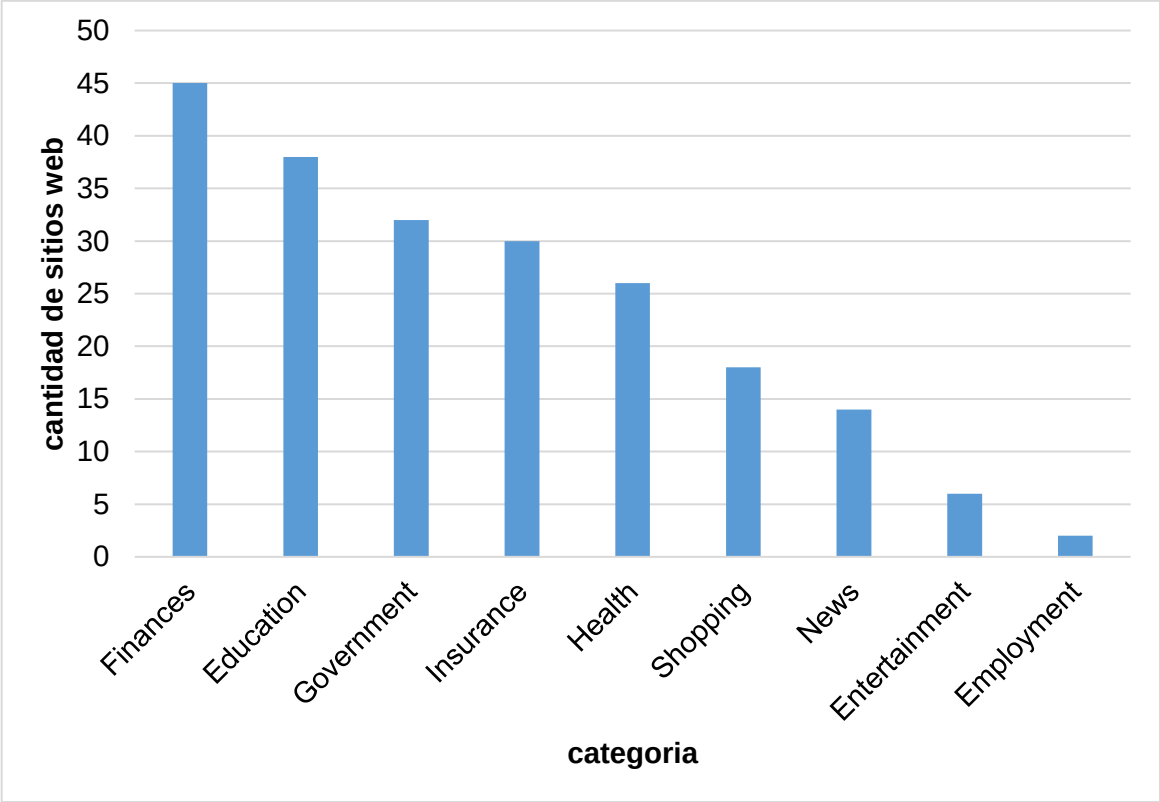

Figura 2. *Cantidad de sitios web ecuatorianos, por categoría, seleccionados para este trabajo.*

3.1.2 Sitios web con y sin políticas de privacidad

Luego de explorar manualmente los 211 sitios web ecuatorianos, se encontró que un número significativo de ellos carecía de políticas de privacidad. En concreto, se determinó que el 31.27% de estos sitios no contaba con una política, mientras que el 68.73% (145) sí tenía una política que trataba de informar a los usuarios sobre cómo se manejaban sus datos personales.

3.1.3 Sitios web con política de privacidad accesible en la página

De los 145 sitios web ecuatorianos con política de privacidad, el 56.55% tenía una sección visible en su página principal que llevaba a su política de privacidad. Estos resultados sugieren que una parte significativa de los sitios web evaluados no proporcionaba una sección visible para acceder a su política de privacidad, aunque sí contaban con una.

Por otro lado, el 63.44% de las URLs evaluadas contaba con una URL de política de privacidad visible en su página principal. Además, se observó que el 55.86% de los sitios web tenían su política de privacidad visible tanto en su página principal como en su código fuente.

3.1.4 Sitios web con política de privacidad ubicada fuera de la página principal

Se encontró que el 43.45% de las URLs evaluadas que no tenían una URL de política de privacidad visible en su página principal.

Para poder obtener la política de privacidad de estos sitios, fue necesario utilizar herramientas como *Google Dorks*, tal como se mencionó en la 2.1.3.

3.1.5 Tipos de formatos de publicación de la política de privacidad

La mayor parte de políticas de privacidad estaban publicadas en formato HTML. Una décima parte estaba publicada en formato PDF, y una mínima porción en formato de imagen.

3.1.6 Palabras más comunes en los parámetros de las URLs de las políticas de privacidad

Del análisis de las URLs de las políticas de privacidad, y particularmente de la última sección de éstas, se encontró que las palabras clave más populares fueron: `privacidad, politica, terminos, datos, privacy, policy, policies, informacion, politicas, personales, etica, legal y transparencia`. Tal como se observa en la Fig. 3**,** las palabras `privacidad y privacy` aparecen 86 veces; la palabra `politica`, 19 veces; `terminos`, 18 veces; y `personales,` 16 veces en total.

.

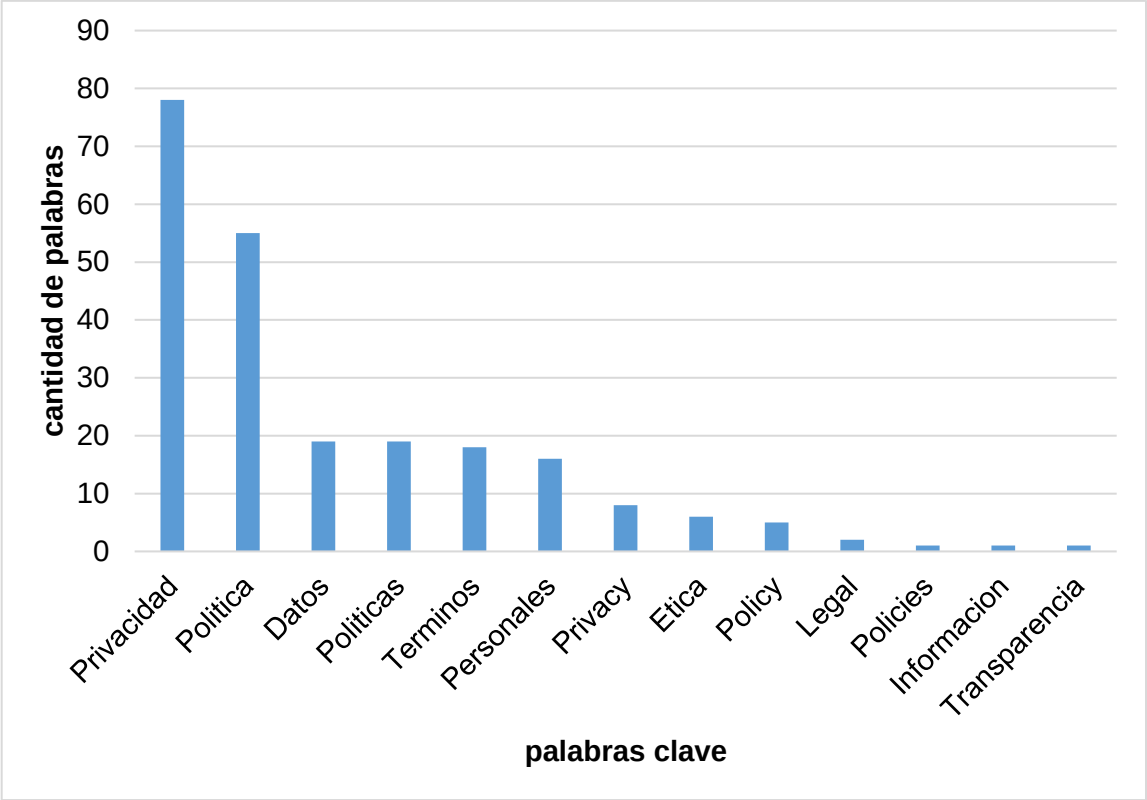

Figura 3. *Prevalencia de palabras clave en las URLs de las políticas de privacidad ecuatorianas.*

Estas palabras clave (las más prevalentes) se usaron para construir el filtro que se explica en 2.3.2 y que permiten estimar si una URL corresponde a la de una política de privacidad.

3.1.7 Precisión de la recolección de las políticas de privacidad en Ecuador

Luego de ejecutar el script para la identificación y descarga de políticas de privacidad, se descargaron correctamente 115 políticas de privacidad. Por otro lado, no se descargó ningún documento de 50 sitios que no albergaban ninguna política de privacidad.

Con relación a los 211 sitios web ecuatorianos, esto corresponde a una precisión en el filtrado de 78.2%. Como parte del error en la operación del script, un grupo de políticas de privacidad no se descargaron, y hubo un grupo de falsos positivos (documentos descargados que no eran políticas).

3.1.8 Tiempo de recolección de políticas de privacidad

Se evaluó el tiempo de ejecución del Script encargado de obtener las políticas de privacidad de 211 URLs en Ecuador. Para ello, se midió el tiempo de exploración y

descarga de las páginas principales, así como el tiempo de procesamiento de los archivos HTML, y la obtención de las URLs de las políticas de privacidad.

Los resultados obtenidos indican que el tiempo promedio de obtención de políticas de privacidad fue de poco más de un minuto (1.13). Hacer este trabajo de forma manual, podría tomar hasta 5 minutos.

3.1.9 Resultados obtenidos con las URLs de México

Durante la investigación se analizaron las páginas web más visitadas en México, tomando como base un total de 78 URLs. Dichas URLs fueron aplicadas en varias categorías, incluyendo `finanzas`, `educación`, `compras`, `salud`, `noticias`, `gobierno`, `entretenimiento`, `organizaciones no gubernamentales` y `empleo`, como se muestra en la Fig. 4.

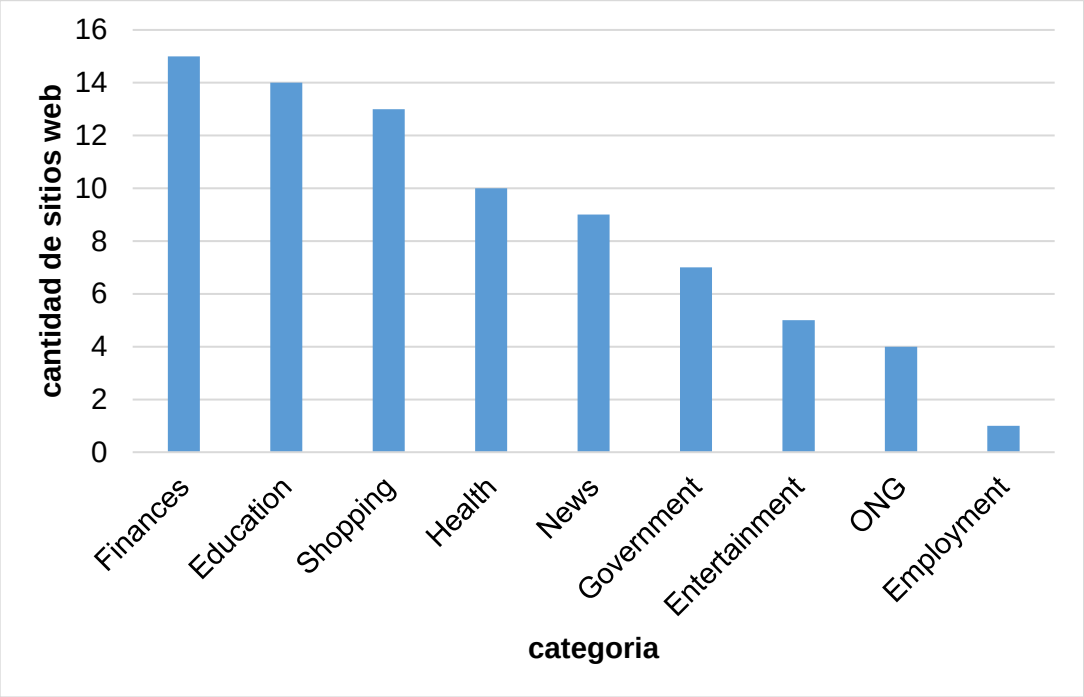

Figura 4. *Cantidad de sitios web mexicanos, por categoría, seleccionados para este trabajo.*

Se incluyeron 78 sitios web mexicanos, y, del análisis manual realizado, se encontró que el 97.43% de ellos tenían política de privacidad.

De estos sitios, casi el 80% tiene dicha política visible en la página principal. En consecuencia, alrededor del 20% de políticas de privacidad se enlazaban fuera de la página principal.

En cuanto a las palabras más comunes, se notó que la mayoría de URLs de políticas de privacidad en cada parámetro de esta URL contenía la palabra `privacidad.`

Las etiquetas usadas para contener el texto de la política de privacidad fueron esencialmente las mismas que en el caso de Ecuador. También, en un par de casos, se generó una prueba de captcha que impidió la descarga de las políticas de privacidad.

Al aplicar el mismo script que se utilizó para el caso ecuatoriano en los sitios web mexicanos, se obtuvo una precisión de 80.76% en la identificación y descarga de políticas de privacidad.

El tiempo de procesamiento fue similar al de los sitios web ecuatorianos, ya que se utilizó la misma lógica y no se modificó nada del script.

3.1.10 Modelo de clasificación para identificar una política de privacidad

Como resultado del proceso de entrenamiento del modelo de aprendizaje automático, usando como entrada el texto de los documentos descargados de los sitios ecuatorianos, se obtuvo una *accuracy* de alrededor del 80%.

Considerando que en el conjunto de datos se tenían 115 registros de políticas de privacidad (70%) de un conjunto de 165 documentos descargados, la *accuracy* del modelo obtenido (80%) obtiene una ganancia del 10% sobre la línea base (70%). Este resultado se muestra a continuación en la Fig. 5.

Figura 5. *Resultado exacto obtenido en Python de acurracy del modelo de aprendizaje automático*

3.2 Conclusiones

- En Ecuador, la prevalencia de políticas de privacidad en los sitios web es aún baja. Esto podría ser indicador de que las entidades aún no se empiezan a preocupar por cumplimiento de LOPDP.

- Casi la mitad de los sitios web que sí tienen una política de privacidad no la enlazan en la página principal, lo que evidenciaría otros problemas, como falta de transparencia, o falta de mantenimiento de los sitios.

- La identificación y descarga automática de políticas de privacidad en el Ecuador presenta ciertas limitaciones, pues la publicación de este tipo de documentos no sigue ningún estándar. Así, la falta de homogeneidad en la forma en que las empresas y organizaciones publican sus políticas de privacidad dificulta la auditoría a gran escala de ese parámetro ligado al cumplimiento de la LOPDP.

- El script, aunque usa un método heurístico sencillo, mantiene la precisión en la identificación y descarga de políticas de privacidad cuando se evalúan sitios de otro escenario como el mexicano. Con pequeñas mejoras, podría mejorar su precisión.

- La sencillez del filtro se evidencia en la descarga de cierto porcentaje de documentos que no son políticas (falsos positivos), pero la inclusión de una etapa de clasificación de documentos permite identificar buena parte de esos casos.

- La falta de acceso a algunos enlaces en las políticas de privacidad de los sitios web puede ser una señal de preocupación para los usuarios. Las políticas de privacidad son una herramienta vital para proteger la información personal en línea, por lo que es importante que los enlaces estén disponibles y sean fácilmente accesibles. Es necesario que los sitios web mantengan sus políticas de privacidad actualizadas y funcionales, para brindar una experiencia de navegación segura y confiable a los usuarios.

- La automatización del proceso de identificación y descarga de las políticas de privacidad es una herramienta muy útil para facilitar las auditorías en materia de privacidad. Esta automatización no solo ahorra tiempo y esfuerzo manual, sino que también permite que estas auditorías se realicen a mayor escala y con mayor frecuencia, lo que resulta en una mayor eficacia en el control de la protección de los datos personales.

3.3 Recomendaciones

- La precisión del filtro puede mejorarse incluyendo otras palabras clave, dependiendo del escenario en el que se realice la recolección de políticas de privacidad. Se recomienda explorar otros escenarios similares, para incluir las palabras clave que se usan para identificar las políticas de privacidad.

- Al filtrar los enlaces de las páginas principales con base en las palabras clave, es posible obtener varios enlaces que cumplan el filtrado. Actualmente, se escoge el primero. Se recomienda agregarle cierta “inteligencia” al filtro para seleccionar el enlace que mejor se ajuste a una política de privacidad.

- Con respecto al modelo de clasificación de políticas, se recomienda enriquecer el conjunto de datos con más documentos que no sean políticas de privacidad, para reducir el sesgo inherente al tener pocos documentos de ese tipo.

- Es importante ampliar el conjunto de datos utilizados para entrenar sus modelos de aprendizaje automático en relación con las políticas. A menudo, los conjuntos de datos están compuestos únicamente por políticas de privacidad, lo que puede generar un sesgo en los resultados y decisiones automatizadas. Por esta razón, se recomienda incluir una variedad de documentos que no sean políticas de privacidad, como términos y condiciones de uso, políticas de seguridad, y otros documentos legales relevantes. Al incluir una mayor diversidad de documentos, se puede reducir el sesgo en los modelos de aprendizaje automático, lo que a su vez puede mejorar la precisión y confiabilidad de las decisiones automatizadas.

- Se recomienda realizar periódicamente el proceso de recolección de políticas de privacidad para monitorear la evolución en el tiempo de la prevalencia y de la calidad de estos documentos.

- La publicación del documento de política de privacidad debería estandarizarse. Por ejemplo, todos los sitios web deberían publicarlo en la página principal y en un lugar visible. Además, debería facilitarse su identificación en el código fuente de la página para hacer posible una auditoría más precisa y veloz.

- Los sitios web deben realizar un seguimiento constante y regular de los enlaces en sus políticas de privacidad para garantizar que estén siempre disponibles y accesibles para los usuarios.

4. REFERENCIAS BIBLIOGRÁFICAS

[1] S. Chen Mok, «Privacidad y protección de datos: un análisis de legislación comparada,» Iberdrola, Agosto 2010. [En línea]. Disponible: https://www.scielo.sa.cr/scielo.php?script=sci_arttext&pid=S1409-469X2010000100004. [Último acceso: 10 Diciembre 2022].

[2] A. Clay, «Estadísticas de la situación digital de Ecuador en el 2020-2021,» Branch, 05 Mayo 2021. [En línea]. Disponible: https://branch.com.co/marketing-digital/estadisticas-de-la-situacion-digital-de-ecuador-en-el-2020-2021/. [Último acceso: 11 Diciembre 2022].

[3] E. Tapia, «Gobierno lanzará un plan de reducción de tarifas de Internet el 27 de noviembre del 2019,» El Comercio, 26 Noviembre 2019. [En línea]. Disponible: https://www.elcomercio.com/actualidad/negocios/gobierno-plan-reduccion-tarifa-internet.html. [Último acceso: 11 Diciembre 2022].

[4] Universia, «¿Cuáles son las ventajas y desventajas de las redes sociales?,» Universia 2020, 19 Agosto 2014. [En línea]. Disponible: https://orientacion.universia.edu.pe/infodetail/consejos/orientacion/cuales-son-las-ventajas-y-desventajas-de-las-redes-sociales--1302.html. [Último acceso: 12 Diciembre 2022].

[5] C. Botero Marino, F. Guzmán Duque, S. Jaramillo Otoya y S. Gómez Upegui, «El derecho a la libertad de expresión,» Julio 2017. [En línea]. Disponible: https://www.dejusticia.org/wp-content/uploads/2017/07/El-derecho-a-la-libertad-de-expresi%C3%B3n-PDF-FINAL-Julio-2017-1-1.pdf. [Último acceso: 13 Diciembre 2022].

[6] Asamblea Nacional, «Ley Orgánica De Protección De Datos Personales,» 26 Mayo 2021. [En línea]. Disponible: https://www.telecomunicaciones.gob.ec/wp-content/uploads/2021/06/Ley-Organica-de-Datos-Personales.pdf. [Último acceso: 15 Diciembre 2022].

[7] R. Amos, G. Acar, E. Lucherini, M. Kshirsagar, A. Narayanan y J. Mayer, «Privacy Policies over Time: Curation and Analysis of a Million-Document Dataset,» 26 Mayo 2021. [En línea]. Disponible: https://doi.org/10.1145/3442381.3450048. [Último acceso: 15 Diciembre 2022].

[8] Real Academia de la Lengua Española, «Diccionario panhispánico del español jurídico,» 11 Diciembre 2022. [En línea]. Disponible: https://dpej.rae.es/lema/privacidad. [Último acceso: 28 Febrero 2023].

[9] Comisión Europea, «¿Qué son los datos personales?,» [En línea]. Disponible: https://ec.europa.eu/info/law/law-topic/data-protection/reform/what-personal-data_es. [Último acceso: 28 Febrero 2023].

[10] Digital Guide IONOS, «La política de privacidad de las páginas web,» 09 Febrero 2022. [En línea]. Disponible: https://www.ionos.es/digitalguide/paginas-web/derecho-digital/politica-de-privacidad-gana-la-confianza-de-tus-usuarios/. [Último acceso: 11 Diciembre 2022].

[11] Naciones Unidas, «La Declaración Universal de Derechos Humanos,» 1984 Diciembre 10 .

[12] A. Roig Batalla, « Tecnología, libertad y privacidad,» 2018. [En línea]. Disponible: https://ddd.uab.cat/pub/caplli/2011/149207/libexpinf_a2011p45.pdf. [Último acceso: 28 Febrero 2023].

[13] Asamblea Nacional del Ecuador, «Ley Orgánica de protección de datos personales,» 21 Mayo 2021. [En línea]. Disponible: https://www.asambleanacional.gob.ec/sites/default/files/private/asambleanacional/filesasambleanacionalnameuid-29/Leyes%202013-2017/920-lmoreno/ro-459-5to-sup-26-05-2021.pdf. [Último acceso: 28 Febrero 2023].

[14] L. G. Esquivel-Quirós, E. G. Barrantes y F. Esponda Darlington, «Marco de medición de la privacidad,» 14 Enero 2019. [En línea]. Disponible: https://www.proquest.com/openview/b1b9522ad5d32d45a59196a912c857ae/1?pq-origsite=gscholar&cbl=1006393#:~:text=El%20objetivo%20de%20las%20m%C3%A9tricas,los%20due%C3%B1os%20de%20los%20datos.. [Último acceso: 28 Febrero 2023].

[15] J. P. Ramírez Ramírez, «Análisis automatizado de políticas de privacidad en Ecuador,» Febrero 2022. [En línea]. Disponible: https://bibdigital.epn.edu.ec/handle/15000/22378. [Último acceso: 27 Febrero 2023].

[16] P. Galiana, «IEBS,» 14 Noviembre 2022. [En línea]. Disponible: https://www.iebschool.com/blog/que-es-el-web-scrapping-y-como-se-utiliza-en-los-negocios-digital-business/. [Último acceso: 28 Febrero 2023].

[17] WebTAP Princeton University, «Software: OpenWPM,» Center for Information Technology Policy, 2020. [En línea]. Disponible: https://webtap.princeton.edu/software/. [Último acceso: 28 Febrero 2023].

[18] OpenWPM, «Using OpenWPM,» OpenWPM, 2021. [En línea]. Disponible: https://openwpm.readthedocs.io/en/latest/Using_OpenWPM.html. [Último acceso: 28 Febrero 2023].

[19] E. Klein y S. Bird, Natural Language Processing with Python, Sebastopol: CA: O'Reilly Media, 2009.

[20] A. Moreno, «Procesamiento del lenguaje natural ¿qué es?,» Instituto de ingeniería del conocimiento, Noviembre 2018. [En línea]. Disponible: https://www.iic.uam.es/inteligencia/que-es-procesamiento-del-lenguaje-natural/. [Último acceso: 28 Febrero 2023].

[21] D. Huarcaya Taquiri, «Traducción automática neuronal para lengua nativa,» 2020. [En línea]. Disponible: https://repositorio.upeu.edu.pe/handle/20.500.12840/4143. [Último acceso: 28 Febrero 2023].

[22] Python Software Foundation, «Contenido de la documentación de Python,» Python Software Foundation, 2023. [En línea]. Disponible: https://docs.python.org/es/3/contents.html. [Último acceso: 28 Febrero 2023].

[23] Amazon Web Services, «Alexa Top Sites,» Amazon Web Services, 2023. [En línea]. Disponible: https://aws.amazon.com/es/alexa-top-sites/. [Último acceso: 24 Febrero 2023].

[24] Support Google, «Usar herramientas de desarrollo de Chrome para revisar etiquetas,» Google, 2023. [En línea]. Disponible: https://support.google.com/campaignmanager/answer/2828688?hl=es#:~:text=Haga%20clic%20con%20el%20bot%C3%B3n,alt%2Bcomando%2Bi%22.. [Último acceso: 24 Febrero 2023].

[25] J. Campderrós Vilá, «Ataques y vulnerabilidades web,» Diposit Digital, 2019. [En línea]. Disponible: http://diposit.ub.edu/dspace/handle/2445/143419. [Último acceso: 24 Febrero 2023].

[26] OpenWPM, «Browser and Platform Configuration,» OpenWPM, 2021. [En línea]. Disponible: https://openwpm.readthedocs.io/en/latest/Configuration.html#browser-configuration-options. [Último acceso: 28 Febrero 2023].

5. ANEXOS

ANEXO I. Código ha sido desarrollado utilizando Python.

ANEXO I. Código que ha sido desarrollado utilizando Python

El código generado para llevar a cabo el análisis automatizado se puede encontrar en el enlace que se adjunta.

https://github.com/J-Andres-Canar-R/recoleccion_automatica_de_politicas_de_privacidad