INSTITUTO POLITÉCNICO NACIONAL

CENTRO DE INVESTIGACIÓN EN COMPUTACIÓN

Clasificación textual de información personal sensible

T E S I S

QUE PARA OBTENER EL GRADO DE:
MAESTRÍA EN CIENCIAS DE LA COMPUTACIÓN

PRESENTA:

Ing. Sara De Jesús Sánchez

DIRECTORES DE TESIS:

Dr. Eleazar Aguirre Anaya
Dr. Francisco Hiram Calvo Castro

Ciudad de México
Enero de 2023

SIP-13
REP 2017

INSTITUTO POLITÉCNICO NACIONAL
SECRETARIA DE INVESTIGACIÓN Y POSGRADO

ACTA DE REGISTRO DE TEMA DE TESIS
Y DESIGNACIÓN DE DIRECTOR DE TESIS

Ciudad de México, a 25 de agosto del 2021

El Colegio de Profesores de Posgrado del **Centro de Investigación en Computación** en su Sesión
(Unidad Académica)

Ordinaria No. 05 celebrada el día 28 del ~~mes~~ mayo de 2021, conoció la solicitud presentada por el (la) alumno (a):

Apellido Paterno:	DE JESÚS	Apellido Materno:	SÁNCHEZ	Nombre (s):	SARA

Número de registro: A 2 1 0 3 5 7

del Programa Académico de Posgrado: **Maestría en Ciencias de la Computación**

Referente al registro de su tema de tesis; acordando lo siguiente:

1.- Se designa al aspirante el tema de tesis titulado:

"Clasificación textual de información personal sensible"

Objetivo general del trabajo de tesis:

Diseñar una herramienta para la identificación y clasificación textual de información personal sensible, mediante técnicas de inteligencia artificial, para prevenir su exposición en espacios públicos.

2.- Se designa como Directores de Tesis a los profesores:

Director: **Dr. Eleazar Aguirre Anaya** 2° Director: **Dr. Francisco Hiram Calvo Castro**

No aplica: ☐

3.- El Trabajo de investigación base para el desarrollo de la tesis será elaborado por el alumno en:

Centro de Investigación en Computación

que cuenta con los recursos e infraestructura necesarios.

4.- El interesado deberá asistir a los seminarios desarrollados en el área de adscripción del trabajo desde la fecha en que se suscribe la presente, hasta la aprobación de la versión completa de la tesis por parte de la Comisión Revisora correspondiente.

Director de Tesis

Dr. Eleazar Aguirre Anaya

2º Director de Tesis

Dr. Francisco Hiram Calvo Castro

Aspirante

Sara De Jesús Sánchez

Presidente del Colegio

Dr. Marco Antonio Moreno Ibarra

INSTITUTO POLITECNICO NACIONAL
CENTRO DE INVESTIGACION EN COMPUTACION
DIRECCION

SIP-14
REP 2017

INSTITUTO POLITÉCNICO NACIONAL

SECRETARÍA DE INVESTIGACIÓN Y POSGRADO

ACTA DE REVISIÓN DE TESIS

En la Ciudad de México siendo las 10:00 horas del día 16 del mes de diciembre del 2022 se reunieron los miembros de la Comisión Revisora de la Tesis, designada por el Colegio de Profesores de Posgrado de: Centro de Investigación en Computación para examinar la tesis titulada:

"Clasificación textual de información personal sensible"

del (la) alumno (a):

Apellido Paterno:	DE JESÚS	Apellido Materno:	SÁNCHEZ	Nombre (s):	SARA

Número de registro: A 2 1 0 3 5 7

Aspirante del Programa Académico de Posgrado: **Maestría en Ciencias de la Computación**

Una vez que se realizó un análisis de similitud de texto, utilizando el software antiplagio, se encontró que el trabajo de tesis tiene 07 % de similitud. **Se adjunta reporte de software utilizado.**

Después que esta Comisión revisó exhaustivamente el contenido, estructura, intención y ubicación de los textos de la tesis identificados como coincidentes con otros documentos, concluyó que en el presente trabajo **SI** ☐ **NO** ☒ **SE CONSTITUYE UN POSIBLE PLAGIO.**

JUSTIFICACIÓN DE LA CONCLUSIÓN: *(Por ejemplo, el % de similitud se localiza en metodologías adecuadamente referidas a fuente original)*

El porcentaje hace referencia al formato, referencias y definiciones formales

****Es responsabilidad del alumno como autor de la tesis la verificación antiplagio, y del Director o Directores de tesis el análisis del % de similitud para establecer el riesgo o la existencia de un posible plagio.**

Finalmente y posterior a la lectura, revisión individual, así como el análisis e intercambio de opiniones, los miembros de la Comisión manifestaron **APROBAR** ☒ **SUSPENDER** ☐ **NO APROBAR** ☐ la tesis por **UNANIMIDAD** ☒ o **MAYORÍA** ☐ en virtud de los motivos siguientes:

Cumple con los requisitos indispensables de una tesis de maestría.

COMISIÓN REVISORA DE TESIS

Dr. Eleazar Aguirre Anaya
Director de Tesis

Dr. Moisés Salinas Rosales

Dr. Raúl Acosta Bermejo

Dr. Francisco Hiram Calvo Castro
2°Director de Tesis

Dra. Gina Gallegos García

Dra. Sandra Dinora Orantes Jiménez

ESTADOS UNIDOS MEXICANOS
INSTITUTO POLITÉCNICO NACIONAL
CENTRO DE INVESTIGACIÓN EN COMPUTACIÓN
DIRECCIÓN
IPN-CIC

Dr. Francisco Hiram Calvo Castro
PRESIDENTE DEL COLEGIO DE PROFESORES

INSTITUTO POLITÉCNICO NACIONAL

SECRETARÍA DE INVESTIGACIÓN Y POSGRADO

CARTA DE AUTORIZACIÓN DE USO DE OBRA PARA DIFUSIÓN

En la Ciudad de México el día 4 del mes de enero del año 2023, el (la) que suscribe Sara De Jesús Sánchez, alumno(a) del programa Maestría en Ciencias de la Computación, con número de registro A210357, adscrito(a) al Centro de Investigación en Computación, manifiesta que es autor(a) intelectual del presente trabajo de tesis bajo la dirección del Dr. Eleazar Aguirre Anaya y Dr. Francisco Hiram Calvo Castro y cede los derechos del trabajo intitulado “Clasificación textual de información personal sensible”, al Instituto Politécnico Nacional, para su difusión con fines académicos y de investigación.

Los usuarios de la información no deben reproducir el contenido textual, gráficas o datos del trabajo sin el permiso expresado del autor y/o director(es). Este puede ser obtenido escribiendo a las siguiente(s) dirección(es) de correo. sdejesuss2021@cic.ipn.mx, eaguirre@cic.ipn.mx, hcalvo@cic.ipn.mx. Si el permiso se otorga, al usuario deberá dar agradecimiento correspondiente y citar la fuente de este.

Sara De Jesús Sánchez

Resumen

En esta tesis se muestra el desarrollo de ICIS, un modelo que utiliza técnicas de procesamiento de lenguaje natural y aprendizaje automático para identificar y clasificar la información personal sensible presente en textos, lo que permitirá prevenir su exposición en los medios públicos de las organizaciones gubernamentales.

Se propone una taxonomía con 55 tipos datos personales, agrupados en 10 categorías. A partir de esta taxonomía se identifican los datos personales en los documentos, utilizando técnicas de procesamiento de lenguaje natural. La identificación considera el contexto, no sólo el formato o las palabras asociadas, sino que los datos personales estén relacionados con un titular, que se refieran a alguien en un mismo segmento de texto. El modelo identifica, en paralelo, los datos personales en cada segmento de texto y forma vectores con los que se hace la clasificación de información sensible.

Se propone una definición de la Información Personal Sensible en términos computacionales, con base en cuatro clasificaciones de los tipos de datos personales: clasificación de datos sensibles unitarios o C_DSU, clasificación de datos sensibles no unitarios o C_DSNU, clasificación de datos personales identificadores o C_DPI y clasificación de datos personales o C_DP. Utilizando algoritmos de aprendizaje automático, el modelo hace las cuatro clasificaciones a los vectores de los segmentos de texto y, con estos resultados, genera un nuevo vector con el que hace la clasificación de información sensible o C_IS.

Este modelo forma parte del proyecto PICIS, Plataforma de Identificación, Clasificación y Monitoreo de Información sensible, que es uno de los ganadores del Fondo de Innovación en Ciberseguridad de Latinoamérica en 2021, patrocinado por la OEA, Cisco y la Fundación Citi. El proyecto PICIS es desarrollado por el Laboratorio de Ciberseguridad del CIC, IPN. PICIS representa un control de seguridad enfocado a prevenir la exposición de información sensible en documentos, en los medios públicos del gobierno federal de México.

Abstract

This thesis shows the development of ICIS, a model that uses natural language processing and machine learning techniques to identify and classify sensitive personal information present in texts, to prevent its exposure in the public media of government organizations.

A taxonomy is proposed, it's made up of 55 personal data types grouped by 10 categories. The personal data in the documents is identified using natural language processing techniques. The identification considers the context, not only the format or the associated words, but that personal data is related to a holder, that it refers to someone in a text segment. The model identifies the personal data in each text segment in parallel and builds vectors, the sensitive information classification is made with them.

A Sensitive Personal Information definition is proposed, in computational terms, based on the personal data type classifications: unit sensitive data classification or C_DSU, non-unit sensitive data classification or C_DSNU, identifier personal data classification or C_DPI, and personal data classification or C_DP. Using machine learning algorithms, the model makes the four classifications; with these results a new vector is built and sensitive information classification or C_IS is made.

This model is part of the PICIS project, Platform for the Identification, Classification and Monitoring of Sensitive Information, which is a winner of the Latin American Cybersecurity Innovation Fund in 2021, sponsored by the OAS, Cisco, and the Citi Foundation. PICIS project is developed by the Cybersecurity Laboratory at CIC, IPN. PICIS represents a security control focused on preventing the sensitive information exposure in documents, in the public media of the federal government of Mexico.

Agradecimientos

A mis directores, Dr. Eleazar Aguirre Anaya y Dr. Francisco Hiram Calvo Castro, por su generosa guía en la realización de este proyecto, por sus invaluables conocimientos y consejos compartidos.

A todos mis profesores y personal administrativo del CIC, por sus transformadoras enseñanzas, amabilidad y orientación.

Al equipo del proyecto PICIS, por su confianza, su ejemplo, por todo el conocimiento compartido y por brindarme la oportunidad de participar en un proyecto internacional.

A mis amigos y compañeros de estudio del CIC, por su motivación, interés, por compartir su experiencia y hacer más llevadera la estancia.

Al Centro de Investigación en Computación, al Instituto Politécnico Nacional, al Consejo Nacional de Ciencia y Tecnología, por su apoyo social, académico y económico, por permitirme continuar con mi formación y alcanzar nuevos objetivos.

Y sobre todo, a mi familia, por su incondicional amor, apoyo y motivación.

Índice

Índice de ilustraciones

Glosario de términos

IA. Inteligencia artificial.
Clasificar. Ordenar o dividir un conjunto de elementos en clases o grupos, a partir de un criterio determinado.
C_DP. Clasificación de datos personales, indica si el texto contiene datos personales.
C_DPI. Clasificación de datos personales identificadores, como la CURP, el RFC, el nombre.
C_DSNU. Clasificación de datos sensibles no unitarios, como los datos de salud o ideología, son no unitarios porque deben estar relacionados con una persona.
C_DSU. Clasificación de datos sensibles unitarios, son sensibles por sí mismos, como el RFC.
C_IS. Clasificación de información sensible, puede ser personal o sensible o no identificada.
Dato. Lo dado. Presentación simbólica de un atributo o variable cuantitativa o cualitativa.
Datos personales. Los concernientes a una persona física identificada o identificable.
Datos sensibles. Los datos personales que por su naturaleza atenten contra las libertades fundamentales o la intimidad y cuyo mal uso provoque discriminaciones o ponerles en grave riesgo.
DT. Árbol de decisiones o *decision tree*.
Encargado. Persona física o moral que trata los datos personales en posesión de un responsable.
Identificar. Distinguir una cosa de otras, por las características que la diferencian.
Información. Dar forma o realidad sustancial a una cosa. Conjunto de datos con significado.
LR. Regresión logística o *logistic regression*.
Metadatos. Datos que describen a otros datos. Grupo de datos que describen el contenido informativo de un objeto al que se denomina recurso. El nombre, tipo, tamaño, ubicación, fecha de creación, autor, son ejemplos de metadatos de un archivo.
ML. Aprendizaje automático o *machine learning*.
nB. Bayes ingenuo o *naive Bayes*.
NER. Reconocimiento de entidades nombradas o *named entities recognition*.
Particular. Responsable del sector privado. Cualquier persona física o moral privada, responsable de los datos personales que se le entregan.
PLN. Procesamiento de lenguaje natural.
Responsable. Persona física o moral a quien se le entregan los datos personales, que decide sobre su tratamiento, qué hacer con ellos y cómo utilizarlos. Ya sea un particular o un sujeto obligado.
Sujeto obligado. Responsable del sector público. Cualquier autoridad, entidad, órgano y organismo de gobierno, órgano autónomo, partido político, fideicomiso y fondo público, responsable de los datos personales que se le entregan.
SVM. Máquina de vectores de soporte o *support vector machine*.
Texto estructurado. Texto contenido en estructuras como tablas, diagramas o cuadros.
Texto no estructurado. Texto escrito en prosa. Párrafos formados por frases y oraciones.
Titular. Persona física a quien pertenecen los datos. El titular acepta compartir sus datos personales con una institución a cambio de un servicio.
Tratamiento. Se refiere a la obtención, almacenamiento, modificación, eliminación y toda operación que se hace sobre los datos personales en una institución.

1. Introducción

Este capítulo incluye la introducción al presente trabajo de investigación, el marco teórico de la protección de datos personales tanto a nivel internacional como nacional y cómo disminuir los riesgos de exposicion. Por último se presenta el problema de investigación.

Con el auge de la digitalización y el intercambio de información, aumenta el riesgo de que la información personal sensible sea expuesta públicamente. Esto afecta tanto a los individuos, titulares de sus datos personales, como a las organizaciones, responsables del tratamiento de estos datos.

Las instituciones tienen la responsabilidad de proteger los datos personales de quienes confían en ellas y deciden compartirles sus datos para recibir un servicio. Las organizaciones están obligadas a proteger los datos personales que poseen, más aún, tratándose de datos personales sensibles. Para disminuir los riesgos de exposición de la información personal sensible es preciso identificarla y clasificarla, así, tanto los responsables como los titulares podrán tomar las medidas correspondientes para mantener la confidencialidad de los datos.

ICIS es un modelo para la identificación de datos personales y la clasificación de información sensible presente en textos no estructurados, escritos en lenguaje natural, en el ámbito de las organizaciones gubernamentales, donde la escritura sigue, generalmente, las reglas gramaticales y ortográficas de la lengua castellana, aunque los documentos pueden ser de formatos y tamaños muy variados.

La tesis forma parte del proyecto PICIS, Plataforma de Monitoreo de Información sensible para entidades del Gobierno Federal de México, que es uno de los doce ganadores del Fondo de Innovación en Ciberseguridad de Latinoamérica en 2021, patrocinado por la OEA, Cisco y la Fundación Citi. El proyecto PICIS es desarrollado por el Laboratorio de Ciberseguridad del CIC, IPN.

En este documento se presenta el desarrollo del modelo ICIS, el cual se llevó a cabo siguiendo la metodología que se describe en la *ilustración 1*. A lo largo de la esta tesis se presentarán cada uno de los pasos de la metodología de investigación, cómo se efectuaron estos pasos y los resultados obtenidos.

Los productos generados por este proyecto son: un artículo aceptado en el congreso CORE 2021, a publicarse en la revista indexada Research in Computing Science; un póster del trabajo “Clasificación de información personal sensible”, persentado en el Décimo tercer Encuentro de la Red de Computación en 2021. En el Anexo 1 se muestran dichos productos.

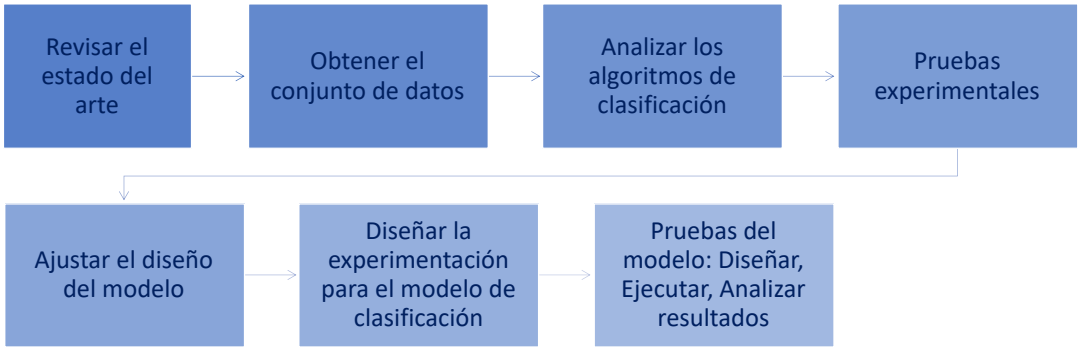


Ilustración 1. Metodología seguida en la investigación

1.1. Protección de datos personales

La protección de las personas con respecto al tratamiento automatizado de los datos personales tiene origen en el Convenio 108 del Consejo de Europa en 1981. Este convenio es una herramienta global para el intercambio efectivo y seguro de información [1], y ha sido adoptado en América por países como Argentina, Uruguay y México, además de los países miembros de la Unión Europea.

Los **datos personales** son los concernientes a una persona física identificada o identificable cuya manifestación sea textual, gráfica, acústica o fotográfica [2]. Los datos personales textuales (numéricos o alfanuméricos) pueden ser de diferentes tipos: identificativos, laborales, académicos, de salud, de patrimonio.

La Comisión Europea define a los **datos personales sensibles** como los datos que por su naturaleza atenten contra las libertades fundamentales o la intimidad, y están sujetos a condiciones de tratamiento específicas [3]. Los siguientes datos personales se consideran sensibles: datos personales que revelen el origen racial o étnico, opiniones políticas, creencias religiosas, filosóficas y morales, la afiliación sindical, datos genéticos, datos biométricos, datos relativos a la salud, datos relativos a la vida sexual o a la orientación sexual de una persona [4].

Uno de los puntos principales del Convenio 108 es garantizar la confidencialidad de los datos personales sensibles por parte de las organizaciones responsables de ellos. Las organizaciones deben brindar protección a la información dependiendo de la sensibilidad, valor y criticidad de ésta [5].

1.2. Protección de datos personales en México

En la *ilustración 2* se muestra la línea de tiempo de la legislación en materia de protección de datos personales en México.

En 2002 se crea la Ley Federal de Transparencia y Acceso a la Información Pública Gubernamental LFTAIPG, así como el Instituto Federal de Acceso a la Información IFAI.

En 2007 se modifica el artículo 6 de la Constitución Política de los Estados Unidos Mexicanos, "La información que se refiere a la vida privada y los datos personales será protegida en los términos y con las excepciones que fijen las leyes."

A partir de 2008 se modifican los artículos de la Constitución 16 y 73, señalando que "Toda persona tiene derecho a la protección de sus datos personales, así como el derecho de acceder a los mismos y, en su caso, obtener su rectificación, cancelación y manifestar su oposición en los términos que fijen las leyes." y que "El Congreso tiene facultad para legislar en materia de protección de datos personales en posesión de particulares…, para expedir las leyes generales reglamentarias que desarrollen los principios y bases en materia de transparencia gubernamental, acceso a la información y protección de datos personales en posesión de las autoridades, entidades, órganos y organismos gubernamentales de todos los niveles de gobierno".

En 2010 se crea la Ley Federal de Protección de Datos Personales en Posesión de Particulares (LFPDPPP), aplicable a las personas físicas o morales de carácter privado.

En 2015 se sustituyen la LFTAIPG por la Ley General de Transparencia y Acceso a la Información Pública LGTAIP, y el IFAI por el Instituto Nacional de Transparencia, Acceso a la Información y Protección de Datos Personales INAI.

En 2017 se crea la Ley General de Protección de Datos Personales en Posesión de sujetos obligados (Ley General), aplicable a cualquier autoridad, entidad, órgano y organismo de los poderes ejecutivo, legislativo y judicial, órganos autónomos, partidos políticos, fideicomisos y fondos públicos.

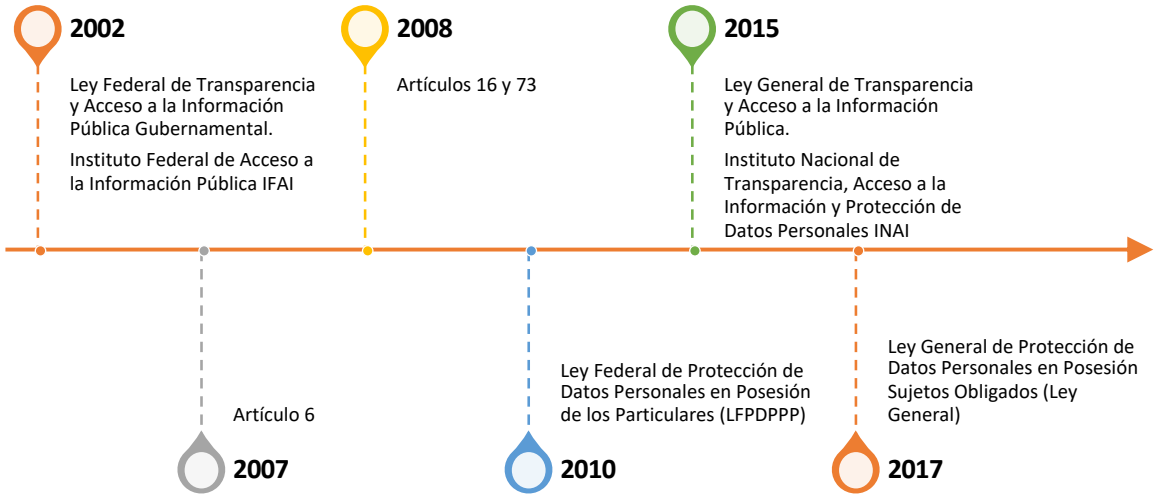


Ilustración 2. Línea de tiempo de la legislación en materia de Protección de datos personales en México

Tanto la LFPDPPP como la Ley General, regulan el tratamiento de los datos personales. El tratamiento se refiere a la obtención, almacenamiento, modificación, eliminación, divulgación y cualquier operación sobre los datos personales. Las personas involucradas en el tratamiento de los datos personales son el titular, el responsable y el encargado, como se describe en la *ilustración 3*.

Titular. Es la persona física a quien pertenecen los datos personales.

Responsable. Es la persona física o moral con quien el titular comparte sus datos a cambio de un servicio. El responsable decide sobre el tratamiento de los datos personales. Ya sea un particular o un sujeto obligado por la Ley General.

Encargado. Es la persona física o moral que trata los datos personales.

Ilustración 3. Personas involucradas en el tratamiento de datos personales

Las diferentes legislaciones en materia de protección de datos personales garantizan los derechos de las personas, titulares de sus datos, a la privacidad y obligan a los responsables y encargados, tanto en las organizaciones públicas como privadas, a proteger los datos personales que tratan.

En el caso de los responsables en las organizaciones públicas, el artículo 23 de la LGTAIP señala que los sujetos obligados deben "transparentar y permitir el acceso a su información y proteger los datos personales que obren en su poder".

1.3. Riesgos de exposición de la información sensible

Al transparentar la información y durante el tratamiento de los datos personales, existe el riesgo de que los datos personales sensibles sean expuestos. Las organizaciones de gobierno cuentan con sitios públicos de internet donde, en cumplimiento con su obligación de transparencia, publican documentos que, de manera no intencional, pudieran contener información sensible expuesta.

Las consecuencias de la exposición de información sensible llegan a ser graves, tanto para los titulares, como para los responsables, los encargados y el mismo gobierno; no solo en la

economía[1] de los responsables y encargados, sino en la privacidad de los titulares y en la pérdida de la confianza en las instituciones.

Por otro lado, como se mencionó en este capítulo, existen convenios internacionales y legislaciones nacionales para el intercambio seguro de información y para garantizar a cualquier persona física el respeto de sus derechos y libertades fundamentales, concretamente su derecho a la vida privada, con respecto al tratamiento automatizado de los datos de carácter personal correspondientes a dicha persona.

Un mecanismo de seguridad que disminuya los riesgos de exposición de la información personal sensible en medios públicos es identificar la información personal y clasificarla de acuerdo con su sensibilidad, utilizando técnicas de inteligencia artificial (IA), como el procesamiento de lenguaje natural (PLN) y aprendizaje automático (*machine learning* ML), para hacerlo de forma automática.

1.4. Identificación

En los textos publicados por los responsables se pueden identificar los datos personales, utilizando técnicas de PLN, como la extracción de datos.

El PLN es una rama de la IA y la extracción de datos en textos es una de sus tareas. La extracción de datos permite obtener datos significativos del texto, como fechas, entidades, números, moneda, entre otros. Algunas técnicas para extraer datos son las expresiones regulares y las dependencias gramaticales.

Las expresiones regulares son patrones conformados por secuencias de caracteres que permiten buscar combinaciones que coincidan con ese patrón, en una cadena de texto [6].

Las dependencias gramaticales son patrones conformados por secuencias de palabras, de acuerdo con su categoría gramatical (*part of speech tagging* POST). Las dependencias gramaticales permiten buscar combinaciones de palabras que coincidan con ese patrón en una cadena de texto.

En este trabajo, se utilizan técnicas de PLN, como las expresiones regulares y las dependencias gramaticales para identificar los datos personales en un texto, como se describe en las secciones 4.2 y 4.4.5.

[1] La LFPDPPP [16] establece que las multas para los responsables de los Datos Personales van desde los 100 hasta 320,000 días de salario mínimo vigente en la CDMX; en el caso de Datos Personales Sensibles las sanciones podrán incrementarse hasta por dos veces los montos establecidos [17].

1.5. Aprendizaje automático y clasificación

Una vez que se identifican los datos personales en los textos, estos se pueden interpretar como información sensible gracias al uso de técnicas de ML.

Una rama de la IA es el ML, que es convertir los datos en información, dar sentido a los datos. A partir de un conjunto de datos ejemplo y mediante ciertos algoritmos, la máquina aprende, encontrando alguna relación entre los datos con los que fue entrenada. Con más datos y entrenamiento, los resultados del ML son mejores, así como los humanos mejoran con la práctica.

La clasificación es una de las tareas del ML, se utiliza para colocar un elemento desconocido en un grupo conocido, a partir de un conjunto de datos de entrenamiento.

En el ML, el conjunto de datos está formado por elementos o instancias de datos. Las instancias suelen ser renglones en el conjunto de datos. Las características o atributos son las medidas individuales que, cuando se combinan con otras características, forman una instancia. Las características suelen ser columnas en un conjunto de datos.

En la clasificación, cada instancia de datos se compone de una serie de características y una variable objetivo que indica la clase a la que pertenece esa instancia. La máquina aprende, encontrando patrones en las características y la clase. La tarea de la clasificación es predecir a qué clase debe pertenecer una instancia de datos.

Hay algoritmos de ML para la clasificación, como las máquinas de vectores de soporte (SVM), Bayes ingenuo (naive Bayes nB), k vecino más próximo (k nearest neighbor kNN) o árbol de decisiones (decision tree DT).

Para probar los algoritmos de ML, generalmente se tiene un conjunto de datos de entrenamiento y un conjunto de datos separado, llamado conjunto de prueba. Inicialmente, el programa se alimenta con ejemplos de entrenamiento; aquí es cuando se lleva a cabo el aprendizaje automático. A continuación, el conjunto de prueba alimenta al programa. La variable de destino para cada ejemplo del conjunto de prueba no se le da al programa, y el programa decide a qué clase debe pertenecer cada ejemplo. Luego, la variable objetivo, o la clase a la que pertenece el ejemplo, se compara con el valor predicho y podemos tener una idea de qué tan preciso es el algoritmo [7].

En este trabajo se utilizan algoritmos de clasificación de ML para que los datos personales, identificados en los textos, sean clasificados como información sensible. El conjunto de datos está basado en los posibles datos personales a identificar en un texto. El diseño de la clasificación de información sensible se describe en las secciones 4.3, 4.4.7 y 4.4.8.

1.6. Problema de investigación

Los datos personales y los datos personales sensibles son difíciles de identificar en el contenido textual de un documento, porque no sólo depende de que los textos contengan ciertas palabras o se apeguen a un formato, sino también del contexto en que están redactados, para clasificarse como información sensible. La identificación de los datos personales está relacionada completamente con el contexto del discurso, de no considerarse, propicia la ocurrencia de falsos positivos.

Los datos personales contenidos en los textos pueden ser identificados, considerando el contexto, mediante técnicas de procesamiento de lenguaje natural, y clasificados como información sensible, utilizando técnicas de aprendizaje automático.

El objetivo del presente trabajo es diseñar un modelo para la identificación y clasificación de información personal sensible en textos, mediante técnicas procesamiento de lenguaje natural y de aprendizaje automático, para prevenir su exposición en espacios públicos de sujetos obligados.

2. Estado del arte de la clasificación de información sensible

La clasificación de información sensible se abordó como un problema en 2005 [8] y desde entonces se han propuesto numerosos métodos para resolverlo. Los métodos para la clasificación de información sensible que actualmente han logrado mayor precisión son los basados en ML y los basados en aprendizaje profundo (deep learning DL).

Abhinav Nagpal et al. [9] reportaron, en 2022, la clasificación detallada o de grano fino de entidades de datos personales con modelos de lenguaje. Clasificaron entidades de datos personales (personal data entities PDE) con redes neuronales (neural networks NN) y con DistilBERT. Agregaron tipos de entidades de grano grueso y utilizaron reconocimiento de entidades nombradas (named entity recognition NER) y análisis gramatical (part of speech POS). Utilizan más de 100 clases de PDE, sin distinguirlas como datos sensibles.

Ji-Sung Park et al. [10] desarrollaron, en 2020, un sistema de prevención de pérdida de datos (data loss prevention DLP), identificando datos sensibles con NER en textos estructurados. Aunque este artículo está orientado a textos estructurados, se considera el uso de NER para identificar entidades como nombres, lugares, organizaciones que podrían clasificarse como información sensible.

En 2019, Huimin Jiang et al. [11] hicieron una clasificación de datos médicos sensibles, basada en la clasificación de textos, su experimento utilizó 736 muestras de textos de registros clasificados como sensibles y normales, obteniendo un 90% de exactitud con el algoritmo de SVM, 85% con Bayes ingenuo (nB) y 80% con el algoritmo k vecino más próximo (kNN). Su proceso consistió en un preprocesamiento con segmentación de palabras y selección de características basada en la frecuencia, no se especifica el tamaño de los vectores ni de los textos. En este trabajo se muestra la factibilidad de utilizar clasificación de textos para clasificar datos médicos sensibles. Sin embargo concluyen que, para elegir el mejor clasificador, se deben hacer investigaciones específicas.

El mismo año, Gousheng Xu et al. [12] utilizaron redes neuronales de circonvolución (CNN) para detectar información sensible en textos no estructurados con un 95% de exactitud y del 94%, utilizando redes neuronales recurrentes (RNN). El conjunto de datos contiene textos sensibles de su propio laboratorio del ámbito político y militar, los textos no sensibles son el conjunto de datos de noticias de Sohu, en total 14,000 elementos. Para el preprocesamiento utilizaron eliminación de imágenes, símbolos e información irrelevante, segmentación de textos y vectorización con 50 a 200 características

En 2019, también, Yan Liang et al. [13] utilizaron el aprendizaje incremental y comparación de similitud (ILSC), logrando un 86% de exactitud, 84% con bosque aleatorio en línea (ORF), 82% con nB y SVM incremental (ISVM) con 87%. Con un conjunto de datos de 6,330 documentos etiquetados como restringido, confidencial, secreto y ultra secreto, de acuerdo con las oraciones sensibles y etiquetas de seguridad contenidas en ellos. No se indica el tamaño de los vectores ni de los documentos.

Graham McDonald et al. [14]en 2017 evaluaron SVM para la clasificación de sensibilidad utilizando POS, obteniendo una efectividad del 90%. En 2015 habían utilizado n-gramas de clases gramaticales, obteniendo una efectividad del 84% de precisión [15]. El uso de n-gramas y POS para el preprocesamiento es una referencia para identificar datos personales.

Los trabajos anteriores a 2022 corresponden a diferentes ámbitos específicos, como el médico, el organizacional o el militar. En ellos, los métodos que lograron mayor precisión en la clasificación de infomación sensible de documentos, son los basados en CNN, con un conjunto de datos de entrenamiento de 14,000 elementos, que es el mayor de los trabajos aquí analizados. En estos trabajos se reporta la clasificación binaria y multiclase de datos sensibles, mediante ML. Utilizan identificación de datos para la extracción de características de los vectores. Solo Nagpal reporta una clasificacion detallada de entidades de datos personales, sin considerar su sensibilidad.

El presente trabajo muestra el diseño de un modelo para clasificar información personal sensible, identificando los distintos tipos de datos personales en cada uno de los segmentos de texto de un documento, utilizando PLN y ML. Se retoma el uso de las técnicas segmentación de textos y eliminación de palabras auxiliares en el preprocesamiento; de categorías que agrupan diferentes tipos de datos personales; de técnicas de PLN, como NER, POS y diccionarios para la identificación de datos personales; de la vectorización o extraccion de características; así como del uso de algoritmos de ML para clasificar los datos como información sensible, en textos no estructurados de sujetos obligados, donde la escritura es de carácter formal y los documentos son de tipos y tamaños variados.

3. Pruebas experimentales

Las pruebas experimentales consistieron en la obtención de un conjunto de datos, el análisis de los algoritmos de aprendizaje automático y las pruebas experimentales. En este capítulo se muestran los resultados de estos pasos.

3.1. Clasificación de información sensible en textos

Las pruebas experimentales ofrecen una clasificación, con algoritmos de aprendizaje automático, que indica si un texto contiene o no información sensible. La clasificación se hizo en un conjunto de datos formado por textos cortos.

En esta sección se describe la primera experimentación de este proyecto para clasificar información sensible en textos.

3.2.. Conjunto de datos

El primer ejercicio de clasificación de información sensible se hizo en un conjunto de datos formado a partir de 160 textos cortos tomados de redes sociales, el 60% de ellos contenía información personal sensible, el 40% no.

El conjunto de datos se almacenó en un archivo de texto separado por tabuladores, con el formato: identificador -> texto -> etiqueta. Los textos se etiquetaron con 1 si el texto contenía información sensible, con 0 en caso contrario.

3.3. Preprocesamiento

Los textos fueron preprocesados utilizando algunas técnicas de PLN, tales como la normalización y la eliminación de palabras auxiliares de la biblioteca Stanza.

La normalización consiste en convertir las palabras en su forma normal, es decir, como aparecen en el diccionario.

La eliminación de las palabras auxiliares, descarta las palabras que no aportan significado a un texto, como los artículos, las preposiciones, los pronombres.

3.4. Clasificación mediante aprendizaje automático

Para la clasificación se utilizó el modelo de espacio vectorial bolsa de palabras (BOW por sus siglas en inglés), los valores de los vectores fueron asignados por la frecuencia de términos o Tf por sus siglas en inglés. El conjunto de datos se dividió en 75% entrenamiento, 25% pruebas.

Se utilizaron los algoritmos de aprendizaje automático naive Bayes (NB), regresión logística (LR) y máquina de vectores de soporte (SVM), de la biblioteca Sklearn.

3.5. Pruebas

Los resultados de los experimentos fueron cercanos a lo reportado en el estado del arte, como se muestra en la *ilustración 4*.

RL	SVM	NB
0.840	0.853	0.839

Ilustración 4. Exactitud de los 3 algoritmos de aprendizaje automático del experimento inicial

Es importante el preprocesamiento de los textos con técnicas de PLN, para reducir el tamaño de los vectores.

Los resultados de los experimentos fueron cercanos a lo que se reporta en el estado del arte, con los mismos algoritmos de aprendizaje automático.

No obstante, al analizar estos resultados, se plantean varias interrogantes.

- ¿Qué información es sensible?
- ¿Hay diferentes tipos de datos sensibles?
- ¿Cómo detectar información sensible?
- ¿Qué tipos de archivos serán clasificados?
- ¿Sólo archivos de contenido textual?
- ¿La precisión se podría mejorar?

En respuesta a estas interrogantes se propone el diseño del modelo ICIS, que se describe en el capítulo 4.

4. Modelo ICIS

Tras analizar los resultados de las pruebas experimentales, se procedió a ajustar el diseño. En este capítulo se muestra el diseño del modelo de identificación y clasificación de información sensible ICIS.

En la sección 4.1 se presenta el análisis derivado de las pruebas exploratorias y que sustenta el diseño del modelo.

En la sección 4.2 se detallan los elementos que se diseñaron para identificar datos personales y clasificarlos como información sensible.

En la sección 4.3 se describe el modelo propuesto para la Identificación y clasificación de información sensible ICIS y cada uno de los módulos que lo componen.

4.1. Análisis

Es necesario considerar que los textos a procesar son los contenidos en los documentos disponibles en los sitios web de las instituciones, en los que se espera que tengan un tipo de redacción formal, apegado a las normas lingüísticas del idioma español. Los documentos están escritos en lenguaje natural no estructurado y son de tamaños y formatos variados.

Además de clasificar la información como sensible, se requiere identificar los datos personales que contienen los documentos, indicando en qué parte del documento se encuentran y por qué el texto contiene información sensible.

Es necesario determinar cuáles son los datos personales a identificar en los documentos y de qué manera detectarlos. Si los datos no identifican ni hacen identificable a una persona, no son datos personales; También se requiere analizar cómo clasificar esos datos como información sensible. Si el dato personal está relacionado a la esfera más íntima de la persona se considera sensible.

Y sobre todo, se requiere considerar el contexto de los datos. Así, un documento que contiene un dato sensible, pero sin relacionarlo con una persona en particular, como en un texto informativo, no es considerado como información sensible. Para ser información sensible, en el texto debe existir un dato sensible relacionado con una persona en particular, ya sea por su nombre o por algún número de identificación.

Por otro lado, el análisis gramatical también nos ayuda a contextualizar la información para saber si un número o una palabra son un dato personal en particular, no solo por su formato, sino también por la estructura gramatical en la que se encuentran.

4.2. Elementos para la identificación de datos personales

En esta sección se presentan los elementos que fueron diseñados para la Identificación de datos personales, que son la taxonomía y las técnicas de procesamiento de lenguaje natural PLN para su identificación.

4.2.1. Diseño de la taxonomía

Se propone una taxonomía con los tipos de datos personales a identificar en un documento. La taxonomía de datos personales propuesta en la *Ilustración 5* se basa en los datos de identificación y los datos enunciados como sensibles. También incluyen los correspondientes a los servicios ofrecidos por las instituciones, por ejemplo la educación, la salud, la movilidad.

Taxonomía de datos personales									
Identifiativos	**Electrónicos**	**Laborales**	**Tránsito/ Migratorio**	**Patrimonio**	**Salud**	**Académicas**	**Ideológicos**	**Intimidad**	**Características físicas**
Nombre	Email	Empresa	Pasaporte	Sueldo	Estado de salud	Escuela	Religión	Preferencia sexual	Iris
Dirección	Contraseña	Puesto	Licencia	Impuestos	Historial clínico	Calificación	Afiliación sindical	Hábito	ADN
Fecha nacimiento	Usuario	Dirección laboral	Visa	Créditos	Enfermedad	Título	Preferencia política	Relación personal	Color de piel
RFC		Email laboral	Placas	Número de tarjeta	Tratamiento	Certificado	Organización civil		Huella dactilar
CURP			NIV	Inversiones	Estudio clínico	Número de cédula			Cicatriz
Teléfono				Afore	Alergia				Tipo de sangre
Firma				Seguros	Condición psicológica				Peso
INE					NSS				Altura

Total tipos de datos:	55
Total de categorías:	10

Ilustración 5. Taxonomía de los datos personales.

La taxonomía incluye 55 tipos de datos personales, agrupados en 10 categorías. Esta es la taxonomía utilizada para identificar los datos personales y los datos personales sensibles en un texto. Sin embargo, es posible agregar más tipos de datos a la taxonomía o adaptarla.

4.2.2. Técnicas para la identificación de datos personales

Para poder identificar los 55 tipos de datos personales, se utilizan técnicas de PLN como expresiones regulares, diccionarios, análisis gramatical y formatos de archivos. En la *ilustración 6* se muestran las técnicas utilizadas para la identificación de los datos personales.

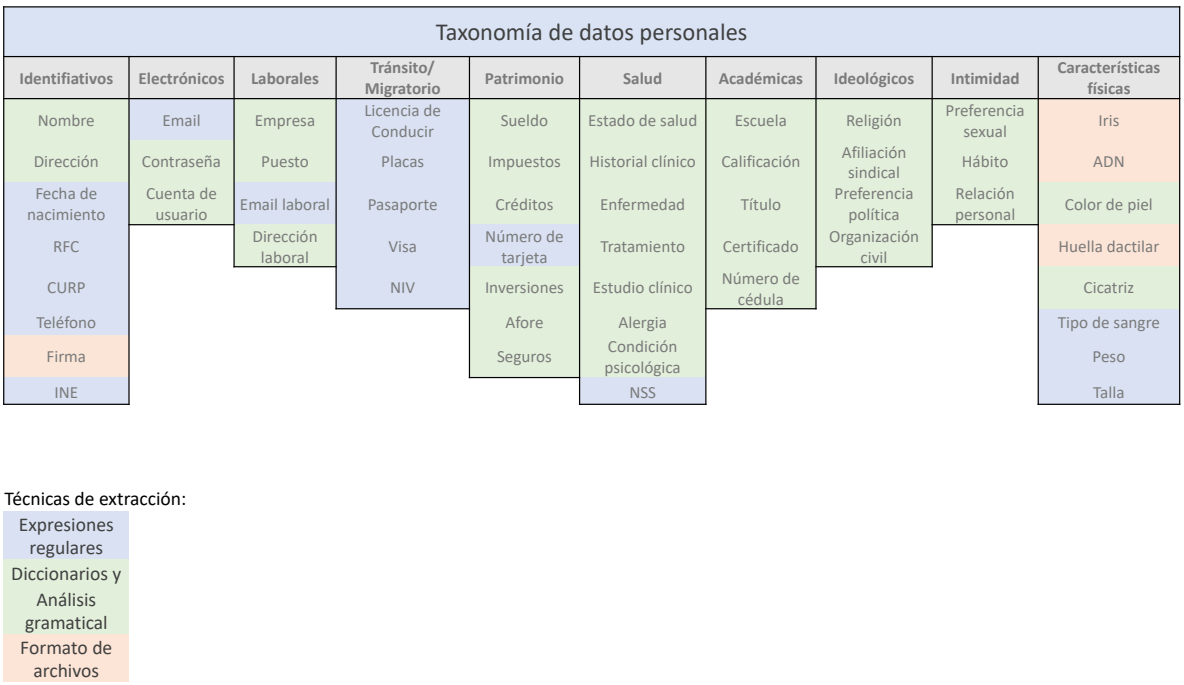


Ilustración 6. Técnicas utilizadas para la identificación de datos personales.

Las expresiones regulares se utilizan para identificar datos con un formato definido, tales como CURP, RFC, números telefónicos, correos electrónicos o placas. Las expresiones regulares se diseñaron de acuerdo con la normatividad nacional.

Los diccionarios se emplean para identificar tipos de datos que utilizan alguna de las palabras dentro de un conjunto definido, como los nombres, religiones, preferencias políticas.

El análisis gramatical, se utiliza para identificar datos en los segmentos de texto con cierto patrón gramatical, como estados de salud, alergias, hábitos o relaciones personales. El análisis gramatical es en lengua castellana.

Para considerar el contexto, se utiliza el análisis de dependencias gramaticales en conjunto con las expresiones regulares, pues no solo se identifican los formatos de los datos, sino la estructura gramatical en la que se encuentran. Por ejemplo, si solo se consideraran las expresiones regulares, la cadena “DEL 2022”, sería identificada como una placa de automóvil, cuando en realidad es parte de una fecha. Por eso es importante considerar el contexto y esto se logra agregando a las expresiones regulares el análisis de dependencias gramaticales. A los identificadores por diccionario, también se les agrega el análisis gramatical, pues es importante considerar el contexto, no solamente las palabras que contiene.

4.3. Elementos para la clasificación de datos personales y de información sensible

En esta sección se presentan los elementos diseñados para la Clasificación de información sensible, como son los vectores de datos personales, las reglas de clasificación de datos personales y de información sensible, así como las etiquetas de clasificación. Estos módulos forman parte de la arquitectura general del modelo ICIS, que se describe en la sección 4.4.

4.3.1. Vector de datos personales

El vector de entrada es una combinación de 55 características binarias, correspondientes a los 55 tipos de datos de la taxonomía de datos personales. Las características son binarias porque indican si un texto contiene o no un tipo de dato personal. Por ejemplo, el texto:

"Hola, soy Juan Pérez López, mi curp es PELJ800709HDFK01 y mi tipo de sangre es O+."

tiene un vector asociado de 55 características binarias con el valor 1 en las posiciones 0, 4 y 52, y con el valor 0 en el resto de las posiciones [1,0,0,0,1,0,0 …, 0,1,0,0], pues contiene un nombre, una CURP y un tipo de sangre, como se muestra en la *Ilustación 7*.

NOMBRE	DIRECCION	FECHA NAC.	RFC	CURP	TELEFONO	INE	FIRMA	EMAIL	CONTRASEÑA	USUARIO	EMPRESA	PUESTO	…	SANGRE	PESO	ALTURA
1	0	0	0	1	0	0	0	0	0	0	0	0	…	1	0	0

Ilustración 7. Vector de datos personales con 55 características binarias.

En el Anexo 2 se muestra la lista ordenada de los 55 tipos de datos, cada uno de ellos corresponde a una característica del conjunto de datos: la posición 0 al nombre del titular, la 1 a la dirección, y así hasta la posición 54, que corresponde a la altura.

4.3.2. Clasificaciones de datos personales

Las clasificaciones de datos personales son la base para la clasificación de información sensible, considerando el contexto.

Al analizar los 55 tipos de datos, encontramos algunos que son sensibles, como los de las categorías de salud o ideología, de acuerdo a lo definido por la Comision Europea [4], porque se refieren a la esfera más íntima del titular o ponen en riesgo sus derechos y libertades. Para considerar que un texto contiene información sensible, los datos sensibles no están aislados, pertenecen a un titular. Los titulares son identificados directamente por su nombre o un número de identificación oficial. Algunos de estos identificadores oficiales contienen información sensible por sí mismos. Por otro lado, se considera que hay textos

con información personal que no es sensible, que contienen datos personales identificables, porque no identifican directamente a una persona, pero son atribuidos a un titular, en el contexto.

Por estas razones, se proponen las clasificaciones de Datos Personales Identificadores (C_DPI), Datos Sensibles Unitarios (C_DSU), Datos Sensibles No Unitarios (C_DSNU) y Datos Personales (C_DP).

Cada vector de datos personales es clasificado en estas cuatro formas. En las secciones 4.3.2.1 a la 4.3.2.4 se especifican los cuatro clasificadores de datos personales.

4.3.2.1. C_DSU o Clasificador de Datos Sensibles Unitarios

Algunos tipos de datos personales son por sí mismos sensibles, como el RFC, la CURP, etc., porque, accediendo a ciertas herramientas públicas, es posible obtener información confidencial de una persona. A este tipo de datos se les ha denominado, para este proyecto, Datos Sensibles Unitarios o DSU.

El primer clasificador C_DSU es para saber si un vector contiene DSU. Las etiquetas son DSU, NO_DSU o NI.

- DSU o Dato Sensible Unitario. Indica que el texto sí contiene un DSU. Si el texto contiene alguno de los siguientes tipos de datos: RFC, CURP, INE, PASAPORTE, VISA, LICENCIA, VISA, PLACA, NIV, NSS, CEDULA, TARJETAS, SEGURO, AFORE, entonces se asigna la etiqueta DSU.

- NO_DSU. Indica que el texto No contiene un DSU. Si el texto contiene algún otro dato personal, es decir, si hay un valor 1 en cualquiera de las características restantes del vector, distintas a los DSU, entonces se asigna la etiqueta NO_DSU.

- NI. Indica que contiene datos No Identificados. Si el texto no contiene ninguno de los tipos de datos personales indicados en el vector, es decir, si todos sus valores son 0, entonces se asigna la etiqueta NI.

El vector del ejemplo contiene una CURP, por lo tanto, tendrá la etiqueta DSU en el clasificador C_DSU, como se muestra en la *Ilustración 8*.

NOMBRE	DIRECCION	FECHA NAC.	RFC	CURP	TELEFONO	INE	FIRMA	EMAIL	CONTRASEÑA	USUARIO	EMPRESA	PUESTO	...	SANGRE	PESO	ALTURA	C_DSU
1	0	0	0	1	0	0	0	0	0	0	0	0	...	1	0	0	DSU

Ilustración 8. Ejemplo de Clasificación de Datos sensibles unitarios C_DSU.

4.3.2.2. C_DPI o Clasificador de Datos Personales Identificadores

Los tipos de datos personales con los que se identifica directamente o indirectamente a una persona, como su nombre o un número de identificación, se han denominado Datos Personales Identificadores o DPI.

El segundo clasificador C_DPI es para saber si un vector contiene DPI. Las etiquetas son DPI, NO_DPI o NI.

- DPI o Dato Personal Identificador. Indica que el texto sí contiene un DPI. Si el texto contiene al menos uno de los siguientes datos: NOMBRE, RFC, CURP, INE, PASAPORTE, VISA, LICENCIA, VISA, PLACA, NIV, NSS, CEDULA, TARJETAS, SEGURO, AFORE, entonces se asigna la etiqueta DPI.

- NO_DPI. Indica que el texto No contiene un DPI. Si el texto contiene algún otro dato personal que no sea DPI, es decir, si hay un valor 1 en cualquiera de las características restantes, distintas a los DPI, entonces se asigna la etiqueta NO_DPI.

- NI. Indica que los datos son No Identificados. Si el texto no contiene ninguno de los tipos de datos personales indicados en el vector, entonces se asigna la etiqueta NI.

Al vector del ejemplo, por contener un nombre y una CURP, se le asignará la etiqueta DPI en el Clasificador C_DPI, como se muestra en la *Ilustración 9*.

NOMBRE	DIRECCION	FECHA NAC.	RFC	CURP	TELEFONO	INE	FIRMA	EMAIL	CONTRASEÑA	USUARIO	EMPRESA	PUESTO	...	SANGRE	PESO	ALTURA	C_DPI
1	0	0	0	1	0	0	0	0	0	0	0	0	...	1	0	0	DPI

Ilustración 9. Ejemplo de Clasificación de Datos Personales Identificadores C_DPI.

4.3.2.3. C_DSNU o Clasificador de Dato Sensible No Unitario

Los datos personales sensibles son los que tienen que ver con la intimidad de una persona, podrían ponerla en riesgo o discriminarla. Principalmente los que corresponden a las categorías de datos ideológicos, intimidad, características físicas y salud. A estos datos se les ha denominado Datos Sensibles No Unitarios o DSNU. Son No Unitarios pues por sí mismos no son sensibles, para serlo requieren asociarse a una persona mediante un identificador.

El tercer clasificador C_DSNU es para saber si un vector contiene DSNU. Las etiquetas de clasificación son DSNU, NO_DSNU o NI.

- DSNU. Indica que el texto sí contiene un DSNU. Si el texto contiene al menos uno de los siguientes tipos de datos: SUELDO, IMPUESTOS, EDOSALUD, HCLINICO, ENFERMEDAD, TRATAMIENTO, ESTUDIOCLINICO, ALERGIA, CPSICOLOGICA, RELIGION, ASINDICAL, PPOLITICA, ORGCIVIL, PSEXUAL, HABITO, RPERSONAL, IRIS, ADN, COLORPIEL, HUELLAD, CICATRIZ, SANGRE, PESO, ALTURA, FECHANAC, entonces se asigna la etiqueta DSNU.

- NO_DSNU. Indica que el texto No contiene un DSNU. Si el texto contiene algún otro tipo de dato personal, que no sea DSNU, es decir, si hay un valor 1 en cualquiera de las características restantes, distintas a los DSNU, entonces se asigna la etiqueta NO_DSNU.

- NI. Indica que el texto contiene datos No Identificados. Si el texto no contiene ninguno de los tipos de datos personales correspondientes con el vector, entonces se asigna la etiqueta NI.

El vector del ejemplo contiene un tipo de sangre, por lo tanto, tendrá la etiqueta DSNU, en el clasificador C_DSNU, como se muestra en la *Ilustración 10*.

NOMBRE	DIRECCION	FECHA NAC.	RFC	CURP	TELEFONO	INE	FIRMA	EMAIL	CONTRASEÑA	USUARIO	EMPRESA	PUESTO	...	SANGRE	PESO	ALTURA	C_DSNU
1	0	0	0	1	0	0	0	0	0	0	0	0	...	1	0	0	DSNU

Ilustración 10. Ejemplo de Clasificación de Datos Sensibles No Unitarios C_DSNU.

4.3.2.4. C_DP o Clasificador de Datos Personales

Los tipos de datos personales que corresponden a una persona, como su dirección o un número de identificación, se han denominado Datos Personales o DP.

El cuarto clasificador C_DP es para saber si un vector contiene DP. Las etiquetas son DP y NO_DP.

- DP o Dato Personal. Indica que el texto sí contiene un DP. Si el texto contiene cualquiera de los 55 tipos de datos personales, entonces se asigna la etiqueta DP.

- NO_DP. Indica que el texto No contiene un DP. Si el texto no contiene ningún dato personal DP, se asigna la etiqueta NO_DP.

Al vector del ejemplo, por contener un nombre, una CURP y un tipo de sangre, se le asignará la etiqueta DP en el Clasificador C_DP, como se muestra en la *Ilustración 11*.

NOMBRE	DIRECCION	FECHA NAC.	RFC	CURP	TELEFONO	INE	FIRMA	EMAIL	CONTRASEÑA	USUARIO	EMPRESA	PUESTO	...	SANGRE	PESO	ALTURA	C_DP
1	0	0	0	1	0	0	0	0	0	0	0	0	...	1	0	0	DP

Ilustración 11. Ejemplo de Clasificación de Datos Personales C_DP.

4.3.3. Clasificación de información sensible

Con los resultados de las cuatro clasificaciones se forma un nuevo vector de cuatro características correspondientes a los cuatro clasificadores de datos personales, para la clasificación de la información sensible, como se muestra en la *Ilustración 12*.

C_DPI	C_DSU	C_DSNU	D_DP	C_IS
DPI	DSU	DSNU	DP	IPS

Ilustración 12. Vector de Clasificación de Información Sensible.

En el modelo ICIS se proponen reglas para clasificar la información y etiquetarla como información personal sensible IPS, como información personal IP o como información no identificada NI.

4.3.3.1. Información personal sensible IPS

Un texto tiene Información Personal Sensible si en su contenido existe la combinación de un Dato Personal Identificador con algún Dato Personal Sensible, o un Dato Sensible Unitario combinado o no con algún Dato Personal.

IPS = (C_DPI=DPI ^ C_DSNU=DSNU) V (C_DSU=DSU) V (C_DSU=DSU ^ C_DP=DP)

Donde:

IPS: Contiene información personal sensible
DPI: Contiene algún dato personal identificador
DSNU: Contiene un dato sensible no unitario
DSU: Contiene un dato sensible unitario
DP: Contiene un dato personal

El algoritmo OCAT con heurísitca AR1 obtiene una regla equivalente, e incluso reducida, para la información sensible:

$$IPS = DSU \vee DPI \wedge DSNU.$$

4.3.3.2. Información personal IP

Un texto tiene información personal si contiene un Dato Personal Identificador con un Dato Personal.

$$IP = (C\_DPI=DPI) \ \hat{} \ (C\_DP=DP) \ \hat{} \ \neg IPS$$

4.3.3.3. Información no identificada NI

Si un texto no contiene información personal, ni información personal sensible, se etiqueta como no identificada.

$$NI = \neg IP \ \hat{} \ \neg IPS$$

Al vector del ejemplo, por contener un DPI con un DSNU, o bien, por contener un DSU, se le asignará la etiquea IPS, que indica que el texto contiene información personal sensible.

4.4. Diseño del modelo ICIS

Una vez descritos los elementos necesarios para identificar los datos personales y clasificar la información sensible, se presenta en la *Ilustración 13* el modelo de Identificación de Datos Personales y Clasificación de Información Sensible ICIS.

El modelo procesa un Archivo de entrada con contenido textual. Como resultado del procesamiento, se obtiene un objeto json con los hallazgos de datos personales en el texto, es decir, los tipos de datos identificados y su ubicación dentro del documento, así como los resultados de la clasificación de información sensible.

En esta sección se describen cada uno de los módulos que componen al modelo ICIS.

Ilustración 13. Modelo ICIS para identificar datos personales y clasificar información sensible.

4.4.1. Archivo de entrada

La entrada del modelo es la ruta con la ubicación de un archivo con contenido textual en alguno de los siguientes formatos: pdf, docx, xlsx, pptx, o de texto plano como txt.

4.4.2. Separación

Es un módulo dentro del modelo ICIS, cuyo objetivo es separar los metadatos y el contenido textual de los archivos. Su entrada es la ruta del archivo a analizar, su salida es la tupla (metadatos, contenido textual).

El archivo se separa en contenido textual y metadatos. Sin importar el formato de los archivos, siempre se extraen sus metadatos. Para obtener su contenido textual, deben ser archivos de tipo pdf, docx, xlsx, xls, pptx y texto plano, como txt, de lo contrario, el contenido textual estará vacío.

La función de Separación de un archivo devuelve una tupla metadatos, contenido.

metadatos, contenido = separador_datos_metadatos(ruta)

Donde:

ruta: Contiene la ubicación del archivo en una cadena de texto.
metadados: Contiene los metadatos del archivo, en una lista de elementos de la forma "llave: valor" (diccionario), cuyas llaves representan los tipos de metadatos del archivo. Si los metadatos no son leídos, porque el archivo está dañado, devuelve un objeto None.
contenido: Es una cadena del texto contenido en el archivo. Representa el texto del documento, se le han agregado marcadores para su posterior segmentación en páginas. Si el archivo no contiene texto, o no es leído, devuelve un objeto None.

4.4.3. Segmentación y análisis contextual

Segmentación es un módulo de ICIS cuyo objetivo es segmentar los textos en páginas, párrafos y oraciones.

Una vez que se tiene el contenido textual, éste es segmentado a diferentes niveles, como páginas, párrafos u oraciones. El texto es segmentado por niveles con dos fines: para conocer la ubicación de los datos personales identificados y para facilitar el análisis contextual.

El análisis contextual basado en la segmentación nos permite relacionar, por ejemplo, un Dato Personal Identificador con un Dato Sensible No Unitario en una misma oración, en un mismo párrafo o en una misma página.

El Segmentador recibe el contenido textual, que es resultado del módulo Separación de contenido y metadatos de ICIS. El contenido textual es segmentado, ya sea en páginas, párrafos, oraciones o tablas, las tablas a su vez son segmentadas en párrafos u oraciones.

Con el contenido textual se construye un objeto de la clase Segmentos, cuyas propiedades son la ruta del documento, el texto y la lista de páginas. La clase Segmentos incluye métodos

para obtener cada una de las segmentaciones (páginas, párrafos, oraciones, párrafos de tablas y oraciones de tablas).

Cada segmentación es una lista de objetos que la representan. Por ejemplo, la segmentación por páginas es una lista de objetos de la clase Página, cuyas propiedades son el índice de ubicación de la página, el texto y la estampa de tiempo en que se hizo la separación de contenido y metadatos. De la misma forma, la segmentación por párrafos es una lista de objetos de la clase Párrafo, con su índice de ubicación, su texto y su estampa de tiempo, y así sucesivamente.

Los índices de ubicación de los segmentos se forman con los marcadores de nivel 'p:', 'pr:', 'o:', 't:', 'c:' y los índices numéricos de cada nivel al que pertenecen. En la *ilustración 14* se muestran ejemplos de los índices de ubicación, de acuerdo al tipo de segmentación del texto.

Tipo de Segmentación	Índice de ubicación ejemplo
Página	p:1
Párrafo	p:1:pr:1
Oración	p:1:pr:1:o1
Párrafo de tabla	p:1:t:1:c:1:pr:1
Oración de tabla	p:1:t:1:c:1:pr:1:o:1

Ilustración 14. Identificadores ejemplo para los tipos de segmentación.

Cuando los documentos contienen texto dentro de tablas, la segmentación se hace por párrafos de tabla o por oraciones de tabla. Los índices de ubicación de los segmentos por párrafos de tabla incluyen la página, la tabla, la celda y el párrafo. Los índices de ubicación de los segmentos por oración de tabla incluyen, además, la oración.

Posteriormente, tanto la Identificación de datos personales, como la construcción de vectores y las Clasificaciones de datos personales e información sensible, se efectuarán de acuerdo al tipo de segmentación. Por ejemplo, si el texto fue segmentado por oraciones, cada oración tendrá su lista de datos personales identificados, su vector y sus etiquetas de clasificación de información sensible.

4.4.4. Filtrado

Filtrado es un módulo dentro de ICIS cuyo objetivo es filtrar o limpiar los textos de las páginas, párrafos y oraciones.

El Filtrado recibe una lista de segmentos de texto, ya sea una lista de objetos de la clase Página o una lista de objetos de la clase Párrafo, por ejemplo, que son producto de la Segmentación de texto de ICIS.

Los segmentos de texto son filtrados, eliminando los caracteres especiales, los espacios en blanco consecutivos, los tabuladores consecutivos y los cambios de línea consecutivos.

4.4.5. Identificación de datos personales

Es un módulo de ICIS cuyo objetivo es identificar, en los segmentos de texto, los diferentes tipos de datos personales incluidos en la taxonomía.

Los segmentos de texto son procesados para identificar en ellos 55 tipos de datos personales. Se tienen 55 identificadores especializados para cada tipo, que utilizan técnicas de PLN como expresiones regulares, análisis gramatical y diccionarios, como se describe en la sección 4.2.2.

La identificación de datos se ejecuta en paralelo. Se tienen 10 procesos de identificación, uno por categoría, los procesos de identificación por categoría agrupan a los 55 identificadores de acuerdo con la taxonomía. Los procesos de identificación por categoría se ejecutan en paralelo, mediante hilos, en cada uno de los segmentos de texto, como se muestra en la *Ilustración 15*.

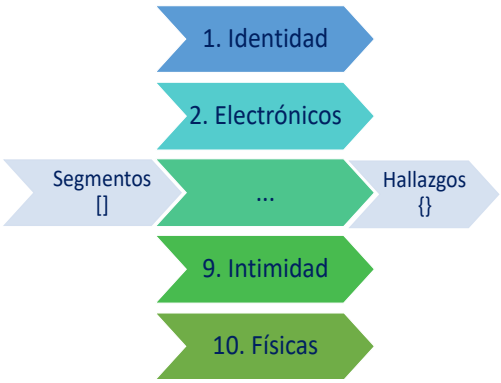


Ilustración 15. Ejecución en paralelo de los procesos de Identificación por categorías de datos personales.

Como resultado se tiene una estructura de datos tipo diccionario con los hallazgos en los segmentos de texto, es decir, las listas de los datos personales identificados en cada segmento de texto. El diccionario de hallazgos contiene para cada segmento: la ubicación, el texto, los tipos y los datos identificados de cada tipo, como lo muesra la *Ilustración 16*.

```
{
':p:1:pr:1:o:3:': {
'TEXTO': 'texto1', 'NOMBRE': ['nombre1']},

':p:1:pr:1:o:4:': {
'TEXTO': 'texto2', 'NOMBRE': ['nombre2'], 'PREFPOLITICA' : ['prefpolitica1']}
}
```

Ilustración 16. Formato del diccionario con los hallazgos de datos personales identificados en los segmentos de texto.

4.4.6. Construcción del vector

Es un módulo dentro de ICIS cuyo objetivo es construir los vectores para la clasificación de información sensible de los segmentos de texto.

Los vectores de clasificación tienen 55 características correspondientes a los tipos de datos personales. Las características son binarias, cada una indica si el segmento de texto contiene un tipo de datos personales en particular, como se detalla en la sección 4.3.1.

Inicialmente se asignan ceros a todas las características del vector de un segmento de texto. Dentro de los procesos de identificación por categoría se modifican los valores de las características, asignándoles un uno, en caso de identificar el tipo correspondiente a su posición en el vector.

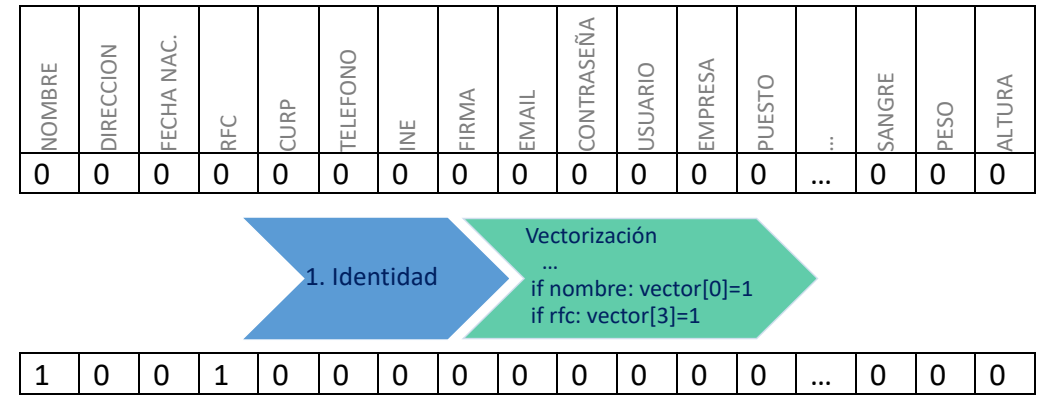


Ilustración 17. Construcción del vector de un segmento, de acuerdo a los datos identificados en su categoría.

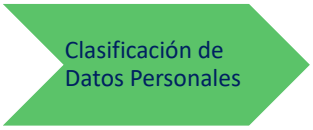
Por ejemplo, si en el proceso de la categoría Identidad, en un segmento de texto se identifican un nombre y un RFC, los valores 0 de la primera y de la cuarta característica se reemplazan por 1, como se muestra en la *Ilustración 17*. Así, en cada proceso de identificación por categoría se construye el vector, modificando solo las características de los tipos de datos correspondientes.

4.4.7. Clasificación de datos personales

Es el módulo dentro de ICIS cuyo objetivo es clasificar los segmentos de texto como Datos Personales, Datos Identificadores, Datos Sensibles Unitarios y Datos Sensibles No Unitarios.

El vector de datos personales de 55 características es la entrada a cada uno de los cuatro clasificadores, C_DSU, C_DPI, C_DSNU y C_DP. Como resultado, se tienen las cuatro etiquetas asignadas, como se muestra en la *Ilustración 18* y se indica en la sección 4.3.2.

NOMBRE	DIRECCION	FECHA NAC.	RFC	CURP	TELEFONO	INE	FIRMA	EMAIL	CONTRASEÑA	USUARIO	EMPRESA	PUESTO	...	SANGRE	PESO	ALTURA
1	0	0	0	1	0	0	0	0	0	0	0	0	...	1	0	0



C_DSU	C_DPI	C_DSNU	C_DP
DSU	DPI	NO_ DSNU	DP

Ilustración 18. Ejemplo de Clasificación de datos personales.

Para las Clasificaciones de Datos Personales se utlizaron los algoritmos de aprendizaje automático SVM, nB, LR y RF, con un conjunto de datos de 17,000 elementos sintéticos, generados de acuerdo a las reglas de los clasificadores descritas en la sección 4.3.2.

En el diccionario resultante del módulo de Identificación, que contiene los hallazgos de los segmentos de texto, se agregan las cuatro etiquetas de los clasificadores de datos personales.

4.4.8. Clasificación de información sensible

Es el módulo de ICIS cuyo objetivo es clasificar los segmentos de texto como Información Sensible. Las etiquetas asignadas a los vectores de datos personales permiten clasificarlos como Información Sensible.

Con las cuatro etiquetas de los clasificadores de datos personales, que se asignaron al vector de un segmento de texto, se genera un nuevo vector de cuatro características, este vector se clasifica como Información Sensible, como se muestra en la *Ilustración 19* y se indica en la sección 4.3.3.

C_DSU	C_DPI	C_DSNU	C_DP
DSU	DPI	NO_ DSNU	DP

C_IS
IS

Ilustración 19. Ejemplo de Clasificación de Información Sensible.

Para la clasificación de Información Sensible se utilizaron los algoritmos de aprendizaje automático RF, LR, nB y SVM, con un conjunto de datos de 17,000 elementos, producto de la Clasificación de datos personales, etiquetados de acuerdo a la regla de clasificación de información sensible descrita en la sección 4.3.3.

En el diccionario resultante del módulo de Identificación, que contiene los hallazgos de los segmentos de texto y las etiquetas de los clasificadores de datos personales, se incluye también la etiqueta de la clasificación de información sensible.

4.4.9. Generación de la estructura de salida tipo JSON

Con el diccionario de los hallazgos se genera una estructura json, que incluye la ruta del archivo analizado, la ubicación de cada segmento de texto con hallazgos y dentro de ésta, el texto, los datos identificados y las etiquetas de clasificación de datos personales e información sensible del segmento. La *Ilustración 20* muestra un ejemplo de la salida json.

{'Archivo.txt': {
':p:1:pr:8:or:1': {
'TEXTO': 'Juan Perez Lopez tiene una alergia al fenileno, tiene un número de cédula 1234567',
'NOMBRES':'Juan Perez Lopez'], 'CEDULA': ['1234567'], 'ALERGIA': ['fenileno']],
'CLASIFICADORES': ['DSU', 'DPI', 'DSNU', 'DP', 'IS']}
}}

Ilustración 20. Ejemplo de la estructura json que devuelve el modelo ICIS.

5. Análisis de pruebas

Siguiendo la metodología, después de ajustar el diseño del modelo, se diseñó la experimentación para el modelo de clasificación, se diseñaron y ejecutaron las pruebas y se analizaron sus resultados.

En este capítulo se presenta el diseño de las pruebas al modelo ICIS, los resultados de la ejecución de las pruebas realizadas y el análisis de sus resultados.

Se hicieron pruebas para los módulos de preprocesamiento, para la clasificación de datos personales y para la clasificación de información sensible del modelo ICIS.

5.1. Pruebas del Preprocesamiento

Los módulos de preprocesamiento son la Separación, la Segmentación y el Filtrado. En esta sección se muestran los resultados del diseño, la ejecución y el análisis de las pruebas del preprocesamiento.

5.1.1. Diseño de las pruebas del preprocesamiento

Se utilizó un conjunto de 50 archivos de diferentes formatos, como entrada al preprocesamiento, con el objetivo de evaluar su correcto funcionamiento en la salida de cada módulo, pues los módulos de preprocesamiento son procesos en cascada.

La entrada al módulo de Separación es un archivo, su salida es la tupla formada por el contenido textual y los metadatos del archivo.

El contenido textual es ahora la entrada al módulo de Segmentación, parametrizando el tipo de segmentación (páginas, párrafos, oraciones, párrafos de tabla y oraciones de tabla) y la salida de este módulo es la lista de los segmentos de texto, es decir, una lista de páginas, párrafos u oraciones.

La lista de segmentos es la entrada al módulo de Filtrado y su salida es también una lista de segmentos, libre de caracteres especiales y de tabuladores, líneas y espacios en blanco consecutivos.

Con cada archivo se requiere probar que separe, segmente y filtre correctamente. Por esta razón se debe analizar la salida de cada uno de los tres módulos del preprocesamiento.

5.1.2. Ejecución de las pruebas del preprocesamiento

En el Anexo 3 se muestran los resultados de la ejecución de las pruebas del preprocesamiento.

5.1.3. Análisis de las pruebas del preprocesamiento

Como podemos observar, el preprocesamiento funciona adecuadamente en sus 3 etapas. Posiblemente la más crítica es la primera, pues si el archivo está dañado no se obtienen los metadatos ni el conenido textual. Además, para leer el contenido textual es necesario que el archivo contenga texto que pueda ser seleccionado, pues en el caso de los archivos pdf con texto obtenido con reconocimiento óptico de caracteres (OCR), el contenido es ilegible o incompleto.

Una vez que se ha separado el contenido textual de los metadatos, la segmentación y el filtrado se realizan adecuadamente, preparando los textos para la identificación.

5.2. Pruebas de la Clasificación de datos personales

La clasificación de datos personales se lleva a cabo mediante aprendizaje automático. Consiste en cuatro clasicaciones: C_DSU, C_DPI, C_DSNU, C_DP, descritas en la sección 6.2.2. En cada una de ellas se aplican los algoritmos SVM, nB, *Logistic Regresion* y *Decision tree.*

Los vectores de entrada tienen 55 características binarias, es decir, 2^{55} combinaciones posibles. El conjunto de datos es sintético y está formado por 17,000 elementos o vectores generados aleatoriamente y etiquetados conforme a las reglas de clasificación definidas en la sección 6.2.2.

5.2.1. Diseño de las pruebas de clasificación de datos personales

El conjunto de datos se particiona en 70% para el entrenamiento y el 30% para las pruebas.

La distribución del C_DSU es la siguiente: el 24% fueron etiquetados como NO_DSU, el 64% como DSU y el 12% como NI.

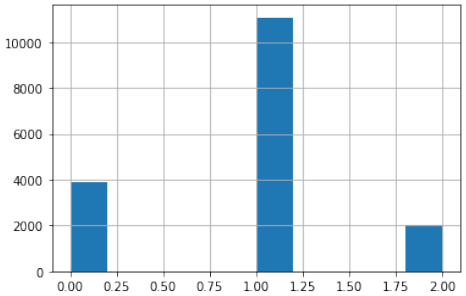


Ilustración 21. Distribución de la clasificación C_DSU.

La distribución del C_DPI es de 21% como NO_DPI, 67% como DPI y 12% como NI.

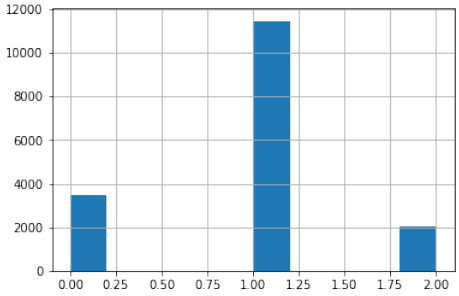


Ilustración 22. Distribución de la clasificación C_DPI.

La distribución del C_DSNU es del 6% como NO_DSNU, 82% como DSNU y 12% como NI.

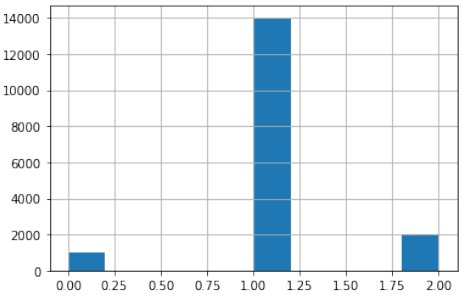


Ilustración 23. Distribución de la clasificación C_DSNU.

La distribución del C_DP es de 11% como NO_DP, 89% como DP.

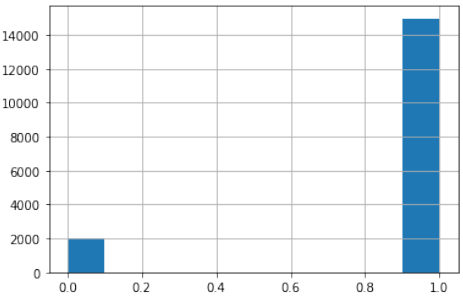


Ilustración 24. Distribución de la clasificación C_DP.

5.2.2. Ejecución de las pruebas de clasificación de datos personales

En cada una de las clasificaciones se aplican los algoritmos *SVM, naive Bayes, Logistic Regresion* y *Decision tree*.

5.2.2.1. Ejecución de las pruebas del clasificador C_DSU (y personales multiclase)

En la *ilustración 25* se muestran las matrices de confusión de los algoritmos aplicados en la clasificación C_DSU. El algoritmo con más positivos verdaderos y con menos falsos negativos en todas las clases es naive Bayes.

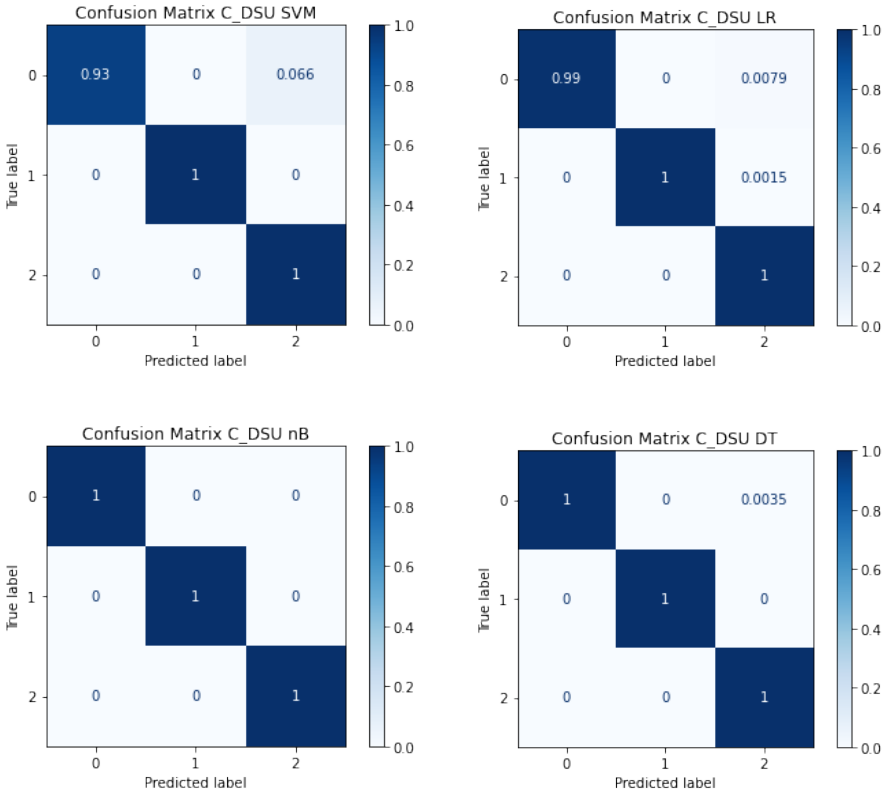


Ilustración 25. Matrices de confusión de los algoritmos aplicados en la clasificación C_DSU.

En la *ilustración 26* se muestran las métricas de precisión, exhaustividad, exactitud y f1. No es relevante la medida de exactitud, pues las clases están desbalanceadas. Nos interesa la exhaustividad, pues al clasificar datos como sensibles unitarios, no se aceptan falsos negativos. El índice más alto de exhaustividad también se obtiene con el algoritmo naive Bayes.

C_DSU	SVM	LR	nB	DT
Precisión	0.9795	0.9969	1.0	0.998839
Exhaustividad	0.9782	0.9968	1.0	0.998835
Exactitud	0.9782	0.9968	1.0	0.998835
Valor-F1	0.9788	0.9969	1.0	0.998837

Ilustración 26. Métricas de los algoritmos en la clasificación C_DSU.

Los clasificadores, C_DPI y C_DSNU tienen métricas similares, pues sus distribuciones son similares a las del calsificador C_DSU.

5.2.2.2. Ejecución de las pruebas del clasificador binario C_DP

El clasificador C_DP tiene una distribución distinta a C_DSU, C_DPI y C_DSNU, pues es un clasificador binario. En la *ilustración 27* se muestran las matrices de confusión de los cuatro algoritmos. nB es el que tiene más verdaderos positivos y menos falsos negativos.

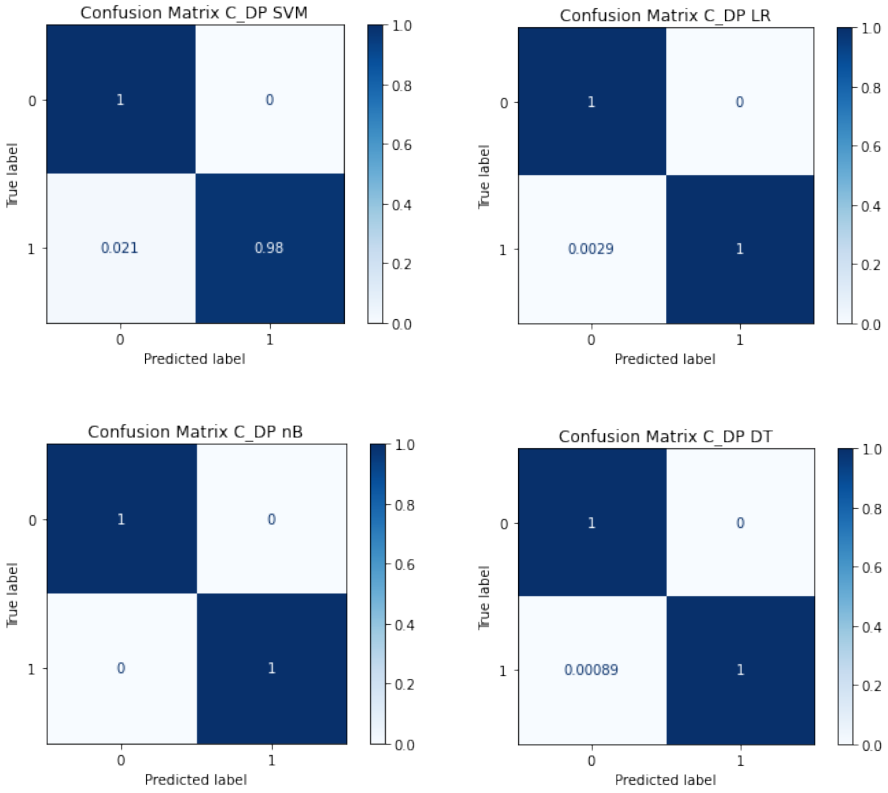


Ilustración 27. Matrices de confusión de los algoritmos de clasificación C_DP.

En la *ilustración 28* se muestran las métricas de precisión, exhaustividad, exactitud y f1. El conjunto de datos no está balanceado, por lo tanto también descartamos exactitud. Y elegimos exhaustividad, pues un falso negativo sería inaceptable. La exhaustividad más alta se obtiene con el algoritmo nB.

C_DP	SVM	LR	nB	DT
Precisión	0.989858	0.998557	1.0	0.9995552
Exhaustividad	0.989648	0.998553	1.0	0.9995548
Exactitud	0.989648	0.998553	1.0	0.9995548
Valor-F1	0.989753	0.998555	1.0	0.9995549

Ilustración 28. Métricas de los algoritmos del clasificador C_DP.

5.2.3. Análisis de las pruebas de clasificación de datos personales

Las métricas de desempeño de los cuatro algoritmos en los clasificadores de datos personales son muy altas, tanto para los multiclase como para el binario.

Los conjuntos de datos están desbalanceados, por lo que la exactitud podría parecer mejor de lo que realmente es. Por esta razón la precisión, la exhaustividad y f1 son más representativos. Sin embargo, la métrica elegida es la exhaustividad (*recall*), pues en los

clasificadores de datos personales, C_DSU, C_DPI, C_DSNU y C_DP, no se aceptan los falsos negativos, es decir, no se aceptan datos que sean personales pero que la predicción sea falsa, que no sean detectados.

5.3. Pruebas de la Clasificación de información sensible

La clasificación de información sensible se realiza mediante aprendizaje automático. Consiste en la clasifiación C_IS, descrita en la sección 5.2.3, aplicando los algoritmos *Decision Tree, Logistic Regresion, SVM* y *naive Bayes.*

5.3.1. Diseño de las pruebas de clasificación de información sensible

Los vectores de entrada tienen 4 características, que son las etiquetas de los clasificadores de datos personales. Debido al pequeño número de características, se redujo el conjunto de datos al 10%.

El conjunto de datos se integra con las clasificaciones de datos personales de 100 vectores, los cuales son etiquetados siguiendo las reglas de clasificación de información sensible, definidas en la sección 5.2.3.

Para probar la clasificación de información sensible el conjunto de datos se divide en 70% para el entrenamiento y el 30% para las pruebas.

La distribución de la clasificación C_IS es de 36% como NI, 33% como IP y 31% como IS.

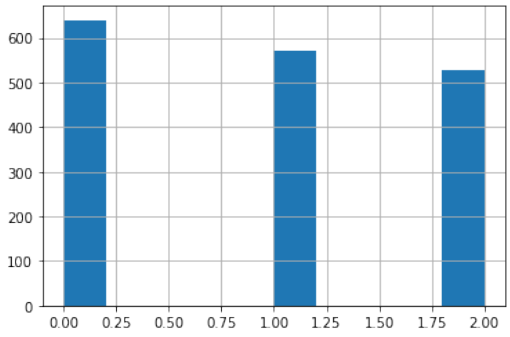


Ilustración 29. Distribución de la clasificación C_IS.

5.3.2. Ejecución de las pruebas de clasifiación de información sensible

Debido a que los tamaños de los vectores y de la muestra son muy pequeños, el algoritmo con mejores resultados es DT. En la *ilustración 30* se observa que la matriz de confusión con más verdaderos positivos y menos falsos negativos, en todas sus clases, es la del algoritmo DT.

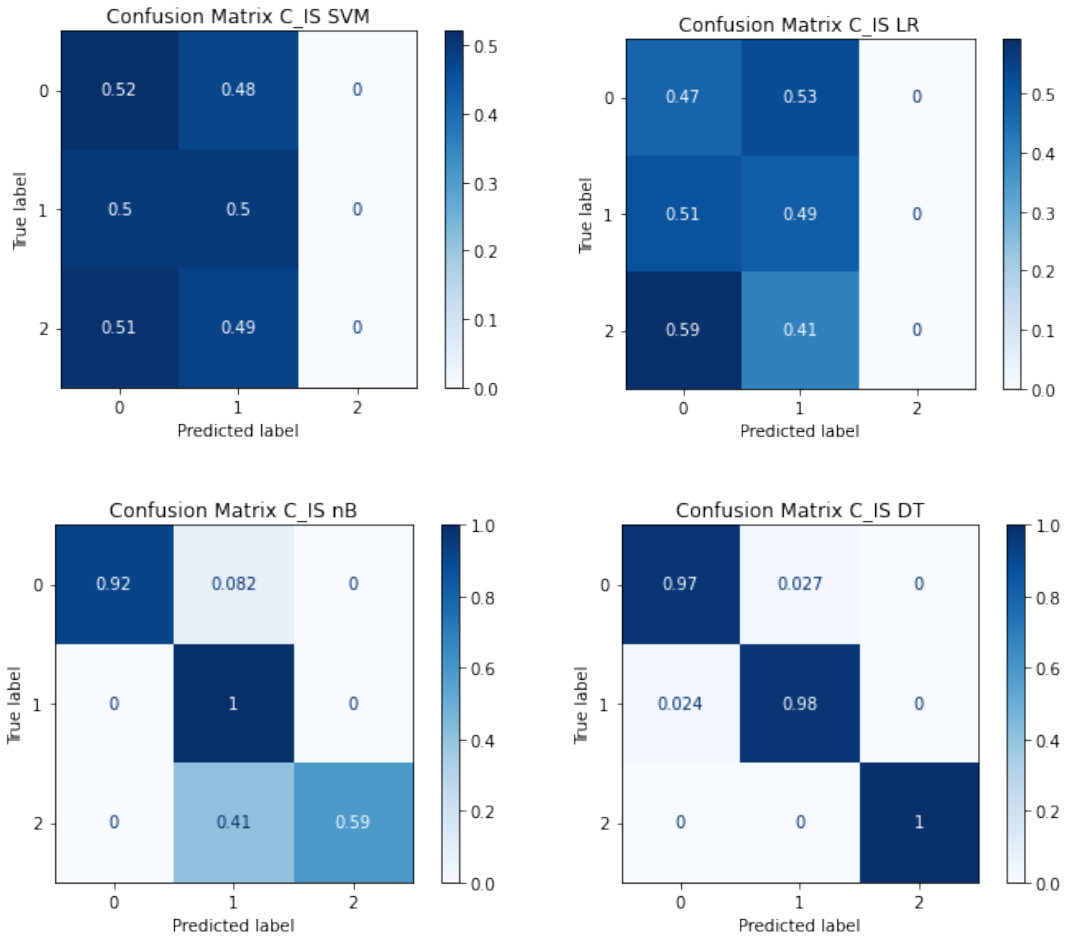


Ilustración 30. Matrices de confusión de los algoritmos del clasificador C_IS.

En el clasificador de información sensible C_IS, el conjunto de datos sí está balanceado, por lo que la exactitud se considera confiable. Sin embargo, nuevamente la métrica que nos interesa el exhaustividad, porque no es posible aceptar falsos negativos. En la *ilustración 31* se muestran las métricas de los algoritmos del clasificador C_IS. El algoritmo con las mejores métricas, incluida la exhaustividad, es DT.

C_IS	SVM	LR	nB	DT
Precisión	nan	Nan	0.891	0.983
Exhaustividad	0.227	0.215	0.741	0.983
Exactitud	0.255	0.240	0.737	0.983
Valor-F1	nan	Nan	0.809	0.983

Ilustración 31. Métricas de los algoritmos del clasificador C_IS.

5.3.3. Análisis de las pruebas clasifiación de información sensible

Debido a que el conjunto de datos y el vector son pequeños, los algoritmos nB y DT tuvieron el mejor desempeño.

La muestra está balanceada, por lo que la exactitud es confiable. Sin embargo, debido a que no se aceptan falsos negativos, la métrica seleccionada es exhaustividad. El algoritmo seleccionado es DT, que tiene las mejores métricas, incluyendo la exhaustividad.

6. Conclusiones

Con este trabajo de investigación se comprobó que los datos personales pueden ser identificados considerando el contexto, mediante técnicas de PLN y que la información puede ser clasificada como sensible, utilizando técnicas de ML. Con ICIS se cumple el objetivo de diseñar un modelo para la identificación y clasificación de información sensible en textos, que puede ayudar a prevenir su exposición en espacios públicos de sujetos obligados.

El modelo, a diferencia del estado del arte, no solo indica si un texto corto contiene información sensible o no, y no solo detecta diferentes tipos de datos personales. Indica los tipos de datos personales que fueron identificados en cada segmento de texto, sus valores, su ubicación en el documento y si el segmento de texto es clasificado como información sensible, derivado de la regla definida para ello y de las cuatro clasificaciones de datos personales propuestas.

Las principales aportaciones de este trabajo son la taxonomía de los tipos de datos personales, la identificación de datos personales basada en el contexto, la clasificacion de información sensible basada en el contexto y la definición de la información personal sensible, en términos computacionales.

Con la taxonomía de datos personales propuesta, se identificaron los diferentes tipos de datos en documentos de distintos formatos, mediante técnicas de PLN, como expresiones regulares, diccionarios y análisis gramatical. Es posible agregar tipos de datos a la taxonomía o adaptarla, agregando el proceso de identificación, aumentando el tamaño del vector y adpatando en conjunto de datos para las cuatro clasificaciones de datos personales, pero sin alterar la clasificación de información sensible.

Los datos personales fueron detectados, no solo por su formato, sino por su contexto. La segmentación facilitó la tarea de contextualización, pues el modelo considera que los datos son personales solamente si están relacionados con una persona física en una misma oración, en un mismo párrafo o página. El análisis gramatical también facilitó que se considere el contexto, al identificar datos personales tomando en cuenta la estructura gramatical, no solo el formato o las palabras características de ese tipo de datos. El tiempo de ejecución de los procesos de identificación se optimizó, al agruparlos por categoría y ejecutarlos en paralelo, en lugar de ejecutar los 55 procesos de manera secuencial en cada segmento de texto.

Para clasificar la información como sensible, se definieron cuatro diferentes clasificaciones de datos, los sensibles unitarios (DSU), los sensibles no unitarios (DSNU), los personales identificadores (DPI) y los personales (DP). Para este mismo fin, se propuso la regla de clasificación de información sensible. Esta regla de clasificación define, en términos computacionales, a la información personal sensible.

Los cuatro algoritmos de aprendizaje automático evaluados tuvieron un excelente desempeño para las clasificaciones de datos personales, con vectores de 55 características binarias y un conjunto de datos de 17,000 elementos. En cambio, el algoritmo de aprendizaje automático DT tuvo el mejor desempeño para la clasificación de información sensible, con un conjunto de datos y un vector pequeño, de 4 características. Los resultados de la clasificación superan a lo reportado en el estado del arte, en el capítulo 2.

6.1. Trabajo a futuro

En cuanto a la implementación del modelo en la plataforma PICIS, representa un control de seguridad orientado a la confidencialidad de los datos personales de los sujetos obligados, que ayuda a los encargados y responsables, en las instituciones gubernamentales de nuestro país, a evitar la exposición de información sensible en espacios públicos. Se tienen acuerdos con dos instituciones federales, pero se espera que el uso se extienda a nivel nacional, inicialmente.

Este modelo ICIS además de ayudar a los sujetos obligados a prevenir la exposición de información sensible, puede ser implementado en una herramienta, como un control de seguridad para prevenir la exposición de información sensible también de los particulares, es decir, de instituciones privadas, pues el modelo es independiente de que el tipo de redacción sea legal, comercial o de servicios; se basa en los datos personales que se identifican en un texto, para clasificarlos como información sensible.

El modelo ICIS también puede aplicarse como un control de seguridad orientado a coadyuvar en la prevención de la exposición o fuga de los datos personales de los titulares, para verificar que los documentos en sus dispositivos, estén libres de información sensible.

Además de analizar el contenido textual de los documentos, es posible analizar los metadatos de los archivos de cualquier tipo, no solo de texto, sino imágenes, audio, video. En los metadatos podría haber información personal, como el autor, el lugar, el equipo con el que fue creado el documento, por ejemplo. También podrían vincularse los metadatos con el contenido textual del documento.

La identificación de algunos tipos de datos se hizo de acuerdo a la normatividad de México, como las placas de autos, el RFC, CURP o INE, sin embargo, es posible agregar los formatos de los números de identificación utilizados en otras regiones o paises.

La clasificación basada en la segmentación de los textos por páginas, párrafos y oraciones, podría mejorarse con el uso de una ventana deslizante de tamaño variable, que considere un segmento y, además, n segmentos anteriores y n segmentos posteriores, para generar el vector a clasificar.

Las pruebas fueron realizadas al módulo de clasificación, utilizando un conjunto de datos sinético, sin embargo, hace falta probar con un conjunto de datos reales, derivado del etiquetado de documentos descargados de los sitios públicos de los sujetos obligados.

Se podría aplicar también reconocimiento óptico de caracteres a los documentos con texto digitalizado como imagen, pues es común que se tengan archivos pdf de este tipo. En otro tipo de archivos como videos o audio, el sonido se podría convertir en texto para analizar si se habla en ellos de datos personales o de información sensible.

Bibliografía

[1] a. l. A. d. C. y. a. l. F. T. d. D. Decreto promulgatorio del Protocolo Adicional al Convenio para la Protección de las Personas con respecto al Tratamiento Automatizado de Datos de Carácter Personal, «Diario Oficial de la Federación,» 2018. [En línea]. Available: https://www.dof.gob.mx/nota_detalle.php?codigo=5539474&fecha=28/09/2018. [Último acceso: 30 Junio 2021].

[2] Grupo de trabajo del artículo 29. Dictamen 4/2007 sobre el concepto de datos personales, «¿Qué son los datos personales?,» 2007. [En línea]. Available: https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2007/wp136_es.pdf. [Último acceso: Junio 2021].

[3] Reglamento (UE) 2019/679 del Parlamento Europeo y del Consejo del 27 de abril de 2016, relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos, «EUR-Lex,» 2016. [En línea]. Available: https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX%3A32016R0679#d1e1547-1-1. [Último acceso: Julio 2021].

[4] ¿Qué datos personales se consideran sensibles?, «Comisión Europea,» 2021. [En línea]. Available: https://ec.europa.eu/info/law/law-topic/data-protection/reform/rules-business-and-organisations/legal-grounds-processing-data/sensitive-data/what-personal-data-considered-sensitive_es. [Último acceso: 1 Julio 2021].

[5] Clasificación de datos, 2019. [En línea]. Available: https://www.oas.org/es/sms/cicte/docs/ESP-Clasificacion-de-Datos.pdf. [Último acceso: Junio 2021].

[6] R. Mitkov, The Oxford Handbook of Computational Linguistics, Oxford: Oxord University Press, 2004.

[7] P. Harrington, Machine learning in action, Shelter Island, NY: Manning, 2012.

[8] Hassan Mathkour, Ameur Touir, Waleed Al-Sanie, «Automatic information classifier using rhetorical structure theory,» de *International Conference on Intelligent Information Processing and Web Mining*, 2005.

[9] Nagpal, Abhinav; Dasgupta, Riddhiman; Ganesan, Balaji, «Fine Grained Classification of Personal Data Entities with Language Models,» de *CODS-COMAD*, Bangladesh, India, 2022.

[10] Park, Ji-Sung; Kim, Gun-woo; Lee-Dong-ho, «Sensitive Data Identification in Structured Data through GenNER Model Based on Text Generation and NER.,» de *Proceedings of the 2020 International Conference on Computing Networks and Internet of Things*, Sanya, China, 2020.

[11] Jiang, Huimin; Chen, Chunling; Wu, ShengChen; Guo, Yongan, «Classification of Medical Sensitive Data based on Text Classification,» *2019 IEEE International Conference of Consumer Electronics-Taiwan (ICCE-TW),* pp. 1-2, May 2019.

[12] G. Xu, C. Qi, H. Yu, S. Xu, C. Zhao y J. Yuan, «Detecting Sensitive Information of Unstructured Text Usign Convolutional Neural Network,» *2019 International Conference of Cyber-Enabled Distributed Computing and Knowledge Discovery,* pp. 474-479, October 2019.

[13] Y. Liang, Z. Wen, Y. Tao, G. Li y B. Guo, «Automatic Security Classification Based on Incremental Learning and Similarity Comparison,» *2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC),* pp. 812-817, May 2019.

[14] G. McDonald, N. García Pedrajas, C. Macdonald y I. Ounis, «A Study of SVM Kernel Functions for Sensitivity Classification Ensembles with POS Sequences,» *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval,* pp. 1097-1100, 2017.

[15] McDonald, Graham; McDonald, Craig; Ounis, Iadh, «Using Part-of-Speech N-Grams for Sensitive-Text Classification,» de *Proceedings of the 2015 International Conference on The Theory of Information Retrieval*, Northampton, Massachusetts, USA, 2015.

[16] inai, «INAI,» 2021. [En línea]. Available: https://home.inai.org.mx. [Último acceso: 2021].

[17] Ley Federal de Protección de Datos Personales en Posesión de Particulares, «inai.org.mx,» 2021. [En línea]. Available: https://home.inai.org.mx/?page_id=1870&mat=p. [Último acceso: 08 09 2021].

[18] J. K. e. al., Deep Learning Illustrated, USA: Addison-Wesley Data & Analytics Series, 2019.

[19] Glosario, «Protección de Datos Personales,» 2021. [En línea]. Available: http://www.oas.org/es/sla/ddi/proteccion_datos_personales_glosario.asp. [Último acceso: Julio 2021].

[20] Matic, Srjdan; Iordanou, Costas; Smaragdakis, Georigios; Laoutaris, Nikolaos, «Identifying Sensitive URLs at Web-Scale,» de *Proceedings of the ACM Internet Measurement Conference*, 2020.

[21] Shi, Pu; Xiong, Li; Fung, Benjamin C.M., «Anonymizing Data with Quasi-Sensitive Attribute Values,» de *Proceedings of the 19th ACM International Conference of Information and Knowledge Management*, Toronto, ON, Canada, 2010.

Clasificación de Información Personal Sensible*

Sara De Jesús Sánchez[1,2], Jorge Enrique Coyac Torres[1],
Eleazar Aguirre Anaya[1], Francisco Hiram Calvo Castro[2]
sdejesuss2100@alumno.ipn.mx, jcoyact1900@alumno.ipn.mx,
eaguirrea@ipn.mx, fcalvo@ipn.mx

[1] Laboratorio de Ciberseguridad
[2] Laboratorio de Ciencias Cognitivas Computacionales
Centro de Investigación en Computación
Instituto Politécnico Nacional

Resumen La exposición de la información personal sensible en medios públicos representa grandes riesgos tanto para los individuos, titulares de sus propios datos personales, como para las empresas que requieren tratar los datos personales de sus clientes, proveedores y empleados, siendo las mismas empresas responsables de mantener la privacidad de dichos datos, más aún, tratándose de datos personales sensibles. Para disminuir los riesgos de exposición de la información personal sensible es preciso clasificarla, así los responsables o titulares podrán tomar las medidas correspondientes para mantener la confidencialidad de los datos. Un caso particular de este problema de clasificación, es la clasificación textual de la información personal sensible. El presente estudio tiene por objetivo diseñar una herramienta de identificación y clasificación textos con información personal sensible, mediante técnicas de aprendizaje automático para prevenir su exposición en medios públicos. En este trabajo se muestra una clasificación utilizando tres algoritmos de aprendizaje automático para identificar textos que contienen datos personales sensibles.

Palabras clave: ciberseguridad, información sensible, aprendizaje automático, clasificación, procesamiento de lenguaje natural.

Classification of Sensitive Personal Information

Abstract. The exposure of sensitive personal information in public media represents great risks both for individuals, holders of their own personal data, and for companies that need to process the personal data of their customers, suppliers and employees, the same companies are responsible for maintaining the privacy of this data, even more, in the case of sensitive personal data. To reduce the risks of exposure of sensitive personal information, it must be classified, so the managers or owners may take the corresponding measures to maintain the confidentiality of the data. A particular case of this classification problem is the textual classification of sensitive personal information. The objective of this study is to design a tool for identifying and classifying texts with sensitive personal information, using machine learning techniques to prevent their exposure in public media. This work shows a classification using three machine learning algorithms to identify texts that contain sensitive personal data.

Keywords: cybersecurity, sensitive information, machine learning, classification, natural language processing.

* Este trabajo fue apoyado por CONACyT, COFAA, IPN, IPN-EDI, IPN-SIP, OEA, Cisco y la Fundación Citi, gracias a los proyectos SIP 20210189, SIP 20211758 y al proyecto PICIS, ganador del Fondo de Innovación en Ciberseguridad.

Ilustración 32. Página 1 del artículo aceptado en el CORE 21, a publicarse en RCS

1 Introducción

La protección de las personas con respecto al tratamiento automatizado de los datos personales tiene origen en el Convenio 108 del Consejo de Europa en 1981. Además de los países miembros de la Unión Europea, en América, países como Argentina, Uruguay y México han adoptado este convenio como una herramienta global para el intercambio efectivo y seguro de información. Uno de los puntos principales de este convenio es garantizar la confidencialidad de los datos personales sensibles por parte de las organizaciones responsables de ellos [3].

Los datos personales, en general, son los concernientes a una persona física identificada o identificable cuya manifestación puede ser textual, gráfica, acústica o fotográfica [1]. Pueden ser de diferentes tipos: identificativos, laborales, académicos, de salud, de patrimonio.

Un subconjunto de los datos personales son los llamados datos personales sensibles, definidos por la Comisión Europea como los datos que, por su naturaleza, puedan atentar contra las libertades fundamentales o la intimidad, y están sujetos a condiciones de tratamiento específicas [2].

Los datos personales se consideran sensibles en caso de que revelen:

–Origen racial o étnico
–Opiniones políticas
–Creencias religiosas, filosóficas y morales
–Afiliación sindical
–Datos genéticos
–Datos biométricos
–Datos relativos a la salud
–Datos relativos a la vida sexual o a la orientación sexual de una persona [4]

Las organizaciones deben garantizar la seguridad de la información personal de la cual son responsables, de acuerdo con el nivel de sensibilidad, valor y criticidad de ésta [12]. Por tal motivo, clasificarla es la base para disminuir los riesgos en su seguridad. Estos riesgos pueden llegar a ser enormes tanto para las personas, como para las organizaciones y los gobiernos.

En México, por ejemplo, la Ley Federal de Protección de Datos Personales en Posesión de Particulares (LFPDPPP) establece que las multas a los responsables de los datos personales van desde los 100 hasta 320,000 días de salario mínimo vigente en la CDMX y que, en el caso de datos personales sensibles, las sanciones podrán incrementarse hasta por dos veces los montos establecidos [5].

Pero las repercusiones no son únicamente económicas, sino también en la credibilidad y confianza en las organizaciones responsables de los datos personales y en las tecnologías que se utilizan para el tratamiento (obtención, almacenamiento, modificación, copia, eliminación, procesamiento, etc.) de los datos.

Ahora bien, como se mencionó anteriormente, la clasificación de la información, de acuerdo al nivel de sensibilidad, es la base para garantizar su seguridad, es decir, para garantizar la condifencialidad, integridad y disponibilidad de esa información. Este tipo de clasificación es desafiante, debido a diversos factores, por ejemplo: no existe consenso sobre los datos que componen la categoría de datos personales sensibles [13]; se especifican de acuerdo con el riesgo que implican, por lo tanto, la información sensible es incierta; es compleja, pues algunas frases pueden ser sensibles en un contexto y no serlo en otro; el conjunto de datos de entrenamiento debe mantenerse actualizado, es cambiante; se requiere un alto grado de precisión al asignar el nivel de seguridad para que las organizaciones responsables puedan tomar las medidas correspondientes. La clasificación de seguridad de la información puede hacerse en dos niveles (sensible o no) o en más (no clasificada, confidencial, secreta y alto secreto, por ejemplo). La clasificación también podría hacerse sobre todo un documento o en las partes que lo componen.

Ilustración 33. Página 2 del artículo aceptado en el CORE 21, a publicarse en RCS

El tipo de información a clasificar en este proyecto es la textual, contenida en documentos o textos que pudieran ser expuestos en algún medio electrónico. La clasificación textual de información personal sensible se aborda como un problema de procesamiento de lenguaje natural o NLP (por sus siglas en inglés), que usa el lenguaje dentro de un documento para clasificarlo en una categoría particular [7]. Para ello se pueden utilizar algoritmos de aprendizaje automático.
La clasificación automática de seguridad es una tecnología que permite predecir el nivel de seguridad de la información de un documento, se abordó como un problema de investigación en 2005 [9] y desde entonces se han propuesto numerosos métodos para resolverlo, tales como el aprendizaje automático y las redes neuronales, principalmente. Empleando distintos tipos de documentos en ámbitos específicos, como el médico, el organizacional y el militar, en lenguas china, coreana e inglesa.
Este proyecto está enfocado en la clasificación de textos en español, de acuerdo a su nivel de sensibilidad en los datos personales, para ello se evalúan distintos algoritmos con el fin de encontrar aquéllos que ofrezcan la mayor precisión en la clasfificación.
El objetivo del proyecto es diseñar una herramienta para la identificación y clasificación textual de información personal sensible, mediante técnicas de inteligencia artificial, para prevenir su exposición en espacios públicos.

2 Estado del arte

En 2005 Hassan Mathkour utilizó Árboles Binarios para la Clasificación de Seguridad basada en Estructuras Retóricas en idioma Árabe [9].
Graham McDonald et al. emplearon, en 2015 y 2017, n-gramas, análisis gramatical (Part of Speech POS) y máquinas de vectores de soporte (SVM) obteniendo una efectividad del 90 % con patrones fijos para la clasificación de sensibilidad de textos en inglés [11].
En 2019, Yan Liang et al. utilizaron aprendizaje incremental y comparación de similitud (ILSC) para la clasificación de textos sensibles en Chino con una efectividad del 86 %, incremental suport vector machine (ISVM) con 87 %, online random forest (ORF) con 84 % y naive Bayes (NB) con 82 % [8].
En ese mismo año, Gousheng Xu et al. emplearon redes neuronales de convolución (CNN) con 95 % de efectividad y las recurrent neural networks (RNN) con 94 % para detectar información sensible en textos no estructurados en chino [15].
Huimin Jiang et al. también en 2019 diseñaron un clasificador de datos médicos sensibles en chino con un 90 % de precisión, no mencionan el algoritmo empleado [6]. En 2020, Srdjan Matic et al. diseñaron un clasificador de URLs sensibles con un 88 % de efectividad, sin mencionar el algoritmo ni el idioma utilizado [10].
Ji-Sung Park et al., en 2020 desarrollaron un sistema de prevención de pérdida de datos (DLP) para clasificar palabras en coreano, en categorías de datos personales sensibles utilizando reconocimiento de entidades nombradas (NER), no mencionan la efectividad obtenida [14].
Como podemos ver, los métodos que actualmente han logrado mayor precisión en la clasificación de infomación sensible, son los basados en aprendizaje automático y los basados en redes neuronales. Los tipos de documentos e idiomas sobre los que se han aplicado son muy variados.
En el presente artículo se hace una clasificación de textos en español, para identificar si contienen información personal sensible, utilizando algoritmos de aprendizaje automático. En este ejercicio la clasificación es binaria, las dos clases de objetos son: sensible y no sensible.

Ilustración 34. Página 3 del artículo aceptado en el CORE 21, a publicarse en RCS

3 Desarrollo de la solución

Para esta solución inicial se tiene un conjunto de datos que consiste en 60 textos, el 60 % de ellos contiene información personal sensible. El conjunto de datos se obtiene de un archivo de texto separado por tabuladores, donde en cada renglón se tiene un texto y su etiqueta asignada con los valores de 1, si contiene información personal sensible, y de 0 en el caso contrario.
El modelo de espacio vectorial que se utiliza está basado en el modelo bolsa de palabras (Bag of Words BOW), donde las palabras del vocabulario son las características de los objetos. Los vectores, en este caso, contienen los valores de frecuencia de término (Tf). Con el fin de reducir el tamaño de los vectores, se hace un preprocesamiento a los objetos con técnicas de PLN, como la normalización y la eliminación de palabras auxiliares.
Posteriormente se toma el 75 % de los objetos para el entrenamiento y el 25 % para las pruebas.
Haciendo uso de la biblioteca sklearn de Python, se aplican tres clasificadores de aprendizaje automático: Regresión Logística, Naive Bayes y Máquina de Vectores de Soporte.

4 Experimentos y resultados

Con una muestra de 20 ejecuciones, los resultados obtenidos en la exactitud (Accuracy) de los algoritmos de Regresión Logística, Máquina de Vectores de Soporte y naive Bayes, son los siguientes:

Cuadro 1. Exactitud obtenida

RL	SVM	nB
0.8401	0.8525	0.8395

Se observan exactitudes muy similares en los tres algoritmos, ligeramente superior con SVM. Los resultados son cercanos a los reportados con algoritmos de aprendizaje automático encontrados en el estado del arte.

5 Conclusiones y trabajo futuro

En el presente artículo se ha mostrado la importancia de encontrar las mejores técnicas de clasificación textual de la información personal sensible, para coadyuvar en su confidencialidad y, por lo tanto, en su seguridad.
Se utilizaron tres algoritmos de aprendizaje automático para la clasificación de textos con información personal sensible, obteniendo resultados similares a los vistos en el estado del arte.
Es necesario considerar que el conjunto de datos empleado es pequeño y es conveniente incrementarlo para los siguientes estudios. Se observa la importancia del PLN para disminuir el tamaño de los vectores. Se debe analizar si es necesario incluir otros tipos de preprocesamiento, tales como el análisis de expresiones regulares (Regular Expressions RE), el reconocimiento de entidades nombradas (Named Entity Recognition NER), la desambiguación del sentido de las palabras (Word Sense Desambiguation WSD), así como formar los vectores utilizando TfIdf como valores de las características de los objetos.

Ilustración 35. Página 4 del artículo aceptado en el CORE 21, a publicarse en RCS

En trabajos posteriores se utilizarán algoritmos de clasificación basados en redes neuronales, que probablemente requieran de algún modelo de espacio vectorial distinto a BOW.
El trabajo por hacer en este proyecto, además de la utilización de distintos algoritmos de clasificación, es el diseño de una herramienta que utilice los algoritmos que funcionan mejor en la clasificación de información personal sensible, para prevenir su exposición en espacios públicos.

Ilustración 36. Página 5 del artículo aceptado en el CORE 21, a publicarse en RCS

Bibliografía

[1] Grupo de trabajo del articulo 29l dictamen 4/2007 sobre el concepto de datos personales. ¿que son los datos personales?, 2007.

[2] Regramento ue 2019/679 del parlamento europeo y del consejo del 27 de abril de 2016, relativo a la proteccion de las personas fisicas en lo que respecta al tratamiento de datos personales y a la libre circulacion de estos datos, 2016.

[3] a. l. A. d. C. y. a. l. F. T. d. D. Decreto promulgatorio del protocolo adicional al convenio para la proteccion de las personas con respecto al tratamiento automatizado de datos de caracter personal, diario oficial de la federacion, 2018.

[4] Comision Europea. *¿Que datos personales se consideran sensibles?* 2021.

[5] INAI. Ley federal de proteccion de datos personales en posesion de particulares. En linea, 2021.

[6] Huimin Jiang, Chunling Chen, ShengChen Wu, and Yongan Guo. Classification of medical sensitive data based on text classification. IEEE, may 2019.

[7] J.K. *Deep Learning Illustrated*. 2019.

[8] Yan Liang, Zepeng Wen, Yizheng Tao, GongLiang Li, and Bing Guo. Automatic security classification based on incremental learning and similarity comparison. IEEE, may 2019.

[9] Hassan Mathkour, Ameur Touir, and Waleed Al-Sanie. Automatic information classifier using rhetorical structure theory. In Mieczyslaw A. Klopotek, Slawomir T. Wierzchon, and Krzysztof Trojanowski, editors, *Intelligent Information Processing and Web Mining, Proceedings of the International IIS: IIPWM'05 Conference held in Gdansk, Poland, June 13-16, 2005*, volume 31 of *Advances in Soft Computing*, pages 229–236. Springer, 2005.

[10] Srdjan Matic, Costas Iordanou, Georgios Smaragdakis, and Nikolaos Laoutaris. Identifying sensitive urls at web-scale. In *Proceedings of the ACM Internet Measurement Conference*, IMC '20, page 619–633, New York, NY, USA, 10 2020. Association for Computing Machinery.

[11] Graham McDonald, Craig Macdonald, and Iadh Ounis. Using part-of-speech n-grams for sensitive-text classification. In *Proceedings of the 2015 International Conference on The Theory of Information Retrieval*, ICTIR '15, page 381–384, New York, NY, USA, 9 2015. Association for Computing Machinery.

[12] OEA. Clasificacion de datos. En linea, 2019.

[13] OEA. *Glosario-Proteccion de Datos Personales*. 2021.

[14] Ji sung Park, Gun woo Kim, and Dong ho Lee. Sensitive data identification in structured data through genner model based on text generation and ner. In *Proceedings of the 2020 International Conference on Computing, Networks and Internet of Things*, CNIOT2020, page 36–40, New York, NY, USA, 4 2020. Association for Computing Machinery.

[15] Guosheng Xu, Chunhao Qi, Hai Yu, Shengwei Xu, Chunlu Zhao, and Jing Yuan. Detecting sensitive information of unstructured text using convolutional neural network. IEEE, oct 2019.

Ilustración 37. Página 6 del artículo aceptado en el CORE 21, a publicarse en RCS

Classification of Sensitive Personal Information

Eleazar Aguirre Anaya[1], Francisco Hiram Calvo Castro[1], Sara De Jesús Sánchez[1]
eaguirrea@ipn.mx, fcalvo@ipn.mx, sdejesuss2100@ipn.mx
[1]IPN, Centro de Investigación en Computación, Ciudad de México, México

Keywords

Information security; Sensitive information; Personal data; Security classification.

Abstract

The exposure of personal information in public media represents great risks for institutions that process personal data, being the same institutions responsible for the security of this data, even more so, in the case of sensitive personal information, which involve greater risks.

This proposal presents a platform for the identification and classification of sensitive information published on public organizations websites, this platform generates alerts due to the publication of sensitive data.

1. Information extraction

The platform will analyze the information obtained from mexican public organizations websites. The information analyzed is the text of the pages published on the websites, the files that contain text and their metadata.

2. Classification

The texts to be analyzed are first preprocessed and then classified.

Preprocessing consists of natural language processing techniques. Classification is based on machine learning algorithms.

After the texts have been classified, the results are saved and, where appropriate, it is alerted if they contain sensitive information, what sensitive information they contain and the location within the texts.

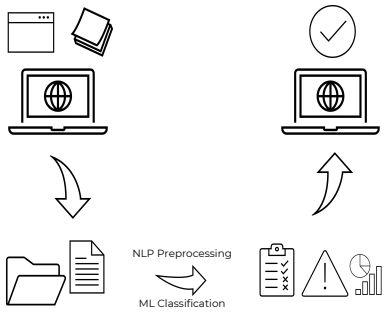


Fig. 1. Process to avoid sensitive personal information exposure.

3. Record

For each site analyzed, a file is generated, where the results are recorded. This facilitates the attention and follow-up of sensitive information exposure alerts.

4. Conclusions

The objective of this proposal is to support the security analysts in avoiding the exposure of sensitive information in public spaces. In order to improve the analysis, different natural language processing and machine learning classification techniques will be applied.

With the goal of evaluating the classification algorithms, statistical tests will be performed.

Ilustración 38. Póster presentado en el décimo tercer Encuentro de la Red de Computación

Anexo 2. Lista ordenada de los tipos de datos personales

1	NOMBRE
2	DIRECCIÓN
3	FECHA NACIMIENTO
4	RFC
5	CURP
6	TELÉFONO
7	INE
8	FIRMA
9	EMAIL
10	CONTRASEÑA
11	USUARIO
12	EMPRESA
13	PUESTO
14	DIRECCIÓN LABORAL
15	EMAIL LABORAL
16	PASAPORTE
17	LICENCIA
18	VISA
19	PLACAS
20	NIV
21	SUELDO
22	IMPUESTOS
23	CRÉDITOS
24	NÚMERO DE TARJETA
25	INVERSIONES
26	AFORE
27	SEGUROS
28	ESTADO DE SALUD
29	HISTORIAL CLÍNICO
30	ENFERMEDAD
31	TRATAMIENTO
32	ESTUDIO CLÍNICO
33	ALERGIA
34	CONDICIÓN PSICOLÓGICA
35	NSS
36	ESCUELA
37	CALIFICACIÓN

38	TÍTULO
39	CERTIFICADO
40	NÚMERO DE CÉDULA
41	RELIGIÓN
42	AFILIACIÓN SINDICAL
43	PREFERENCIA POLÍTICA
44	ORGANIZACIÓN CIVIL
45	PREFERENCIA SEXUAL
46	HÁBITO
47	RELACIÓN PERSONAL
48	IRIS
49	ADN
50	COLOR DE PIEL
51	HUELLA DACTILAR
52	CICATRIZ
53	TIPO DE SANGRE
54	PESO
55	ALTURA

Anexo 3. Pruebas del preprocesamiento

Matriz de Pruebas del Separador

Nombre del archivo	Tipo	Pruebas					Resultados	Comentarios
		Daño del 0%	Daño del 25%	Daño del 50%	Daño del 75%	Daño del 100%		
2717524247346998489.gif	GIF	X					(diccionario, None)	El API con el archivo sin daños regresará los metadatos como "Diccionario" y un objeto "None". Debido a que este formato de archivo no posee contenido de texto.
2717524247346998489_25.gif			X				(None, None)	
2717524247346998489_50.gif				X			(None, None)	
2717524247346998489_75.gif					X		(None, None)	
2717524247346998489_100.gif						X	(None, None)	
4997871131361703284.jpeg	JPEG	X					(diccionario, None)	El API con el archivo sin daños regresará los metadatos como "Diccionario" y un objeto "None". Debido a que este formato de archivo no posee contenido de texto.
4997871131361703284_25.jpeg			X				(None, None)	
4997871131361703284_50.jpeg				X			(None, None)	
4997871131361703284_75.jpeg					X		(None, None)	
4997871131361703284_100.jpeg						X	(None, None)	
18299166394611608593.png	PNG	X					(diccionario, None)	El API con el archivo sin daños regresará los metadatos como "Diccionario" y un objeto "None". Debido a que este formato de archivo no posee contenido de texto.
18299166394611608593_25.png			X				(None, None)	
18299166394611608593_50.png				X			(None, None)	
18299166394611608593_75.png					X		(None, None)	
18299166394611608593_100.png						X	(None, None)	
all.css	CSS	X					(diccionario, string)	El API con el archivo sin daños regresará los metadatos como "Diccionario" y un objeto "String". Con el 25% de daño, sólo se pueden extraer los metadatos. Un daño superior devuelve ambos objetos tipo "None"
all_25.css			X				(diccionario None)	
all_50.css				X			(None, None)	
all_75.css					X		(None, None)	
all_100.css						X	(None, None)	
analytics.js	JS	X					(diccionario, string)	El API con el archivo sin daños regresará los metadatos como "Diccionario" y un objeto "String". Con el 25% de daño, sólo se pueden extraer los metadatos. Un daño superior devuelve ambos objetos tipo "None"
analytics_25.js			X				(diccionario, None)	
analytics_50.js				X			(None, None)	
analytics_75.js					X		(None, None)	
analytics_100.js						X	(None, None)	
npp.8.3.3.Installer.x64.exe	EXE	X					(diccionario, None)	El API con el archivo sin daños regresará los metadatos como "Diccionario" y un objeto "None". Debido a que este formato de archivo no posee contenido de texto.
npp.8.3.3.Installer.x64_25.exe			X				(None, None)	
npp.8.3.3.Installer.x64_50.exe				X			(None, None)	
npp.8.3.3.Installer.x64_75.exe					X		(None, None)	
npp.8.3.3.Installer.x64_100.exe						X	(None, None)	
Oracle_VM_VirtualBox_Extension_Pack-6.1.32.vbox-extpack	GZIP	X					(diccionario, None)	El API con el archivo sin daños regresará los metadatos como "Diccionario" y un objeto
Oracle_VM_VirtualBox_Extension_Pack-6.1.32_25.vbox-extpack			X				(None, None)	
Oracle_VM_VirtualBox_Extension_Pack-6.1.32_50.vbox-extpack				X			(None, None)	