ISSN 1870-4069

Clasificación de información personal sensible

Sara De Jesús Sánchez, Jorge Enrique Coyac Torres,
Eleazar Aguirre Anaya, Hiram Calvo

Instituto Politécnico Nacional,
Centro de Investigación en Computación,
México

sdejesuss2100@alumno.ipn.mx, jcoyact1900@alumno.ipn.mx,
eaguirrea@ipn.mx, fcalvo@ipn.mx

Resumen. La exposición de la información personal sensible en medios públicos representa grandes riesgos tanto para los individuos, quienes son los 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 podrán tomar las medidas correspondientes para mantener la confidencialidad de los datos. Este trabajo muestra la factibilidad de utilizar algoritmos de aprendizaje automático en la identificación y clasificación de textos que contienen datos personales sensibles. Los resultados de los experimentos fueron satisfactorios, empleando los algoritmos naive Bayes, regresión logística y máquina de vectores de soporte y alngunas técnicas de procesamiento de lenguaje natural, como la normalización y eliminación de palabras auxiliares.

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 big risks both for individuals, who are holders of their own personal data, and for companies that must process the personal data of their customers, suppliers and employees. These companies are responsible for maintaining the privacy of such data, even more so, in the case of sensitive personal data. In order to reduce the risks of exposure of sensitive personal information, it must be classified, so those responsible take the corresponding measures to maintain the data confidentiality. This work shows the feasibility of using machine learning algorithms to identify and classify texts that contain sensitive personal data. The results of the experiments were satisfactory, using

naive Bayes, logistic regression and support vector machine algorithms, in addition to some natural language processing techniques, such as lemmatization and stop words elimination.

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

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 [2].

Los datos personales, en general, son los concernientes a una persona física identificada o identificable cuya manifestación puede ser textual, gráfica, acústica o fotográfica [13]. 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 [14].
Los datos personales se consideran sensibles en caso de que revelen: origen racial o étnico, opiniones políticas, creencias religiosas, filosóficas y morales, afiliación sindical, catos genéticos, datos biométricos, datos relativos a la salud, datos relativos a la vida sexual o a la orientación sexual de una persona [3].

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 [10]. 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 [4].

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 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 [11]; 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.

El tipo de información a clasificar en este proyecto es 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 [1].

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 [7] y desde entonces se han propuesto numerosos métodos para resolverlo, tales como el aprendizaje automático y las redes neuronales, principalmente. Se han empleando distintos tipos de documentos en ámbitos específicos, como el médico, el organizacional y el militar, en lenguas china, coreana e inglesa.

El objetivo de este proyecto es identificar y clasificar textos con información personal sensible, mediante técnicas de aprendizaje automático, para prevenir su exposición en espacios públicos. Este proyecto está enfocado en la clasificación de textos en español para saber si contienen datos personales sensibles o no. Para ello se aplicaron tres algoritmos de aprendizaje automático junto con técnicas de NLP, con el fin de encontrar aquéllos que ofrezcan la mayor precisión en la clasfificación.

2. Estado del arte

En 2005 Hassan Mathkour utilizó árboles binarios para la clasificación de seguridad basada en estructuras retóricas en idioma Árabe [7], como se mencionó anteriormente, esta publicación planteó el problema de la clasificación automática de seguridad.

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 [9]. El uso de n-gramas y análisis gramatical para el preprocesamiento es una referencia para identificar secuencias de textos sensibles.

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 % [6]. Se compararon tres algoritmos de aprendizaje automático con buenos resultados, siendo ISVM el más efectivo.

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]. Se compararon las dos principales redes neuronales con una efectividad alta.

Huimin Jiang et al. también en 2019 diseñaron un clasificador de datos médicos sensibles en chino, obteniendo un 90 % de precisión con el algoritmo SVM, 85 % con naive Bayes y 80 % con KNN [5]. Nuevamente vemos que el algoritmo SVM alcanzó mayor precisión.

En 2020, Srdjan Matic et al. diseñaron un clasificador de URLs sensibles con un 88 % de efectividad [8]. No mencionan el algoritmo utilizado, sin embargo la efectividad resultó similar a los otros trabajos revisados.

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) [12]. Se puede considerar el uso de NER para identificar entidades como nombres, lugares, organizaciones que podrían clasificarse como información sensible.

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 y los ámbitos sobre los que se han aplicado son muy variados, así como los idiomas. Se han aplicado técnicas de PLN como n-gramas, POS y NER.

No se encontraron referencias sobre sensibilidad de datos personales en idioma español, tampoco sobre el algoritmo de regresión logística, ni la normalización o eliminación de palabras auxiliares.

En el presente artículo se presenta una clasificación de textos en español, para identificar si contienen información personal sensible, utilizando algoritmos de aprendizaje automático y técnicas de PLN. En esta propuesta la clasificación es binaria, las dos clases de objetos son: sensible (1) y no sensible (0).

3. Desarrollo de la solución

Para esta solución inicial se propone emplear dos técnicas de PLN y tres de los algoritmos más representativos del aprendizaje automático sobre un conjunto de datos de prueba con textos cortos en español para clasificarlos según contengan o no información sensible.

Tabla 1. Exactitud obtenida

RL	SVM	nB
0.8401	0.8525	0.8395

El conjunto de datos consiste en 60 textos, el 60 % de ellos contiene información personal sensible, el 40 % no. El conjunto de datos es almacenado en 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 conjunto de datos está conformado por textos cortos en español, obtenidos de publicaciones en redes sociales y etiquetados manualmente.

El modelo de espacio vectorial 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), es decir qué tanto se repite cada palabra en un texto.

Con el fin de reducir el tamaño de los vectores y con ello facilitar la clasificación de los textos, se hace un preprocesamiento a los objetos antes de entrenar y probar los algoritmos. Este preprocesamiento consiste en aplicar técnicas de PLN a los textos del conjunto de datos.

Las técnicas de PLN empleadas son la normalización y la eliminación de palabras auxiliares. Normalizar un texto es sustituir cada palabra flexionada o derivada (es decir, en plural, conjugada, etc.) por su forma normal o lema. Las palabras auxiliares o stop words son aquellas que, aunque son muy comunes, son poco relevantes (como los artículos, conjunciones, preposiciones, etc.), por tal motivo se filtran. El preprocesamiento se lleva a cabo mediante funciones de la biblioteca Stanza de Python.

Posteriormente el conjunto de datos es dividido de la siguiente manera: 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 de los principales algoritmos 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, se muestran en el Cuadro 1.

Se observan exactitudes 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 factibilidad de emplear algoritmos de aprendizaje automático en la 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 debe analizar si es necesario agregar otras técnicas de PLN en el 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), o el uso de n-gramas, así como formar los vectores utilizando TfIdf como valores de las características de los objetos. Un trabajo posterior es emplear 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.

Agradecimientos. 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 Plataforma de Identificación, Clasificación y Monitoreo de Información sensible para entidades de Gobierno Federal, ganador del Fondo de Innovación en Ciberseguridad de Latinoamérica en 2021.

Referencias

[1] Bassens, A., Beyleveld, G., Krohn, J.: Deep Learning Illustrated (2019)
[2] Diario Oficial de la Federacion: Decreto promulgatorio del protocolo adicional al convenio para la proteccion de las personas con respecto al tratamiento automatizado de datos de caracter personal (2018), https://www.dof.gob.mx/nota_detalle.php?codigo=5539474&fecha=28/09/2018
[3] Europea, C.: ¿Que datos personales se consideran sensibles? (2021), 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
[4] INAI: Ley federal de proteccion de datos personales en posesion de particulares. En linea (2021), https://home.inai.org.mx/?page_id=1870I&mat=p

[5] Jiang, H., Chen, C., Wu, S., Guo, Y.: Classification of medical sensitive data based on text classification. IEEE (may 2019) doi: 10.1109/icce-tw46550.2019.8991726
[6] Liang, Y., Wen, Z., Tao, Y., Li, G., Guo, B.: Automatic security classification based on incremental learning and similarity comparison. IEEE (may 2019) doi: 10.1109/itaic.2019.8785798
[7] Mathkour, H., Touir, A., Al-Sanie, W.: Automatic information classifier using rhetorical structure theory. In: Klopotek, M. A., Wierzchon, S. T., Trojanowski, K. (eds) Intelligent Information Processing and Web Mining, Proceedings of the International IIS: IIPWM'05 Conference held in Gdansk, Poland, June 13-16, 2005. Advances in Soft Computing, vol. 31, pp. 229–236. Springer (2005)
[8] Matic, S., Iordanou, C., Smaragdakis, G., Laoutaris, N.: Identifying sensitive urls at web-scale. In: Proceedings of the ACM Internet Measurement Conference. pp. 619–633. IMC '20, Association for Computing Machinery, New York, NY, USA (10 2020) doi: 10.1145/3419394.3423653 , https://doi.org/10.1145/3419394.3423653
[9] McDonald, G., Macdonald, C., Ounis, I.: Using part-of-speech n-grams for sensitive-text classification. In: Proceedings of the 2015 International Conference on The Theory of Information Retrieval. pp. 381–384. ICTIR '15, Association for Computing Machinery, New York, NY, USA (9 2015) doi: 10.1145/2808194.2809496 , https://doi.org/10.1145/2808194.2809496
[10] OEA: Clasificacion de datos. En linea (2019), https://www.oas.org/es/sms/cicte/docs/ESP-Clasificacion-de-Datos.pdf
[11] OEA: Glosario-Proteccion de Datos Personales (2021), http://www.oas.org/es/sla/ddi/proteccion_datos_personales_glosario.asp
[12] sung Park, J., woo Kim, G., ho Lee, D.: 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. pp. 36–40. CNIOT2020, Association for Computing Machinery, New York, NY, USA (4 2020) doi: 10.1145/3398329.3398335 , https://doi.org/10.1145/3398329.3398335
[13] Union Europea: Grupo de trabajo del articulo 29l dictamen 4, 2007 sobre el concepto de datos personales. ¿que son los datos personales? (2007), https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2007/wp136_es.pdf
[14] Union Europea: Reglamento 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), https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX:32016R0679#d1e1547-1-1.
[15] Xu, G., Qi, C., Yu, H., Xu, S., Zhao, C., Yuan, J.: Detecting sensitive information of unstructured text using convolutional neural network. IEEE (oct 2019) doi: 10.1109/cyberc.2019.00087