2023 IEEE Symposium Series on Computational Intelligence (SSCI)
Mexico City, Mexico. December 5-8, 2023

Context-based classification of sensitive personal information

Sara De Jesús Sánchez
Instituto Politécnico Nacional
Centro de Investigación en Computación
CDMX, México
sdejesuss2100@alumno.ipn.mx

Eleazar Aguirre Anaya
Instituto Politécnico Nacional
Centro de Investigación en Computación
CDMX, México
eaguirrea@ipn.mx

Hiram Calvo
Instituto Politécnico Nacional
Centro de Investigación en Computación
CDMX, México
fcalvo@ipn.mx

Jorge Enrique Coyac Torres
Instituto Politécnico Nacional
Centro de Investigación en Computación
CDMX, México
jcoyact1900@alumno.ipn.mx

Raúl Acosta Bermejo
Instituto Politécnico Nacional
Centro de Investigación en Computación
CDMX, México
racostab@ipn.mx

***Abstract*— Sensitive personal information is at risk of exposure by the institutions it is shared. Institutions are responsible for preserving the privacy of the personal data they hold, even more so, in the case of sensitive data. ICIS, a model for context-based identification and classification of sensitive personal information, considers the context to identify personal data in unstructured texts of government type documents, regardless the size and type, and then classify each text segment as sensitive personal information, using natural language processing and machine learning techniques. ICIS not only indicates whether a text segment contains sensitive information or not, it also indicates personal data identified in each text segment, their location in the document and whether each text segment is classified as sensitive information. The main contributions of this work are both the identification of personal data and the classification of sensitive information based on the context, and the definition of sensitive personal information, in computational terms.**

Organizations must ensure the confidentiality of the personal information for which they are responsible, according to its level of sensitivity, value, and criticality [1], [2]. Personal data, in general, are those concerning an identified or identifiable natural person [3]. Sensitive personal data are a subset of personal data that refer to the most intimate sphere of its holder, or whose misuse may give rise to discrimination or entail a serious risk to it, such as racial or ethnic origin, present or future state of health, genetic information, biometrics, religious, moral, philosophical beliefs, political opinions, and sexual preferences. Sensitive personal data are subject to specific processing conditions, as they require robust protection [4], [5]. In Mexico, the laws that regulate the processing of personal data carried out by public and private institutions, establish high penalties, and in the case of sensitive personal data, the penalties can be increased up to twice [6]. Many NLP and ML techniques have been used for sensitive personal data classification [7], [8], [9], [10], [11], [12].

ICIS identifies 55 personal data types and classify them as sensitive information, in unstructured texts of government documents, where the writing style is formal and the documents vary in types and sizes. ICIS identifies the different personal data types in each text segment of a document using natural language processing (NLP), then vectorizes the personal data found in each text segment, and classifies each vector as sensitive personal information using machine learning (ML). Both identification and classification are completely related to the context of the speech.

In the identification process, NLP helped to contextualize each type of personal data, which were also identified by the grammatical pattern where they were found. ICIS analyzed not only the word or regular expression match, but also the grammatical pattern in the text segment where the data was found.

For the classification process, ICIS considered the personal context of the text segment. If a text segment contained sensitive data which was not related to a particular person, it was not classified as sensitive information. To be classified as sensitive information, there must be sensitive data related to a particular person. Four personal data classifications were proposed for each text segment, if it contains: identifiers (PID), personal data (PD), sensitive personal data (SPD, and sensitive unit data (SUD). With these data classification results, sensitive information classification rules were proposed to define, in computational terms, what sensitive personal information is (1).

$$SPI = SUD \vee PID \wedge SPD \quad (1)$$

where:

SPI: Text segment contains sensitive personal information

SUD: Text segment contains sensitive unit data, like Id numbers that may contain sensitive information within their formats.

PID: Text segment contains identifier personal data, like person names.

SPD: Text segment contains sensitive personal data.

Machine learning algorithms performances evaluated for the four personal data classifications were excellent, nB was the best. For the sensitive information classification, the DT algorithm performance was the best.

ICIS can be used as a system component. A path is the input; a json structure with the personal data findings, and the classification results for each text segment in the document, is the output.

This work was supported by Consejo Nacional de Ciencia y Tecnología (CONACyT), Instituto Politécnico Nacional (IPN), Comisión de Operación y Fomento de Actividades Académicas del IPN (COFAA), Programa de Estímulos al Desempeño de los Investigadores del IPN (EDI), Secretaría de Investigación y Posgrado del IPN (SIP), Convenio IPN-OAG-100-2021, Organization of American States (OAS), Cisco and the Citi Foundation, thanks to projects SIP 20222092, SIP 20211758 and the project Plataforma de Identificación, Clasificación y Monitoreo de Información Sensible (PICIS), winner of the Innovation Fund for Cybersecurity Projects in Latin America and the Caribbean 2021, created by OAS, Cisco and Citi Foundation.

Keywords— personal data, sensitive information, machine learning, classification, natural language processing, cybersecurity

REFERENCES

[1] Diario Oficial de la Federación, «Decreto promulgatorio del Protocolo Adicional al Convenio para la Protección de las Personas con respecto al Tratamiento Automatizado de Datos de Carácter Personal, a las Autoridades de Control y a los Flujos Transfronterizos de Datos», 2018. https://www.dof.gob.mx/nota_detalle.php?codigo=5539474&fecha=28/09/2018

[2] OEA, «Clasificacion de datos», OAS, 2019. https://www.oas.org/es/sms/cicte/docs/ESP-Clasificacion-de-Datos.pdf

[3] Comisión 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

[4] Comisión Europea, «¿Qué 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

[5] UE, «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», 2016. https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX%3A32016R0679#d1e1547-1-1

[6] INAI 2021, «Ley Federal de Protección de Datos Personales en Posesión de Particulares», 2021. https://home.inai.org.mx/?page_id=1870I\ &mat=p.

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

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

[9] Yan Liang, Zepeng Wen, Yizheng Tao, GogLiang Li, Bing Guo, «Automatic Security Classification Based on Incremental Leaning and Similarity Comparison», 2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC), Chongqing, China, 2019.

[10] Guosheng Xu, Chuhao Qi, Hai Yu, Shengwei Xu, Chunlu Zhao, Jing Yuan, «Detecting Sensitive Information of Unstructured Text Using Convolutional Neural Network», 2019 International Conference on Cyber-Enabled Network,» 2019 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), Guilin, China, 2019.

[11] Huimin Jang, Chunling Chen, ShengChen Wu, Yongan Guo, «Classification of Medical Sensitive Data based on Text Classification», 2019 IEEE International Conference on Consumer Electronics - Taiwan (ICCE-TW), Yilan, Taiwan, 2019.

[12] Ji-sung Park, gun-woo Kim, Dong-ho Lee, «Sensitive Data Identification in Structured Data through GenNER Model based on Text Generation and NER», Proceedings of the 2020 International Conference on Computing, Networks and Internet of Things, Sanya, China, 2020.