STClass: A Method for Determining the Sensitivity of Documents

Saturnino Job Morales Escobar[1](✉), José Ruiz Shulcloper[2], Cristina Juárez Landín[3], José-Sergio Ruiz-Castilla[4], and Osvaldo Andrés Pérez García[5]

[1] Centro Universitario UAEM Valle de México, Universidad Autónoma del Estado de México (UAEM), Atizapán de Zaragoza Estado de México, México city, México

[2] Research Group On Logical Combinatorial Pattern Recognition, Investigations, University of Informatics Sciences, Havana, Cuba
jshulcloper@uci.cu

[3] Centro Universitario UAEM Valle de Chalco, Universidad Autónoma del Estado de México (UAEM), Valle de Chalco Solidaridad Estado de México, Valle de Chalco, México

[4] Centro Universitario UAEM Texcoco, Universidad Autónoma del Estado de México (UAEM), Estado de México, Texcoco, México
jsruizc@uaemex.mx

[5] Equipo de Investigaciones de Minería de Datos, CENATAV - DATYS, La Habana, Havana, Cuba
osvaldo.perez@cenatav.co.cu

Abstract. The leakage of sensitive information is a pressing problem when information is processed digitally due to the economic, political and social repercussions that it can cause to its owner. Despite the risks and possible threats, the information must always be kept available to users, therefore, alternatives must be available to protect, detect, and prevent the leakage of sensitive information. A particular case of this problem is the leakage of sensitive textual documents. However, the identification of unstructured sensitive information is a problem whose solution is not totally satisfactory despite the development of methods and applications with promising results. Thus, it is necessary to continue developing methods that contribute to the effective solution of the problem based on a critical analysis of existing techniques and their future projections. In this work we start from a taxonomy of the approaches with which this problem has been approached. From the taxonomy, the critical analysis of the techniques and above all considering the practical needs, a method of solution to the problem of determining the sensitivity of textual documents is proposed from the perspective of Logical Combinatorial Patterns Recognition. The problem is approached as a supervised classification problem with two classes: sensitive and non-sensitive textual documents. The proposal in this work is the STClass method to determine the sensitivity of documents, which consists of two phases: the training phase, where the parameters for classification are defined and the classification phase. With the datasets used, 96% of the well classified documents were reached.

Keywords: Data leakage · Sensitivity of documents · Supervised classification · Logical combinatorial pattern recognition

I. Batyrshin et al. (Eds.): MICAI 2021, LNAI 13068, pp. 140–152, 2021.
https://doi.org/10.1007/978-3-030-89820-5_11

1 Introduction

Human activity, especially the ones that involve automated processes, generates and stores a large amount of data, textual documents (hereinafter, documents), images, videos, audios, etc., being undoubtedly one of the most valuable resources for any organization. For this reason, it must be protected, both to be preserved and to prevent its loss, and to prevent its dissemination at unauthorized instances. Documents that can be considered sensitive are of particular interest, but what should be understood by a sensitive document? A sensitive document "It is one that can not be made public" for reasons of personal or organizational privacy [1], or because it contains sensitive information and "the sensitivity of the information can be evaluated based on the impact that may result from its leakage." [2]. In the previous definitions it is assumed that once the sensitivity is determined it will not be modified, however, it is common for it to occur. Therefore, for the authors of this work, *the sensitivity of a document is an assessment of its importance, privacy and confidentiality at a given moment*.

Thus, due to their degree of sensitivity, some documents, such as those corresponding to intellectual property, financial information, patient information and personal data, should be restricted in use. However, in practice, they are used in activities that involve the use of computers and mobile devices, making them vulnerable to their theft or inappropriate use.

Unfortunately, due to the enormous generation and accumulation of these documents and our own human limitations, determining how valuable or sensitive they are, and in terms of this, preventing their escape or the commission of computer crimes, is a problem that it has increased dramatically.

On the other hand, document leakage can also occur intentionally or due to human errors, but it can be increased by the area in which its transmission is carried out (internal or external) or by the means used for its dissemination (electronic mail, instant messages, web page forms, among others) [3]. The risk increases when sensitive documents are shared by clients, business partners, external employees or when made available through social media and online services [4].

Under these conditions, sensitive data leakage can be seen as result of malicious attacks or by the accidental or involuntary distribution of sensitive data to an unauthorized entity [3, 6].

Against this background, the leakage of sensitive documents is considered an emerging problem of threat to personal and organizational security, not only because of its continuous growth and the financial losses it implies, but also because of the impact at the legal level, the possible suspension of operations and the loss of credibility and trust of its customers or users. In this work it is presented STClass, an effective method to address this problem.

The rest of the work is organized as follows. In Sect. 2, an analysis of the work related to the determination of the sensitivity of documents is presented and the proposal of the new method is in the environment of the Systems for the Prevention of Data Leakage. The STClass method, proposed in this work, is presented in detail in Sect. 3. Section 4 describes the datasets used in the training and classification phases, as well as the results obtained. Finally, the conclusions are presented in Sect. 5.

2 Related Work

To solve the problem, systems for Data Leakage Prevention (DLP) have been developed, which can identify, monitor and protect confidential data and detect its misuse. Typically, DLP systems add to traditional security measures by working well for well-defined and structured data [3].

To achieve the success of a system that offers a solution to data leakage considering the challenges it represents: the semantics associated with the data, sensitive data created without classifying, information exposure on social media platforms, electronic commerce, government-provided services, unstructured sensitive data problems described in [7–10] on DLP, it is necessary that, regardless of the application of the techniques for solving document leakage, methods are incorporated to automatically determine its sensitivity from the moment it is generated.

In general, the methods for determining the sensitivity of documents have been developed from two approaches: By context and by content [3, 5].

For context analysis, features related to the environment where the document is located are used, such as: document owner and assigned permissions, network protocols, encryption format, user role, web services used, web addresses, information associated with devices, among others. In some scenarios, it is enough to know the origin of the document to classify it as sensitive or not, making content analysis unnecessary. However, when the context is not categorical with respect to the sensitivity of the document, then it is necessary to perform the analysis of contents. In synthesis, the context analysis is more focused on characterizing the users and their environment than on the data that the document contains.

On the other hand, content-based sensitivity is tied to the meaning that the data may have. It is clear that in itself, each piece of data can contain a large amount of information, however, it can be increased or decreased if it is related to other data.

This work presents an analysis of the proposals developed from the content approach. Among the most used methods are those that are based on: regular expressions, classifiers, document fingerprints, *n*-grams, weighting or weighting of terms and natural language processing [3]. Here is an overview of these methods and the advantages and disadvantages of each.

Regular Expressions (RE): A set of terms or characters is searched to form detection patterns, they are normally used for partial or exact detection of social security numbers, credit cards, personal and corporate records. With the incorporation of techniques based on state compression [11] and use of specific dictionaries techniques, detection can be accelerated and improved, as shown in [12]. The RE works adequately by verifying predefined rules and quickly identify known data, among the disadvantages are, the difficulty to express the requirements through an RE, they apply only to regular languages, difficulty of developing finite automaton that recognize the generated language by the RE, only identify isolated strings and, where appropriate, the use of specific dictionaries.

The Classifiers: It is known that they depend considerably on an adequate classification of the data, otherwise the prevention systems will not be able to distinguish between sensitive and normal data. The usual practice is that the owner of the data is responsible for determining its sensitivity and the protection policy to apply. Most of the solutions have been based on labels, word lists and use of probabilities with their inheriting limitations, do not maintain semantic relationships between words, most only allow traits of numerical types, randomness and independence between data are assumed, and some require large corpus of text datasets for the training [13–16].

Fingerprint Methods: They are used especially in unstructured data to detect partial or exact matches. It is the most common technique used to detect information leakage. A wide exposition about fingerprinting approach is present in [17]. DLP systems with hashing functions such as MD5, and SHA1 can achieve a high level of accuracy with complete files without alterations, but changes in parts of the document can make this method ineffective [3]. Proposals have been made to overcome data corruption and maintain detection of sensitive data, for example, use of a fuzzy fingerprint algorithm [3] and the use of k-skip-n-grams [17]. Disadvantages: fail to detect small parts of the document, elimination of stop words which can generate loss of context, use of statistics in the selection of terms, excessive use of indexes and failure when sensitive data is altered or modified.

N-gram: Widely used in natural language processing, machine learning and information retrieval by term weighting. The n-grams depend on the frequency analysis of terms and n-grams in the documents. Its application together with Support Vector Machines have been used to classify business documents into two classes: sensitive and non-sensitive [18]; however, depending on the organization, the number of classes can be increased and labeled with different names, for example: public and private; private business, non-public business, and non-business; unclassified, restricted, confidential, secret and top secret [14, 17, 18]. A disadvantage is that once the value of n is set, it can not be changed for every n-gram, another disadvantage, is the elimination of stop words, and finally, they do not maintain semantic relationships between the terms.

Weighting or Term Weighing: The term weighing is a statistical method that indicates the importance of each term in a document, is called the weight of the term. This method is also used in text classification where each document is represented as a vector of dimension n in a vector space [19, 20] and n is the number of terms present in all documents. For term weighing, have been considered from binary weight schemes, the term is present or not [18], to schemes where functions are used to determine the frequencies of the terms based on which their weight is determined [14]. Improvements to the method include the representation of terms by means of n-grams as referred to in [3] or hybrid approaches that combine graphic and vector representations that include term weighing and classifiers [21, 22].

Natural Language Processing: Models to address this problem include the statistical one, in which each document is treated as a bag of words, where only the words and their relative frequencies are of interest to characterize a document. It is also generally assumed that there is a correctly categorized corpus that is used as a training set for supervised machine learning algorithms [14]. Other methods use weighted adaptive graphs, which allow maintaining the semantic sensitivity of the documents [23]. The application of syntactic analysis, the use of probability, identification of domains and corpus in those domains, are other alternatives that have been used to improve the identification of sensitive data from this perspective [24–26]. Among the limitations are: the use of domain corpus, frequently word order and context are ignored, production rules and use of probabilities are required.

It should be noted that in DLP systems, it is where most work has been done on solving the problem of determining the sensitivity of documents, However, due to the aforementioned limitations, it is necessary to continue the development of methods capable of identifying sensitive documents automatically, detecting the semantic content of the data to protect them in a pertinent manner to their sensitivity and prevent their leakage or loss in any of their states: in use, in transit or at rest.

Under these conditions, an important contribution would be aimed at detecting sensitive documents from the moment they are created and before they are released for use. With this objective, STClass was developed, a supervised method capable of determining whether a document is sensitive or not. This method has the following advantages: it does not require a list of words provided by the owner of the data, or specific corpus, it maintains the relationships between the terms to preserve their semantic relationship, it does not presuppose probability distributions over the terms or chains of terms present in documents, the length of sensible strings can be chosen and finally it can be used for document classification. The STClass method is described in the next section of this work.

Considering that the basic architecture of DLP systems is made up of three modules (see Fig. 1). The first detects whether a document is being sent, created or accessed (for printing, copying, editing, sending over the network, etc.) regardless of its content. The second module analyzes the document detected in the filter, reviews it and sends it to the third module for an assessment in accordance with the established policy. This last module responds by allowing access or blocking, if is necessary, the actions on the document to be protected, issuing the corresponding alert.

Is in the analysis module at the content or context level, or both, where the sensitivity of documents must be assessed and based on which will be the response that the system must issue. This answer is qualified by the level of security desired with the DLP application, thus expressing the security policy defined for the system. It is precisely in this module where the theoretical problems related to the determination of the sensitivity of the documents to be protected are located and where the STClass method can be incorporated.

The STClass method is described in the next section.

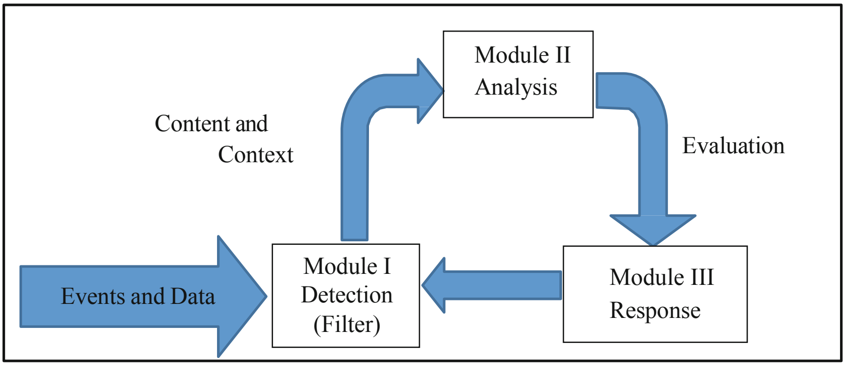

Fig. 1. Basic architecture of the DLP.

3 STClass: Method for Determining the Sensitivity of Documents

Based on the analysis carried out, a new method is presented to determine whether a document is sensitive or not, from the point of view of Logical Combinatorial Pattern Recognition [27, 28]. The method is developed with a content-based approach.

From this perspective, the determination of the sensitivity of documents is approached as a problem of supervised classification considering two stages: training and classification.

In this work, the STClass method is applied with documents classified into two classes: sensitive and non-sensitive documents. Unlike other methods, an initial list of terms is not required.

For the analysis of the document, the point "." is used to delimit the strings of terms that could form sentences or phrases.

In the first instance, as a comparison criterion to decide whether two terms are similar, symbol-by-symbol equality is used, which means that two terms are similar if all the symbols of the term are equal. Without any further modification to the method, other criteria for comparison between terms could be used, e.g., synonyms.

A document will be considered sensitive if it contains a sensitive paragraph and a paragraph is sensitive if it contains a string or substring of terms sensitive, analogously, a string or substring of terms is sensitive if it contains any sensitive term and a term is sensitive, if its frequency in the class of sensitive documents is higher than its frequency in the class of non-sensitive documents by at least a given threshold $\in$, as is show in Eq. (1).

$$S(t) = \begin{cases} 1 if\, |F_s(t) - F(t)| \geq \in \\ 0\, in\, another\, case \end{cases} \tag{1}$$

$Fs(t)$ is the frequency of term t in the class of sensitive documents and $F(t)$ is its frequency in the class of non-sensitive documents.

As can be seen, from Eq. (1), in general, the degree of sensitivity of a term can be expressed as:

$$\mathrm{S_n(t)} = \frac{\mathrm{F_s(t) - F(t)}}{\mathrm{F_t(t)}} \tag{2}$$

$F_s(t)$ y $F(t)$ are defined as in (1) y $F_t(t)$ is the total frequency of the term, given by the Eq. (3)

$$F_t(t) = F_s(t) + F(t) \tag{3}$$

Analogously, the sensitivity of a string or substring of terms, "c", is defined by:

$$S(c) = \begin{cases} 1 if\, |F_s(c) - F(c)| \geq \partial \\ 0\, in\, another\, case \end{cases} \tag{4}$$

Here $F_s(c)$ is the frequency of the string (substring) c in the class of sensitive documents. ∂ is a given threshold and $F(c)$ is its frequency in the class of non-sensitive documents.

It can also define the degree of sensitivity of c as:

$$S_n(c) = \frac{F_s(c) - F(c)}{F_t(c)} \tag{5}$$

$F_s(c)$ y $F(c)$ are defined as (4) y $F_t(t)$ is the total frequency of the term, given by the Eq. (6).

$$F_t(c) = F_s(c) + F(c) \tag{6}$$

The method works as follows, a counter is associated for each term, in such a way that the number of times the term appears in each class is known.

A string of terms is the concatenation of terms delimited by ".".

For the formation of chains, a term is taken and the following terms are linked to it until the delimiter is found. For each substring, the number of times it is present in each class is also stored.

Because of the way the strings are obtained, the relationships between the terms are preserved, which represents an advantage when preserving the information that the terms provide together.

The STClass method consist of 2 phases, and each phase is divided into three steps and in each one the parameters with which the method will work must be provided.

STClass Method

Phase 1. Training

Step 1. Build a training matrix with samples of documents in each class (sensitive and non-sensitive).

Step 2. Construction of chains (substrings) of terms for the training sample.

2.1. Build a list of initial terms of each chain.

2.2. Determine comparison criteria between terms.

2.2.1 Equality

2.2.2 Synonyms

2.3. Build the chains (substrings) of each term in the list.

2.4. For each term and each substring calculate its frequency in each class.

2.5. Determine the sensitivity of each term and each substring of terms using (3) and (4).

Step 3. Definition of parameters for classification.

3.1. Quantity of sensitive terms to use.

3.2. Number of sensitive substrings to use.

3.3. Definition of the similarity function to be used in string comparison.

3.4. Determine sensitivity weights for each substring.

3.4.1. Based on the cardinality of the substrings.

3.4.2. Based on the frequency of terms.

3.5. Classification rule.

Phase 2. Classification

Step 4. Construction of chains (substrings) of terms of the document to classify.

4.1. Build list of initial terms of each chain.

4.2. Build the chains (substrings) of each term in the list.

Step 5. Compare the strings of terms in the document to be classified, with the chains

of terms obtained in step 2 and parameters given in step 3.

Step 6. Classify the document based on the classification rule (step 3.4).

4 Datasets and Experimental Methodology

To test the STClass method a training sample was formed with 80 textual documents and 60 control documents, achieving an efficiency in the 96% classification considering the terms and chains of sensitive terms with an epsilon threshold equal to zero.

Table 1 summarizes the distribution of the documents used in the two phases of the STClass method. For the training phase, a set of 50 sensitive and 30 non-sensitive documents was used. In the classification phase, 60 documents were introduced, of which 40 were sensitive and 20 were non-sensitive. Taking as a function to evaluate the quality the quantity of well classified documents among the total of documents an efficiency of 96% was obtained.

Table 1. Distribution of documents used for training and classification of STClass.

Datasets	Trainning		Classification		Results		Efficiency
	Sensitive	Non-sensitive	Sensitive	Non-sensitive	Well classified	Misclassified	
Documents	50	30	40	20	58	2	96%

In the determination of sensitive documents, seen as a problem of supervised classification, it is essential that the training data set is true and maintained as it is given by the owner of the data. For this reason, the content of the training and classification files should be entered into the program in its original format and content, avoiding the elimination or modification of terms that may be significant in identifying sensitivity and avoiding biases, in some cases, only delimiters were added.

Nevertheless, given the difficulty of obtaining real and current sensitive data, the dataset was formed by taking records published on the internet and were modified to avoid coincidences that compromise some real instance. Below are examples of the contents of the documents used in the training samples and in the classification.

Training sample

Sensitive documents:

$D_{1:}$ Empresa de Viajes Cia. Ltda
Guipuzcoa G2-117 y Moreno.
La marqueza / Quito, Ecuador.
EE. UU. Teléfono: 323-675-874 / consulas@andec.com.
En Guipuzcoa.
Licenciado Leonardo Moreno.
Presidente Constitucional de la República de Ecuador.

D_2: Eduardo Estévez Romero.
Investigador Asociado.
Editor asociado de Transactions on Neural Networks Journal.
Correo electrónico: latevez@ing.ucil.cl.
Eduardo Estévez Medina.
Es secretario de cultura.
El presidente de la República es electo por elecciones directas y por un periodo de 6 años.
No hay reelección a la Presidencia de la República.

Non-sensitive documents:

D_3: Alberto Magaña Mercado.
Director del Departamento de Ingeniería Eléctrica.
Profesor titular (2009-).
Editor asociado de IEEE Transactions on Neural Networks Journal (2000-2005).
Copresidente General Congreso Mundial de IEEE sobre Inteligencia Computacional, Río de Janeiro, Brasil, julio de 2004.

D_4: Gerardo Anaya de Isla Galapagos, trabajó como Marino Mercante, ocupando la función de aceitero.
Acuedo con usted para denunciar al señor José Gabriel Olvera Verlanga, de la compañia island travel cia.
Esta compañía y su embarcación trabaja en la provincia de Isla Galapagos pertenece a la compañía ISLANDIA CON RU. No.1U5737001.

Documents to classify:

D_5: Empresa de Viajes Cia. Ltda
Guipuzcoa G2-117 y Moreno.

D6: Actualmente es profesor de Ingeniería Eléctrica de la Universidad de Chile.
Eduardo Estévez Mediana.
Fue director del Departamento de Ingeniería Eléctrica en los años 2006-2012.

With the previous training documents, the term strings are built and the frequency of the term and the substring in each class is recorded, in such a way that when a document is going to be classified, the strings obtained from the document to be classified are compared with the chains obtained during the training phase.

In this way, all the terms, the order in which they are found and the semantic relationships between them are maintained, it is here that the fulfillment of the similarity functions defined for the comparison of the strings and the comparison criteria of terms is verified.

After applying the method, the strings that were sensitive and that give rise to the classification of the documents as sensitive, are shown below.

This is because the strings were found in the sensitive documents class and are not present in the contents of non-sensitive documents.

As noted in this document, STClass does not require a list of sensitive terms, syntactic rules, specific corpus, or the calculation of probabilities to determine the sensitivity of documents.

Sensitive Chains:
Empresa de Viajes Cia. Ltda.
Guipuzcoa G2–117 y Moreno.
Eduardo Estévez Mediana.

5 Conclusion

The problem of determining the sensitivity of documents is a problem that presents great challenges, in principle by itself's nature, the temporality of their sensitivity, the large number of documents that are generated and most significantly, the dependence of sensitivity to natural language semantics.

In this work, based on an analysis of the most used methods to classify sensitive documents and the need to know if a document should be protected or not, limitations and requirements present in these methods were detected, among which are: it must be provided a lists of sensitive words to be able to identify them, the elimination of words can cause the loss of semantic relationships, the comparison between documents is based on numerical vectors, the order of the words is ignored, use of specific corpus, provide grammar rules and the use of probabilities.

Based on this identification, the STClass method was developed to determine the sensitivity of documents from the point of view of the Logical Combinatorial Pattern Recognition. The method is based on content analysis and is approached as a problem of supervised classification considering two phases: training and classification.

In this work, its application to the problem is presented with two classes: sensitive and non-sensitive documents.

Among the advantages offered by STClass are the following: it does not require an initial list of terms, it maintains the order and the semantic relationship between terms, it does not need specific corpus or a priori probabilities, to evaluate the similarity, different comparison criteria can be incorporated between terms and between strings of terms. In addition, its application can be extended to problems with a larger number of classes.

In the tests carried out, an efficiency of 96% was achieved in the classification of new documents.

As a continuation of this work, it is necessary to test the method with public and articles datasets, make comparisons with other methods, and do the extension to include any number of classes and provide degrees of sensitivity.

References

1. Berardi, G., Esuli, A., Macdonald, C., Ounis, L., Sebastiani, F.: Semi-automated text classification for sensitivity identification. In: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. ACM, pp. 1711–1714, (2015)
2. Alzhrani, K., Ruddy, E., Chow, C., Boulty, T.: Automated U.S diplomatic cables security classification: topic model pruning vs. classification based on clusters. In: Proceedings of the 2017 IEEE International Symposium on Technologies for Homeland Security (HST), pp. 1–6, (2017)
3. Alneyadi, S., Sithirasenan, E., Muthukkumarasamy, V.: A survey on data leakage prevention systems. J. Netw. Comput. Appl. **62**, 137–152 (2016)
4. Salahdine, F., Kaabouch, N.: Social engineering attacks: a survey. Future Internet, **11**(4), 89 (2019)
5. Wynne, N., Reed, B.: Magic quadrant for enterprise data loss prevention. Gartner Group Research Note (2016)
6. Ahmad, N.: Do data almost always eventually leak?: Computer **54**(2), 70–74 (2021)
7. Wadkar, H., Mishra, A., Dixit, A.: Prevention of information leakages in a web browser by monitoring system calls. In: Proceedings of the 2014 IEEE International Advance Computing Conference (IACC), pp. 199–204, (2014)
8. Liu, T., Pu, Y., Shi, J., Li, Q., Chen, X.: Towards misdirected email detection for preventing information leakage. In: Proceedings of the 2014 IEEE Symposium on Computers and Communication (ISCC), pp. 1–6, (2014)
9. Jena, M.D., Singhar, S.S., Mohanta, B.K., Ramasubbareddy, S.: Ensuring data privacy using machine learning for responsible data science. In: Satapathy, S.C., Zhang, Y.-D., Bhateja, V., Majhi, R. (eds.) Intelligent Data Engineering and Analytics. AISC, vol. 1177, pp. 507–514. Springer, Singapore (2021). https://doi.org/10.1007/978-981-15-5679-1_49
10. Ávila, R., Khoury, R., Khoury, R., Petrillo, F.: Use of security logs for data leak detection: a systematic literature review. Secur. Commun. Netw. (2021)
11. Becchi, M., Crowley, P.: An improved algorithm to accelerate regular expression evaluation. In: Proceedings of the 2007 ACM/IEEE Symposium on Architecture for Networking and Communications Systems, pp. 145–154, (2007)
12. Sokolova, M., et al.: Personal health information leak prevention in heterogeneous texts. In: Proceedings of the Workshop on Adaptation of Language Resources and Technology to New Domains, pp. 58–69, (2009)
13. Chen, K., Liu, L.: Privacy preserving data classification with rotation perturbation. In: Fifth IEEE International Conference on Data Mining (ICDM'05), pp. 1–4, (2005)
14. Brown, J.D., Charlebois, D.: Security classification using automated learning (SCALE): optimizing statistical natural language processing techniques to assign security labels to unstructured text. Defense Research and Development Canada, Ottawa (Ontario), (2010).
15. Kowsari, K., Jafari, M., Heidarysafa, M., Mendu, S., Barnes, L., Brown, D.: Text classification algorithms: a survey. Information **10**(4), 150 (2019)
16. Zorarpacı, E., Özel, S.A.: Privacy preserving classification over differentially private data. Wiley Interdisc. Rev. Data Min. Knowl. Discov. **11**(3), e1399 (2021)

17. Shapira, Y., Shapira, B., Shabtai, A.: Content-based data leakage detection using extended fingerprinting. arXiv prepint arXiv:1302.2028 (2013)
18. Hart, M., Manadhata, P., Johnson, R.: Text Classification for data loss prevention. In: Fischer-Hübner, S., Hopper, N. (eds.) PETS 2011. LNCS, vol. 6794, pp. 18–37. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-22263-4_2
19. Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. Commun. ACM **18**, 613–620 (1975)
20. Carvalho, V.R., Balasubramanyan, R., Cohen, W.W.: Information leaks and suggestions: a case study using mozilla thunderbird. In: CEAS 2009-Sixth Conference on Email and Anti-Spam (2009)
21. Xiang, Y., Zhihong, T., Jing, Q., Feng, J.: A data leakage prevention method based on the reduction of confidential and context terms for smart mobile devices. Wirel. Commun. Mob. Comput. (2018)
22. Katz, G., Elovici, Y., Shapira, B.: CoBAn: a context based model for data leakage prevention. Inf. Sci. **262**, 137–158 (2014)
23. Xiaohong, H., Yunlong, L., Dandan, L.: A novel mechanism for fast detection of transformed data leakage. IEEE Xplore Digit. Libr. **6**, 35926–35936 (2018)
24. Yang, Z., Liang, Z.: Automated identification of sensitive data from implicit user specification. Cybersecurity **1**(1), 1–15 (2018). https://doi.org/10.1186/s42400-018-0011-x
25. Neerbek, J., Assent, I., Dolog, P.: Detecting complex sensitive information via phrase structure in recursive neural networks. In: Phung, D., Tseng, V.S., Webb, G.I., Ho, B., Ganji, M., Rashidi, L. (eds.) PAKDD 2018. LNCS (LNAI), vol. 10939, pp. 373–385. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-93040-4_30
26. Briand, A., Zacharie, S., Jean-Louis, L., Meurs, M.-J.: Identification of sensitive content in data repositories to support personal information protection. In: Mouhoub, M., Sadaoui, S., Ait Mohamed, O., Ali, M. (eds.) IEA/AIE 2018. LNCS (LNAI), vol. 10868, pp. 898–910. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-92058-0_86
27. Martínez-Trinidad, J.F., Guzmán-Arenas, A.: The logical combinatorial approach to pattern recognition, an overview through selected works. Pattern Recogn. **34**, 741–751 (2001)
28. Ruiz-Shulcloper, J.: Pattern recognition with mixed and incomplete data. Pattern Recogn. Image Anal. **18**(4), 563–576 (2008)