ARTPHIL: Reversible De-identification of Free Text Using an Integrated Model

Bayan Alabdullah[1,2(✉)], Natalia Beloff[2], and Martin White[2]

[1] Computer Science Department, Princess Nourah Bint Abdul Rahman University, Riyadh 11543, Saudi Arabia
b.alabdullah@sussex.ac.uk

[2] Department of Informatics, University of Sussex, Falmar BN1 4GE, UK

Abstract. Organisations that collect and maintain individual data face the challenge of preserving privacy and security when using, archiving, or sharing these data. De-identification tools are essential for minimising the privacy risk. However, current data de-identification and anonymisation methods are widely used to alter the original data in a way that cannot be recovered. This results in data distortion and, hence, the substantial loss of knowledge within the data.

To address this issue, this paper introduces the concept of reversible data de-identification methods to de-identify unstructured health data under the Health Insurance Portability and Accountability Act (HIPAA) guidelines. The model integrates Philter [9], the state-of-the-art tool for extracting personal identifiers from free-text, to detect confidential information and encrypt them with E-ART, lightweight encryption algorithm E-ART [10]. The performance of the proposed model ARTPHIL is evaluated using i2b2 data corpus in terms of recall, precision, F-measure and execution time. The results of the experiment are consistent with the recent de-identification method with recall of 96.93%. More importantly, the original data can be recovered, if needed, and authenticated.

Keywords: Privacy · De-identification · Pseudonymisation · Reversible · Re-identification

1 Introduction

Preserving data privacy has become a significant issue, while the applications and capabilities of Big Data are expanding dramatically. There is no doubt that this expansion creates enormous opportunities and avenues to understand and solve significant problems over various domains. However, the privacy and security concerns about Big Data are also growing. Legal systems establish laws to protect the privacy of individual information. A well-known example is the General Data Protection Regulation (GDPR), which outlines a specific set of rules for sharing and storing personal data to protect individual privacy. Data minimisation is one of the fundamental principles of GDPR and has strict data retention policies. This protection means that an individual's personal data can be retained for no longer than necessary to carry out the purpose for which the data is

W. Shi et al. (Eds.): SPNCE 2021, LNICST 423, pp. 369–381, 2022.
https://doi.org/10.1007/978-3-030-96791-8_27

processed [1]. GDPR also restrict the use of personal data beyond the purpose for which the data was originally collected (purpose limitations). However, these policies could be relaxed when data is de-identified. GDPR also encourage data controllers and processors to de-identify data to reduce the liability and notification obligations for data breaches [1]. Data security and privacy issues become even more critical when the data is used in healthcare environments, which typically deal with patient-sensitive information. In the United States, standards for protecting healthcare information confidentiality were established in HIPAA's Privacy Rule [2]. HIPAA specifies 18 data elements that consider identifiable health information. Those elements must be removed or generalised from data to be considered de-identified.

Previous work on data de-identification and anonymisation techniques has been developed within the fields of statistical disclosure control [3], privacy-preserving data publishing (PPDP) [4] and privacy-preserving data mining (PPDM). In these fields, sensitive data are shared with untrusted third parties for secondary use but do not disclose information that can be linked to specific individuals. This technique primarily focuses on producing anonymised versions of data by removing, obfuscating or generalising identifiable personal data. The drawback of these methods is that they are usually irreversible; there is no mechanism by which an individual's identification can be recovered.

In many circumstances, it is essential to refer back to the original data without revealing it to the end-user. This allows it to be accessed in emergencies or by those with acceptable levels of access. For example, in the PPDM field, the knowledge obtained from mining de-identified data cannot be verified from the original data, which might cause knowledge uncertainty [5, 6]. In research platforms, where documents are shared between different specialists, including scientists, researchers, and physicians, specific cases are usually chosen for further analysis. For instance, in selected cases where specialists conduct blind diagnosis or annotation procedures, the authorised user can reverse the de-identification process and retrieve the necessary information, making a more accurate assessment [7].

Further, there are cases in clinical trials in which some of the research subjects should be approached again for further study [8]. Therefore, reversible de-identification (e.g., encryption, pseudonymisation) are often preferred. Although robust encryption solutions are available, their application in the face of ever-increasing volume, variety, and speed remains challenging. In addition, the cumbersome key management and distribution of the symmetric encryption algorithm prevent a suitable level of scalability. In addition, more lightweight and practical alternatives must be developed [30].

As a result, there is a need for a new reversible model for de-identifying unstructured data to reduce the risks for data subjects, address the requirements of preserving privacy while supporting data's use. This study addresses this need by integrating Philter [9], the information extracting tool, with the lightweight encryption algorithm E-ART [10]. This process's primary contribution is 1) a reversible, fast de-identification model ARTPHIL to de-identify unstructured textual health data cost-effectively without compromising security. And 2) implementation and evaluation of the proposed model using 2014 i2b2 testing data. The technique achieved a recall of 96.93% to detect and encrypt personal health information (PHI) specified under HIPAA guidelines [2]. However, this work does not claim to generate de-identified data to comply with HIPAA as the PHI is

encrypted, not removed. The method also achieved fast execution time, making it suitable to de-identify a considerable amount of data.

2 Related Work

2.1 De-identification as Named Entity Recognition

The de-identification of structured data has been widely studied, and there are various techniques [11]–[13]. However, de-identifying unstructured data, primarily text data, is complex and requires researchers' manual intervention. The main challenge in de-identifying unstructured data is finding the sensitive attributes that spread throughout the text document.

Several techniques have been proposed to extract those sensitive attributes. Most of them can be seen in the application Named Entity Recognition (NER). NER is a technique that finds and categorises important words inside the text [14]. Many different natural language processing (NLP) applications benefit from the NER technique. Those applications include questioning-answering applications [15, 16], tweet analysis [17, 18], automatic text extraction [11, 12] and data mining applications.

In the de-identification process, NER is a useful tool for extracting identifiable and sensitive attributes from unstructured text. NER techniques can be classified into three main categories: 1) rule-based, 2) machine learning, and 3) hybrid [14]. A brief description of each method is provided below.

Rule-based: The rule-based method consists of rules, such as pattern matching, hand-crafted, heuristics, grammatical and dictionaries to recognise NER in an unstructured text [19]. This method's advantage is that it requires little or no annotated training data and is easy to implement and improve by adding additional rules. However, it is quite expensive, domain-specific and lacks robustness and portability. Examples of de-identification tools that use rule-based and pattern matching methods to detect personal health information (PHI) entities and replace them with tags indicating its category include De-ID [20], HMS Scrubber [21], and PhysioNet de-identification (deid) software [22].

Machine learning: The machine learning approach trains a statistical model to classify words into a PHI or a non-PHI group. Most recent de-identification tools use supervised learning algorithms such as support vector machines [23], conditional random fields [24], and decision trees [25]. The results obtained through supervised machine learning techniques are promising. However, these techniques require extensive, annotated data for training. Creating annotated data is an expensive task because it requires substantial time and effort with domain experts' support.

Hybrid model: The hybrid model approach combines lexicon and rule-based methods to benefit from and overcome the limitations of both [26]. For example, MITRE Identification Scrubber [27] uses pattern matching to extract all numeric data, such as phone numbers and postal codes and uses them in a conditional random fields algorithm.

An example of a recent text identification application is Philter [9]. Philter is a customisable open-source de-identification software developed and evaluated with an extensive collection of unstructured clinical notes from the University of California, San Francisco (UCSF) and 2014 i2b2. The algorithm uses rule-based and statistical NLP

approaches. The algorithm uses an overlapping pipeline of methods that are state-of-the-art in each application. This combination helps classify PHI in a free-text document, including regular expressions, statistical modelling, blacklists, and whitelists, as shown in Fig. 1. The word identified as PHI is replaced with an obfuscated string of precisely the same length (for example, "David Mitchell" becomes "***** ********").

The algorithm's performance compared in [9] to the two strongest real-world competitors, PhysioNet [22] and Scrubber [27], based on recall. Philter demonstrated the highest overall recall on both corpora and the highest recall in each PHI category. A significant drawback of this approach is that it is irreversible. The removed PHI data cannot be re-generated if an authenticated user later requires full access to the data.

2.2 Reversible De-identification

To our knowledge, most of the data de-identification methods have been developed in the fields of PPDP and PPDM. These approaches protect private data from being disclosed during data mining. It often swaps, deletes or modifies the identifiable information and removes the correlation between the original and anonymised data. Consequently, it is unable to recover the original data from the de-identified data, which could create issues such as knowledge uncertainty [5, 6, 28]. For example, if original data is lost, the mined data's knowledge cannot be verified from the original data [5]. To resolve this drawback, Chen et al. [5] used the concept of reversible data hiding in the image and proposed an algorithm called privacy difference expansion (PDE). PDE perturbed and embedded private data with a customised watermark to verify the integrity of the original data. Similarly, Yamac et al. [29] proposed privacy-preserving solutions that combine a multi-level encryption scheme with compressive sensing. This approach can reverse the de-identification so that an authorised person can recover the degraded information using a key. It attempts to simplify the key management issue by watermarking the key into the sensed image. However, this computationally expensive decompression might be challenging to apply to Big Data.

Encryption is also considered an efficient method to obtain reversible de-identification [8]. For example, Landi and Rao [8] proposed a way to de-identify patient data so that only the owners of the original data or legally empowered entities can re-identify. It uses secure public-key encryption technology to generate a public key based on one or more private keys. Gulcher et al. [26] stated the need for reversible de-identification to protect genetic research data. They proposed an approach for de-identifying biological samples for genetic research based on a third-party encryption method using a 128-bit symmetric encryption algorithm-AES. This approach would allow later requested access to the research data. However, The existing encryption standards rely on increasing the key size and the number of rounds to enhance security which could negatively affect the performance [29, 30]. E-ART is a new lightweight encryption algorithm that was proposed in [10] to address the speed requirement by modern application. It uses the concept of a balanced binary tree with ASCII conversion and random key generation. The algorithm compared to existing standard encryption algorithms in terms of performance and security with promising results.

The terms anonymisation and de-identification are often used interchangeably. However, there is a significant difference between these concepts. Clete A. Kushida et al.

[35] state that de-identification of data refers to the process of deleting or obscuring any personally identifiable information from individual records in a way that reduces the risk of disclosure of the data subject identity. However, de-identified datasets are allowed to contain encrypted identifiers where only authorised users have access to the encryption key. The existence of a key makes it possible to recover the original data for the user with correct authorisation. This process is also known as pseudonymisation under GDPR privacy rules. Anonymisation, on the other hand, refers to the process of data de-identification that produces de-identified data that cannot be reversed back to the original data. Under GDPR, anonymous data is not treated as personal data. Therefore, user consent to process and share the data is not required.

Pseudonymisation is a new privacy-preserving concept introduced by the GDPR as "The processing of personal data in such a way that the data can no longer be attributed to a specific data subject without the use of additional information, as long as such additional information is kept separately and subject to technical and organisational measures to ensure non-attribution to an identified or identifiable person" [1].

Pseudonymisation is a particular type of de-identification in which the names and other information that directly identifies an individual are replaced with pseudonyms. Pseudonymisation enables linking personal data through various datasets when all identifiable information is consistently pseudonymised. Pseudonymisation can be reversed when the link between original identities and the pseudonyms is maintained or if the replacement is done with an algorithm whose parameters are known. This provides an option for the de-identification process to be reversed in the future and re-identifying the data subjects. For example, identifiable information can be encrypted with a secret key to create a pseudonym; decrypting the key reversed the pseudonymisation process, recovering the original identifier. Pseudonymised data is still personal data and cannot be equated to anonymised data. Under HIPAA, re-identifying the datasets may only be done by an organisation covered by HIPAA's rules (mostly healthcare providers), known as a covered entity.

Pseudonymisation of data is suggested by GDPR [1] as one of the protective measures that controllers can use to evaluate the feasibility of further processing of personal information for archiving purposes in the public interest, scientific or historical research purposes, or statistical purposes by processing data that do not enable or no longer enable the identification of data subjects. However, GDPR restricts a data handler's potential to benefit from pseudonymised data if re-identification processes are "reasonably likely to be employed, such as singling out, either by the controller or by another person to identify the natural person directly or indirectly" [1]. Hence, data controllers should implement several technical and organisational measures to ensure that pseudonymous data is disconnected from the key enabling re-identification. Furthermore, the risks of re-identification are dynamic and evolve over time, and this implies that data controllers should evaluate these risks on a regular basis and take necessary action when they become significant. For example, changing pseudonyms over time for each use or each type of use as a way to reduce the risk of re-identification through linkability [36].

To sum up, the GDPR provides several regulatory incentives to adopt pseudonymisation. There are, therefore, significant benefits associated with using it, which include enabling data processing for secondary purposes without the need to obtain the explicit

consent of data subjects. However, for this exemption to apply, pseudonymisation should meet the GDPR standard, and the existing pseudonymisation techniques were developed long before GDPR requirements were established. Many implementations of pseudonymisation approaches use static pseudonyms for data subjects, while others may contain indirect identifiers; in both cases, these fail to protect against re-identification due to privacy breaches arising from linkage attacks.

3 Proposed Method

The proposed de-identification system ARTPHIL consists of two key components that are integrated to de-identify unstructured textual data as follow: 1) the core of the Philter package [9] for the PHI detection process onlynot for the replacement, and 2) the E-ART encryption algorithm [10] for replacement strategy. Figure 1 presents an illustration of this method. The steps and dataset details are explained in the following sections.

3.1 PHI Detection

To address the task of locating PHI, we used an overlapping pipeline of multiple state-of-the-art methods provided by Philter [9]. The pipeline includes pattern matching, statistical modelling, blacklist, and whitelist to detect PHI from free-text clinical notes. The detection process involves scanning the unstructured text line by line and dividing it into individual words. First, common words with a high probability of not being PHI are detected using pattern matching with a custom library of 133 "safe" regular expressions. Second, a customised library of 171 regular expressions is used to locate known PHI entities such as salutations, ID numbers, phone numbers, date of birth, email address and zip codes. In both scenarios, the regular expressions look for exact terms, phrases, or numbers to recognise matches using each word's immediate context. The algorithm uses statistical modelling to determine each sentence and document's structure to address the challenge of dealing with words that could be either safe or PHI, such as "white" might be a name or an adjective. To exclude names that are proper nouns, a customised blacklist is used. And the whitelist is used to preserve all the medical terms and common English words. At the end of the pipeline, a token has one of three potential labels: PHI, Non-PHI or unmarked.

3.2 Replacement Strategy

To achieve our goal of developing reversible de-identification, all tokens marked as PHI, and unmarked tokens will be replaced with generated strings that can be retrieved. For this purpose, we used E-ART, the lightweight encryption algorithm proposed in [10]. This algorithm uses the idea of the balanced binary tree along with an ASCII table for character substitution; this increases searching efficiency and reduces processing time as reported in [10]. The algorithm also uses a random key generator based on character position as a seed to generate a dynamic key for increased security. However, reverse engineering of the original values is preserved if needed.

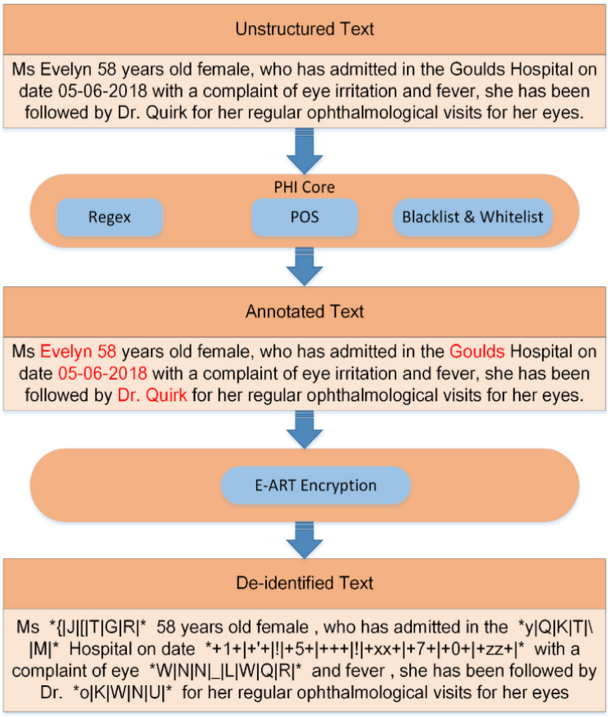

Fig. 1. A conceptual overview of the ARTPHIL process.

The system will look for PHI entities labelled in the PHI detection phase and use the E-ART algorithm to encrypt them. E-ART uses the entity's characters' index as a seed to generate a pseudo-random value. This value will be used as a key to generating an encrypted string to replace the PHI entity. The encrypted PHI entity's start and end will be marked with asterisks to be recognised if recovery of the original data is needed, as shown in the de-identified text in Fig. 1.

During the regeneration process of the original text, the de-identified text will be first scanned line by line, and all words that start and end with asterisks will be labelled as encrypted PHI. Then all encrypted PHI will be decrypted by reversing the process of E-ART encryption as described in [10].

4 Evaluation Metrics

4.1 Precision, Recall and F-measure

Because the proposed de-identification system and most existing de-identification systems treat the PHI identification as a NER task, the evaluation of the identification should use the same metrics as the NER literature [33]. Specifically, we used the 2014 i2b2 annotated data set and report performance, primarily we used three metrics: precision, recall and f-measure. Precision, also called positive predictive value, is the fraction of

the relevant tokens (correctly classified) among the retrieved tokens. recall, also called sensitivity, is the fraction of relevant tokens (correctly classified) that were retrieved. The output of a classifier can be presented in a confusion matrix, which shows the number of true-positive annotations (TP), true-negative annotations (TN), false-positive annotations (FP) and the number of false-negative annotations (FN).

Precision (Eq. 1) and recall (Eq. 2) can be computed from such a matrix. F-measure (Eq. 3) is the weighted mean of precision and recall. recall answers the question, "Did we find all that we were looking for?" and precision answers the question, "Did we only label what we were looking for?" A high recall is generally preferred over high precision because it measures PHI percentage correctly identified; the data subjects' privacy is prioritised over a potential loss of document interpretability [34]. To emphasise sensitivity, we also calculate the F1 measure, which is the harmonic mean of the precision and recall, and the F2 which weighs recall (twice) higher than precision, as defined in (Eq. 4):

$$Recall = \frac{TP}{TP + FN} = \frac{\#\ of\ corrected\ entites}{\#\ of\ corrected\ prediction} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} = \frac{\#\ of\ corrected\ entites}{\#\ of\ expected\ prediction} \tag{2}$$

$$F_1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{3}$$

$$F_{2=\frac{(\beta^2+1)\times Precision\times Recall}{(\beta^2\times Precision+Recall}, \beta=2} \tag{4}$$

4.2 Re-identification Risk

To estimate the re-identification risk of our proposed de-identification system, we calculate the conditional probability of a leak in identifiable information [33] as follows:

$$Pr(Reidentification, leak) = Pr(Reidentification|leak) \times Pr(leak)$$

In a de-identification tool, leaks can occur if a PHI entity is not detected because it can be used to re-identify the individual whom the data describe. So, we assume that:

$$P(Reidentification|leak) = 1$$

As we described above, recall measures the percentage of PHI that is correctly identified in the detection phase. Thus, the probability of a leak in a set of documents is directly related to recall, given by:

$$Pr(leak) = 1 - Recall$$

4.3 Execution Time

We calculated the run time of the ARTPHIL model using batches of 514 notes (3.2 Mb) on a 4-core windows machine with 16 GB of RAM using the Python Time function, 'time', to estimate the feasibility of running Philter at a large scale. We conducted two experiments. In the first experiment, a single batch with 514 notes and a total size of 3.2Mb was run as a single process and timed. In the second experiment, 20 batches of the 514 notes were run as single processes and timed.

5 Results and Discussion

The ARTPHIL de-identification model is written in Python 3.7 (32 bit) using the Anaconda platform.[1] The design used the modified version of the Philter package [9] and integrated it with the E-ART encryption algorithm [10]. To evaluate the model, we used the i2b2/UTHealth 2014 de-identification corpus that was released by Stubbs et al. in [31] as part of the i2b2 National Centre for Biomedical Computing for the NLP Shared Tasks Challenges, whose de-identification guidelines reported by Stubbs and Uzuner [32] comply with the HIPAA Safe Harbor criteria.

The proposed system was evaluated using the evaluation metrics described in the previous section. We computed the overall recall and precision, as shown in Table 1 and the recall for each PHI category, as shown in Table 2, across the publicly-available 514 notes from the 2014 i2b2 test corpus.

From the six main PHI categories shown in Table 3, name, date, age, contact, and ID received a high recall of above 97%. However, location achieved a low recall of 90.92%, which was primarily affected by the tag location-other. This tag does not belong to the HIPAA PHI categories. Consequently, the i2b2 PHI categories' overall result was worse than that of the HIPAA PHI categories. However, the performance in several PHI types was excellent. For example, the recall of medical record, phone, email, fax and zip code achieved 100%, as shown in Table 2.

Table 1 shows that the de-identification method achieves generally good results, with a recall of 96.93%, precision of 79.76%, F1-score of 87.45% and F2-score of 92.92%. The recall of the method shows how likely it is for a PHI to be missed and, thus, how likely it is for re-identification to occur. Precision measures the number of false positives, thereby estimating the amount of information loss that results from applying the method. These results are consistent with recent research. More importantly, patient name achieved 99.86%, as only two names were missed out of 1,447. For contact, medical record and zip code, which most directly identify PHI, a recall of 100% was achieved, while none were missed.

In term of execution time, the time necessary to run 514 notes as a single process was 2.5147 s. The time required to run 20 batches of 514 notes, 10,240 notes total, was 54.675 s, as shown in Table 4. These results indicate the suitability of ARTPHIL for the processing of large-scale data.

In summary, the ARTPHIL system generates de-identified data that comply with the GDPR requirement for strong pseudonymisation as follows:

[1] https://github.com/bayan6060/phiter_eart.

Table 1. Overall model performance.

Metrics	Result (%)
Precision	79.76
Recall	96.93
F1	87.45
F2	92.92
Risk of re-identification	3.07

Table 2. Recall by tag.

Tags	Recall (%)	TPs	FNs	Risk of re-identification (%)
Medical record	100	721	0	0
Device	100	12	0	0
Username	98.91	91	1	1.09
Email	100	3	0	0
Fax	100	6	0	0
Zip code	100	143	0	0
Street	100	160	0	0
Location-Other	60	12	8	40
Patient	99.86	1445	2	0.41
Doctor	99.24	3272	25	0.76
City	98.54	338	2	1.46
Phone	100	407	0	0
State	96.10	197	8	3.9
Date	100	11880	0	0
Age	100	7	0	0
IDNUM	98.42	374	6	1.58

Table 3. Recall by PHI category.

Category	Recall (%)	TPs	FNs
Name	99.44	4718	27
Contact	99.22	416	0
ID	99.46	1113	6
Date	100	11880	0
Location	90.92	850	18
Age	100	7	0

Table 4. Runtime performance.

Experiments	Time/seconds
1- Single batch with 514 notes	2.5147
2- 20 batch with 514 notes	54.675

- Other techniques used static pseudonyms for the identifiers, which make them vulnerable to the mosaic effect[2]. ARTPHIL uses the dynamic encryption E-ART to de-identify personal data. This dynamism helps reduce the risk of re-identification via linkage attacks as it is difficult to correlate data across available datasets
- Unlike most existing pseudonymisation approaches that only remove or obscure direct identifiers, ARTPHIL de-identifies all direct and indirect identifiers with a recall of 96.93%. This decreases the possibility of retaining any personal data that could re-identify the data subject, thus decreasing the potential violation of data subject privacy.

6 Conclusion

De-identification is an essential tool that organisations can use to reduce the cost and the privacy risks associated with collecting, archiving and transferring personal information. However, current data de-identification approaches are a one-way process, and the original data cannot be recovered from the de-identified data if needed. This paper introduced a new reversible data de-identification system to preserve privacy and utility in unstructured data. ARTPHIL de-identification system combined two previously published works, Philter and E-ART, to de-identify all PHI specified under HIPAA guidelines.

However, the proposed system is highly customisable and can be easily modified to cover any domain. Experiments using 2014 i2b2 test data show that ARTPHIL achieved an overall F2 score of 92.92%, with recall of 96.93%. The average estimate of the re-identification risk was 3.07%. The proposed system ensures its suitability for the protection of individuals' privacy and reduces information loss. It helps to comply with GDPR requirements for the lawful processing of personal data.

The proposed method is useful in the data de-identification domain. Especially, when there is a need to re-identify users in the future. For example, in the research domain when the research subjects should be approached again for further study or to reduce the cost and the privacy risk associated with archiving personal data.

Further, the time required to run 20 batches of 514 notes, 10,240 notes total, was 54.675 s. This efficiency indicates the applicability of the applied ARTPHIL to large datasets or for use in delay-sensitive applications in order to reduce the privacy risk for data subject.

[2] The "Mosaic Effect" occurs when it is possible to determine a data subject's identity without having access to the primary identifiers (e.g., name, ID, Email address, etc.) by correlating data pertaining to the individual across multiple data sets.

References

1. European union, regulation 2016/679. Official J. Eur. Commun. **2014**, 1–88, March 2014
2. H.H.S. office for civil rights, department of health and human standards for privacy of individually. Final rule. Fed. Regist. **67**(157), 53181–53273 (2002)
3. Elliot, M.: Statistical disclosure control (2005)
4. Fung, B.C.M., Wang, K.E., Chen, R.U.I., Yu, P.S.: Privacy-preserving data publishing : a survey of recent developments see ACM for the final official version. **42**(4) (2010)
5. Chen, T.S., Lee, W.B., Chen, J., Kao, Y.H., Hou, P.W.: Reversible privacy preserving data mining: A combination of difference expansion and privacy preserving. J. Supercomput. **66**(2), 907–917 (2013)
6. Hong, T.P., Tseng, L.H., Chien, B.C.: Mining from incomplete quantitative data by fuzzy rough sets. Expert Syst. Appl. **37**(3), 2644–2653 (2010)
7. Silva, J.M., Pinho, E., Monteiro, E., Silva, J.F., Costa, C.: Controlled searching in reversibly de-identified medical imaging archives. J. Biomed. Inform. **77**(July 2017), 81–90 (2018)
8. Landi, W., Rao, R.B.: Secure De-identification and Re-identification. AMIA Annu Symp. Proce. Am. Med. Informatics Assoc. **65**(250), 905 (2003)
9. Norgeot, B., et al.: Protected health information fi lter (Philter): accurately and securely de-identifying free-text clinical notes. npj Digit. Med. 1–8 (2020)
10. Alabdullah, B., Beloff, N., White, M.: E-ART: a new encryption algorithm based on the reflection of binary search tree. Cryptography **5**(1), 4 (2021)
11. Wu, Y., Jiang, M., Lei, J., Xu, H.: Named entity recognition in Chinese clinical text using deep neural network. Stud. Health Technol. Inform. **216**, 624–628 (2015)
12. Allahyari, M., Trippe, E.D., Gutierrez, J.B.: A brief survey of text mining: classification, clustering and extraction techniques. arXiv (2017)
13. Bhasuran, B., Murugesan, G., Abdulkadhar, S., Natarajan, J.: Stacked ensemble combined with fuzzy matching for biomedical named entity recognition of diseases. J. Biomed. Inform. **64**, 1–9 (2016)
14. Keretna, S., Lim, C.P., Creighton, D.: A hybrid model for named entity recognition using unstructured medical text. In: Proceedings of the 9th International Conference System Engineering Socio-Technical Perspect. SoSE 2014, pp. 85–90 (2014)
15. Mishra, A., Jain, S.K.: A survey on question answering systems with classification. J. King Saud Univ. - Comput. Inf. Sci. 28(3), 345–361 (2016)
16. Xu, K., Reddy, S., Feng, Y., Huang, S., Zhao, D.: Question answering on freebase via relation extraction and textual evidence (2016)
17. Dugas, F., Nichols, E.: DeepNNNER : applying BLSTM-CNNs and extended lexicons to named entity recognition in tweets. In: Proceedings of the 2nd Work. Noisy User-generated Text, pp. 178–187 (2016)
18. Derczynski, L., et al.: Analysis of named entity recognition and linking for tweets. Inf. Process. Manage. **51**(2), 32–49 (2015)
19. Gkoulalas-Divanis, A., Loukides, G.: Medical data privacy handbook. Med. Data Priv. Handb. 1–832 (2015)
20. Gupta, D., Saul, M., Gilbertson, J.: Evaluation of a De-Identification (De-Id) software engine to share pathology reports and clinical documents for research. Am. J. Clin. Pathol. **121**(2), 176–186 (2004)
21. Beckwith, B.A., Mahaadevan, R., Balis, U.J., Kuo, F.: Development and evaluation of an open source software tool for de-identification of pathology reports. BMC Med. Inform. Decis. Mak. **6**, 1–10 (2006)
22. Neamatullah, I., et al.: Automated de-identification of free-text medical records. BMC Med. Inform. Decis. Mak. **8**, 1–17 (2008)

23. Steinwart, A.C.I.: Support Vector Machines. Springer Science & Business Media, London (2008).https://doi.org/10.1007/978-1-4471-5571-3_16
24. Lafferty, J., Mccallum, A.: Conditional random fields : probabilistic models for segmenting and labeling sequence data. CIS Pap. **2001**(June), 282–289 (2001)
25. Quinlan, J.R.: Induction of decision trees. Mach. Learn. **1**(1), 81–106 (1986)
26. Gulcher, J.R., Kristj, K.: Protection of privacy by third-party encryption in genetic research in Iceland. Eur. J. Hum. Genet. **8**(10), 739–742 (2000)
27. McMurry, A.J., Fitch, B., Savova, G., Kohane, I.S., Reis, B.Y.: Improved de-identification of physician notes through integrative modeling of both public and private medical text. BMC Med. Inform. Decis. Mak. **13**(1), 112 (2013). https://doi.org/10.1186/1472-6947-13-112
28. Herranz, J., Matwin, S., Nin, J., Torra, V.: Classifying data from protected statistical datasets. Comput. Secur. **29**(8), 875–890 (2010)
29. Yamac, M., Ahishali, M., Passalis, N., Raitoharju, J., Sankur, B., Gabbouj, M.: Reversible privacy preservation using multi-level encryption and compressive sensing. Eur. Signal Process. Conf. **27**, 1.5 (2019)
30. Hernández-Ramos, J.L., et al.: Protecting personal data in IoT platform scenarios through encryption-based selective disclosure. Comput. Commun. **130**(July), 20–37 (2018)
31. Stubbs, A., Kotfila, C., Uzuner, Ö.: Automated systems for the de-identification of longitudinal clinical narratives: overview of 2014 i2b2/UTHealth shared task Track 1. J. Biomed. Inform. **58**, S11–S19 (2015)
32. Stubbs, A., Uzuner, Ö.: Annotating longitudinal clinical narratives for de-identification: the 2014 i2b2/UTHealth corpus. J. Biomed. Inform. **58**, S20–S29 (2015)
33. Scaiano, M., et al.: A unified framework for evaluating the risk of re-identification of text de-identification tools. J. Biomed. Inform. **63**, 174–183 (2016)
34. Ferrández, Ó., South, B.R., Shen, S., Friedlin, F.J., Samore, M.H., Meystre, S.M.: Generalizability and comparison of automatic clinical text de-identification methods and resources. AMIA Annu. Symp. Proc. **2012**, 199–208 (2012)
35. Kushida, C.A., et al.: Strategies for de-identification and anonymisation of electronic health record data for use in multicenter research studies. Medical care 50. Suppl S82 (2012)
36. Hintze, M., LaFever, G.: Meeting upcoming GDPR requirements while maximising the full value of data analytics. SSRN 2927540 (2017)