JENRS

Received: 17 February 2022, Revised: 11 May 2022, Accepted: 23 June 2022, Online: 18 July 2022

DOI: https://dx.doi.org/10.55708/js0107002

Privacy Preserving Text Document Summarization

A N Ramya Shree *, **Kiran P**

CSE Department, RNS Institute of Technology, Bengaluru,560098, India

* Corresponding author: A N Ramya Shree, Asst.Professor, CSE Department, RNSIT,8971896318 & ramyashree.a.n@rnsit.ac.in

ABSTRACT: Data Anonymization provides privacy preservation of the data such that input data containing sensitive information is converted into anonymized data. Hence, nobody can identify the information either directly or indirectly. During the analysis of each text document, the unique attributes reveal the identity of an entity and its private data. The proposed system preserves the sensitive data related to an entity available in text documents by anonymizing the sensitive documents either entirely or partially based on the sensitivity context which is very specific to a domain. The documents are categorized based on sensitivity context as sensitive and not-sensitive documents and further, these documents are subjected to Summarization. The proposed Privacy Preserving Text Document Summarization generates crisp privacy preserved summary of the input text document which consists of the most relevant domain-specific information related to the text document without defying an entity privacy constraints with the compression rate of 11%, the precision of 86.32%, and the recall of 84.28%.

KEYWORDS: PHI, PPDP, Generalization, and Sanitization

1. Introduction

Nowadays, the vast volume of electronic data is increasingly growing. It may be structured data such as databases, leggy data of the organization, or unstructured data such as text contents, images, videos, etc. Approximately 85 to 90 percent of the information is available in unstructured form as per the Forbes Survey. Related to this healthcare providers, state and private enterprises are progressively storing vast numbers and types of medical data in both online and offline modes. In recent years developments in healthcare have resulted in requirements like the tremendous number of personal health data to be collected, exchanged, and analyzed by organizations. There has been an increase in health data being produced and processed by health agencies as a result of the increasing adoption of the Electronic Health Records (EHR), profoundly stimulated by the Health Information Technology for Economic and Clinical Health Act (HITECH Act 2009). Although secondary use of Clinical data has greatly improved the consistency and reliability of medical science and healthcare administration, due to the common nature of exchanging health records which results in increasing queries regarding patient privacy. The Health Insurance Portability and Transparency Act (HIPAA) has developed a series of privacy guidelines to address these queries in HITECH Act 2009. The HIPAA Safe Harbor law defines altogether 18 types of features that are specifically called as confidential features [1].

Personal Health Information (PHI), which must be deleted before a third party is released with the health data which leads to a lot of research in Privacy Preserving Data Publishing on structured data means data which has a pre-defined format, where numerous techniques have been proposed and developed. Privacy protection approaches for sharing medical documents, focus on the detection and removal of PHI items from the documents using different PPDP (Privacy Preserving data Publishing) approaches like Data Swapping, Data Randomization, Cryptography, and Data Anonymization. Among these Data Anonymization is the popularly used Privacy Preserving Data Publishing approach. Data Anonymization is achieved by Pattern Matching based approaches and Machine Learning based approaches which mainly focused on structured data. In this paper, a Machine Learning-based approach called Privacy Preserving Text Document Summarization has been proposed to preserve the privacy of unstructured data which uses i2b2 discharge summary documents, which are collections of progression of the release report of patients by the Harvard University. The discharge summary documents contain the subtleties of a specific patient These informational collections are old certifiable data and are of type text documents [2].

2. Related Work

Whenever a transcript or confidential report is made about an entity it must be protected to preserve the privacy of an entity or an individual before publishing to the outside world. In general, the real-world data associated with an individual or the entity mainly belongs to a specific domain, related to this the data anonymization approach is also mainly dependent on domain-specific attribute types associated with an individual or the entity [3]. There are four major types of domain-specific attributes that are used in anonymization. 1. Personally Identifiable Information (PII) - Attributes that are directly used to identify an individual who belongs to a specific domain. 2. Sensitive Attributes / Private Attributes (SA / PA) - Attributes that are very specific to an individual, which are not to be disclosed.3.Quasi Attributes (QA) - Attributes that are indirectly used to identify or recognize an individual belonging to a certain domain.4. Not-sensitive Attributes (NSA) - Attributes that are considered common for all individuals belonging to a certain domain.

The PII is removed before when an individual or entity data is subjected to summarization because it discloses an individual identity. The Quasi Attributes are those which are used by the attacker or a malicious third-party data analyst to identify the individual or entity when it is linked with other publicly available data like voter lists, census data, etc. The major types of data transformation approaches used in anonymization are Generalization and Suppression. The data usage domain plays a vital role in anonymization because either data generalization or suppressed operations depend on a specific domain. In generalization, the individual quasi attributes are generalized based on the usage domain such that they should not reveal the actual value. Example: Date of Birth attribute value generalized to Born in Year such that the birthdate is not disclosed. In Suppression, the sensitive and or quasi attributes are replaced by special symbols or removed before their usage in data publishing operation. Table 1 describes different anonymization approaches [4].

In the healthcare domain, the most challenging aspect is preserving the privacy of patients who undergo various disease treatment processes. The major research related to privacy preserving data publishing focuses on structured data. which also depends on the usage domain. The major works include survey about the De-identification of Sensitive information in a patient note with recurrent neural networks in a detail and how it further reveals an individual identity[5]. Document Sanitization, which is a privacy policy that aims to identify critical attributes like name, dob, etc. which further can be either removed or replaced before it is made public. The government has set specific guidelines for maintaining confidentiality. According to medical data, the Health Insurance and Portability and Accountability Act (HIPAA) prescribes all personal identification information in medical records must be removed before it can be made available to the public [6]. The center thought of the k-anonymization model is that each record in a table is unclear from in any event from other k-1 records regarding the pre-decided quasi identifier where a table is used which excludes all explicit identifiers. The secure data is fetched by publishing information with different tricky credits which turns out to be almost certain than some other distribution styles [7] . The k-anonymization model which has been widely contemplated and upgraded as a feasible meaning of protection in information distribution. The decision about the k-Anonymization model depends on various strategies like Speculation, Concealment, and other hybrid approaches. It changes private information over to public information such that it can be used at different levels of data handling [8]. The patient outline details which is a fundamental need for clinicians to give facilitated care and practice powerful correspondence. The computerized outline can save time, normalize notes, help dynamically, and lessen clinical mistakes. They specify an upper bound on the extractive outline of release notes and build up an LSTM model to successively name the history of present sickness notes [9].

Table 1: Anonymization Approaches

Techniques	Parameters	Applications	Limitations
k-anonymity Sweeny et. al.	Sensitive features	Correlation between the rows	only on structured data
l-diversity Ashwin M et. al.	Quasi & Sensitive attributes	Equivalent sensitive attributes groups	common frequency value for a sensitive attribute
t-closeness Ninghui Li et. al.	PII, Quasi, Sensitive attributes	Measure the distance between two probabilistic distributions	prone to skewness

3. Proposed System

The proposed system generates privacy preserved text document summary which uses sensitivity context aware anonymization which is a machine learning-based approach focused on unstructured text data. It is mainly described in Figure.1, where the raw text documents are fed as input and are subjected to domain-specific extensive focuses on sensitivity context which is varied from one domain to another. The proposed system architecture is natural language pre-processing processes before the document classification due to the unstructured nature of extractive summarization extracts

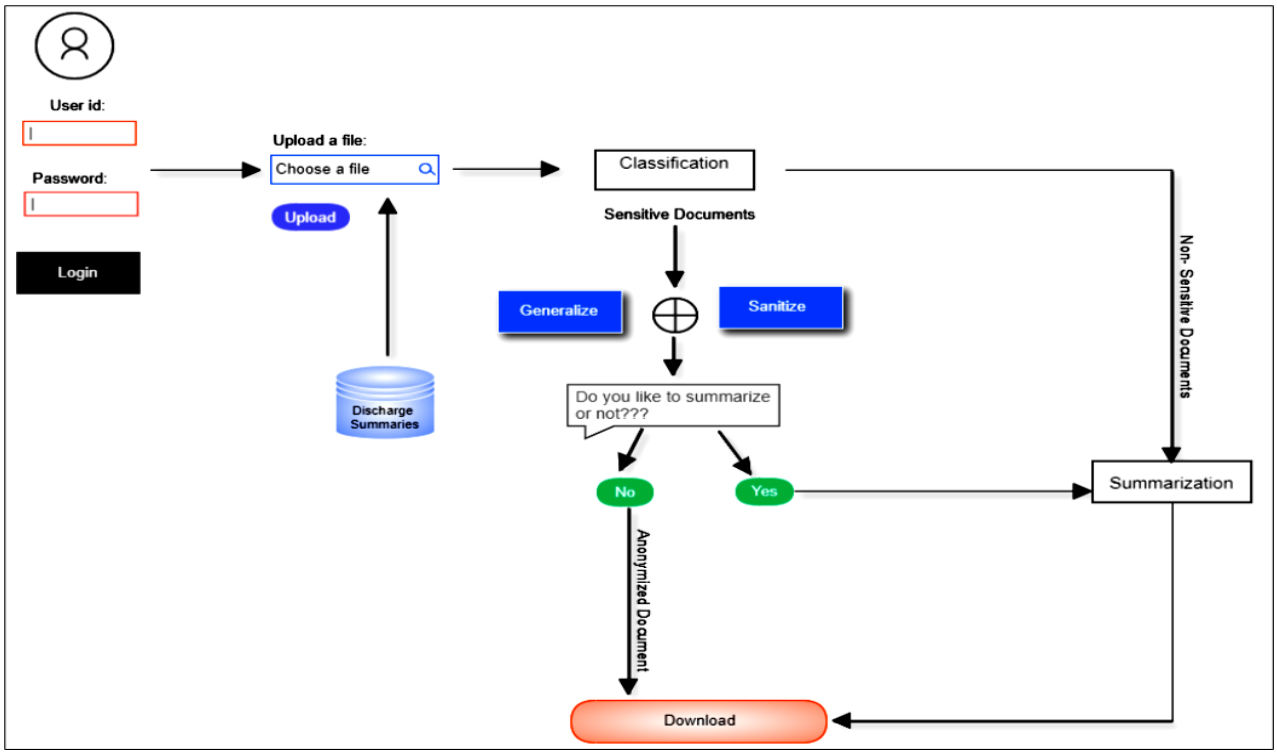

Figure 1: System Architecture of Privacy Preserving Document Summary Generation

a subset of words from a document which treated as most important and specific to the domain to create a single specific summary pertaining to a document. In extractive summarization, weightage is assigned to vital sections of sentences.

Diverse methods and approaches can be used to instrument the sentence weight. Sentence joining is done with relevance and resemblance to domain context to produce a summary. The main features used to generate an extractive summary from patient discharge summary are the past medical history and disease medications on admission related to a particular patient. The discharge summary mainly contains patient individual and disease-specific characteristics as words, sentences, or paragraphs. After preprocessing of discharge summaries, the patient characteristics are available as tokens i.e., words. The sensitive words are those which are identified w.r.t. sensitivity context i.e., determined from the healthcare organization perspective and patients. The sensitivity context required for model development is implemented using a lexicon. It is the knowledge base such that the words in the lexicon are treated as sensitive features which are decided based on the patient's disease details and organizational data regulations. The different phases of the proposed system are as follows:

- Classification-The input text document is classified into a sensitive document or not-sensitive document based on the healthcare domain knowledge and sensitive attributes like disease type present in the given input document.
- Sanitization-This module takes the sensitive document as input and anonymizes the document partially by replacing it with Synthetic data. Generalization- This module takes the sensitive document as input and anonymizes the document completely by replacing it with generic data.
- Summarization- Summarization is the process of highlighting medical information which helps the medical experts efficiently identify the records. The patient's Discharge Summary is given as an input file.

The text document is classified into sensitive or not-sensitive based on the medical terminologies associated with the healthcare domain like sensitive drugs, diseases, etc present in it. The sensitive documents are sent to the anonymization process which is based on the user consent and the anonymization done in two ways they are Generalization and Sanitization. In Generalization, the document is anonymized completely by replacing the quasi attribute with generic data relevant to the healthcare domain. In Sanitization, a document is anonymized partially by replacing the quasi attributes with synthetic data. The anonymized document is subjected to summarization based on the user's choice of whether they are interested to generate a summary or not. If not, then the output document will be either a generalized document or sanitized document. The not-sensitive document is Suppressed, where the Personally Identifiable Information (PII) such as the patient's name, phone number, etc. is suppressed and forwarded to the Summarization process. The input text document is classified into a sensitive document or not-sensitive

document based on the sensitive attributes such as diseases present in the given document [10], [11].

Multinomial Naive Bayes and Logistic Regression Supervised classification techniques are used to predict the target label for the text document as sensitive or not sensitive based on the sensitivity context. Multinomial Naive Bayes is a popular probabilistic classifier based on Bayes Theorem. It uses probability to determine the label of a text grounded on prior knowledge of conditions. It calculates the probability of each tag which is further assigned for a given text and performs label prediction for the tag with the highest probability. Equation 1 describes posterior probability computation to perform the classification of documents based on sensitivity context.

$$P(U|V) = \frac{P(V|U)\,P(V)}{P(U)} \quad (1)$$

where P(U|V) represents the posterior probability of U existence is True with a certain V is True, P(V|U) represents the maximum likelihood of U existence is True certain V is True, P(U) is the prior probability of U existence is true and P(V) is a marginalization of probability V existence is True.

In this approach, to break the sentences as n-grams, the NLTK n-gram tokenizer module is used. The resultant tokens are considered features and the most frequent features are nominated to each predefined class. A feature set is constructed with the union of features that are nominated to predefined classes. The sensitive terms are represented as bag-of-words i.e., a vector $xi = \{xi_1, xi_2, xi_3, \ldots, xi_n\}$, where xi is the number of times the vocabulary term appears in the text document. MNB classifies documents as sensitive or not sensitive based on the posterior probability of the terms with their label occurrence It is determined either by referring to actual sensitive content or consent about data disclosure from an individual patient [12].

Logistic regression is a binary classifier that performs prediction when the target variable is categorical. The Logistic Regression classifier estimates categorical dependent variable relationship with other independent variables and uses binary values of the dependent variable. Logistic Regression refers to predicted values probability scores that are related to the dependent variables i.e., sensitive terms in a range between 0 to 1 & also consider dependent variable natural logs of odds to find refined dependent variable by referring to a logit function. The value nearer to 1 is labeled as a sensitive document otherwise value nearing 0 is labeled as a not sensitive document. In the proposed approach a threshold value of 0.50 is used as a prediction threshold. The odds ratio determines the ratio of success to failure and the same is described in "(2)" where P is the probability of sensitive term occurrences in a document and 1-P is the probability of not occurring of sensitive terms in a document and $0 \leqslant O \leqslant \infty$.

$$O = \frac{P}{1-P} \quad (2)$$

$$Y(1|0) = b + w_i X_i \quad (3)$$

The predicted label Y is categorical and dependent on the independent variables and its co-efficient which is given in "(3)" where $-\infty \leqslant Xi \leqslant \infty$. The logit function is used to predict outcomes as a sensitive document or not sensitive document. The logit function depends on the probability of feature occurrences P and it is $0 < Pi < 1$. Equation (4) and Equation(5) describes the prediction of the test record label as sensitive or not sensitive where, $0 < Pi < 1$ [13].

$$\text{Logit(Pi)} = \ln\left(\frac{Pi}{1-Pi}\right) = f(x) \quad (4)$$

$$Pi = \frac{e^{f(x)}}{1+e^{f(x)}} \quad (5)$$

The sanitization approach takes the sensitive document as input and anonymizes the document partially by replacing it with synthetic data. Synthetic data is the artificial data created by the programmer to preserve the privacy of personal information. Explicit Identifiers are extracted and removed, whereas Quasi Identifiers are extracted and replaced with synthetic data. Feature Extraction is carried out using Regular Expressions which are specific to the healthcare domain [14], [15]. In generalization, both explicit identifiers and quasi-identifiers are extracted. Explicit Identifiers are extracted and removed. Quasi Identifiers are extracted and replaced with generic data. Extraction is carried out using Regular Expressions which are specific to the domain [16] Table 2 describes how the quasi attributes in a sensitive document are anonymized using the proposed approach. In suppression, the explicit identifiers present in the not-sensitive documents are extracted and later suppresses by replacing them with predefined non-readable characters[17], [18]. Summarization is the process of highlighting medical information which helps the medical experts to identify records efficiently. Extractive summaries are created by borrowing phrases or sentences from the original input text [19]. The summarized document in which clinical terms such as dosage, drugs, duration, and frequency of medicine intake are mentioned. For better visual appearance medication strengths are highlighted. The outcome of the proposed approach is described in Figure 2. The text document segmentation breakdown a lengthy document into a shorter one. Shorter segments are sometimes dependent on grammatical rules or dependent on topic continuousness. The term frequency-inverse document frequency-based sentence weightage is used which helps to discriminate and add important domain requirement-specific aspects in the generated summary. The

normalized frequency weights approach is used to discriminate repeated words in a specific document and from a generic corpus which may also contain stop words. To overcome it background information about the healthcare domain is used at the time of stop word removal. The stop word lists are used to eliminate the irrelevant words [20].

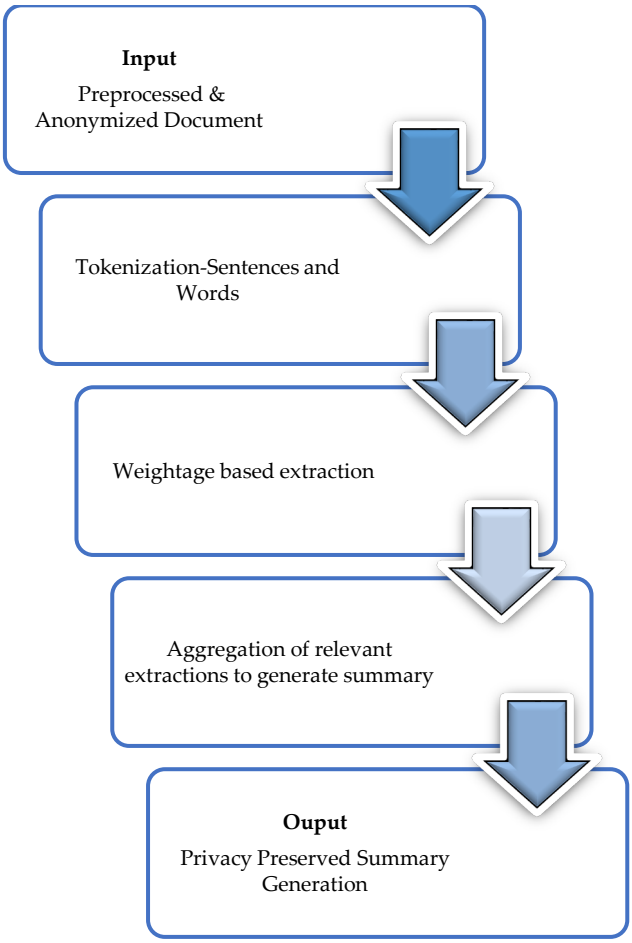

Figure 2: Privacy Preserving Summary Generation Process

The term weightage is approximated using the product of term frequency and inverse document frequency. The inverse document frequency normalization is used to determine the weightage. Let Y = (y1, . . . , yd) be the terms in a document, and idfk is the k^{th} term inverse document frequency. The weight wk can be calculated using "(6)" where the product of terms y_k and idf_k inverse document frequency weight is divided by maximum word frequency outcome is used for document normalization[21], [22].

$$w_k = \frac{y_k \cdot idf_k}{max\{y_i \cdots y_d\}} \quad (6)$$

Terms with weight wk below is a certain threshold is set to weight value 0 and they are treated as not important terms. wk used to score the sentences. The average weight of words related to sentence Sr is used to calculate sentence significance. Average weight μw(Sr) of word w.r.t sentence Sr calculation described in " (7)".

$$\text{Avg. Weight}(Sr) = \frac{\sum_{tk \epsilon Sr} wk}{|\{t:tk \epsilon Sr\}|} \quad (7)$$

Sentences are sorted in descending order w.r.t. to calculated weightage. The sentences with a higher score are selected to produce a summary. Sentence selection for summary generation is dependent on the scoring approach used and the possibility of pairing different selection methods with different scoring methods. Table.2 describes the anonymization of sensitive quasi attributes related to patients in the discharge summary before the summary generation. Sensitive quasi attributes anonymization mainly depends on the domain requirements and consent from the patients regarding the data disclosure [23], [24].

Quasi Identifiers in the discharge summary documents are either hidden or removed based on the type of privacy required by the user. Then for visual representation only the essential text data is extracted and displayed, hence it preserves the privacy of the individual. Algorithm-1 and Algorithm-2 describe procedures associated with the proposed system development.

4. Results and Discussions

The summarization needs to find the important sections of the discharge summary. The summarization assessment can be done using content evaluation. In content evaluation, thoughts of the original document are available in the produced summary which is in turn relevant to human expert generated summary i.e., an ideal summary is analyzed. The Summarization assessment can be done using content evaluation. In content evaluation, thoughts of the original document are available in the produced summary which is in turn relevant to human expert generated summary i.e., an ideal summary is analyzed. Compression Rate (CR), Precision (P), and Recall (R) metrics are used to evaluate the generated summary and the same described in "(8)", "(9)" and "(10)".

Algorithm 1: Privacy Preserving Hospital Discharge Report Generation.

Input: Patient Discharge Summary Document.
Output: Patient Privacy preserved and summarized discharge summary document.

1. Preprocess discharge summary documents.
2. Collect the consent about sensitive data disclosure from the user.

for all preprocessed discharge summary documents **do**

> **Classify** documents as sensitive /not-sensitive based on disclosure consent from the user;
>
> **if** a document is sensitive **then**
>
> > Prompt user for sanitization or generalization.
> > **Generate summary;**

```
    else not-sensitive document then
        Apply suppression.
        Generate summary;
    end
end
```

$$CR = \frac{\text{automated summary length}}{\text{length of the actual document}} \quad (8)$$

$$P = \frac{\text{sentences in system summary+sentences in ideal summary}}{\text{system summary overall sentences}} \quad (9)$$

$$R = \frac{\text{sentences in system summary+sentences in ideal summaries}}{\text{ideal summary overall sentences}} \quad (10)$$

Table 2: Sensitive Quasi attributes Anonymization

Generic Data	Anonymized Data
Name: Mr. John → [NAME] (Entity name)	Name: Mr. John → [NAME] (Entity name)
Age: 43 years → [40-50] years (range)	Age: 43 years → [45.5] years (binning average)
Date: 12-03-2006 → [DATE]	Date: 12-03-2006 →[2006] (year)
Record Identifier 13456	Identifier ##### Suppressed data

Algorithm 2: Privacy Preserving Document Summarization.

Input: Anonymized Document.

Output: Summarized Document.

1. Identify the labels of interest related to the requirement domain.
2. Store domain-specific keywords in a list.
3. Initialize word dictionary WD.
4. Initialize sentence dictionary SD.
5. initialize sentence score dictionary SSD.

for all the anonymized discharge summary documents **do**

```
    Sentence tokenization.
    Word Tokenization;
    if a sentence is not in SD then
        SD[sentence]= WD[word]
    else
        SD[sentence]+= WD[word];
    end
end
```

6.generate histogram with sentence weight

for k exists in WD **do**

WD[k]=WD[k]/maximum(WD.values());

end

7. generate summary w.r.t highest sentence score

The i2b2 data set used for experimentation consists of discharge summaries which have the following details particular to each patient and they are Disease victim or Patient-Attributes

- Reason for admission
- Past medical history
- Medication on admission
- Significant discoveries
- Procedures and treatment offered.
- Patient's discharge state
- Instructions to Patient/family
- Physician's details – one who attends patient during treatments.

Figure 3 shows only the required details like a past medical history of a patient in the generated summary and details of drugs used in the patient medication are highlighted in colors.

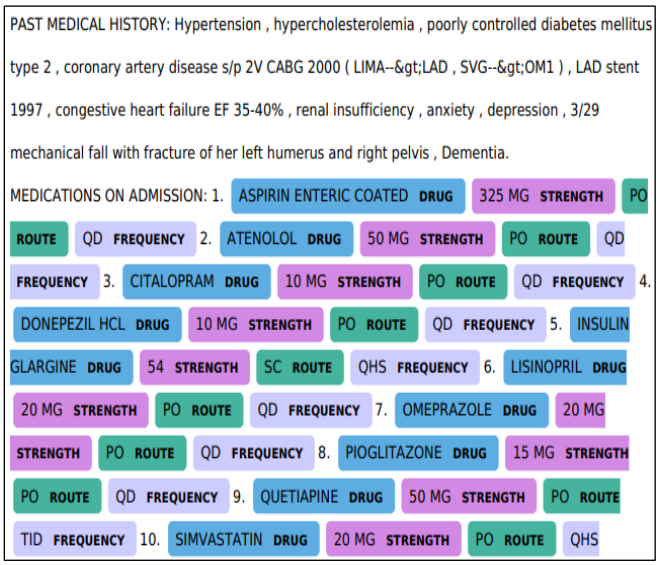

Figure 3: Summarized Medical History and Medicine prescription

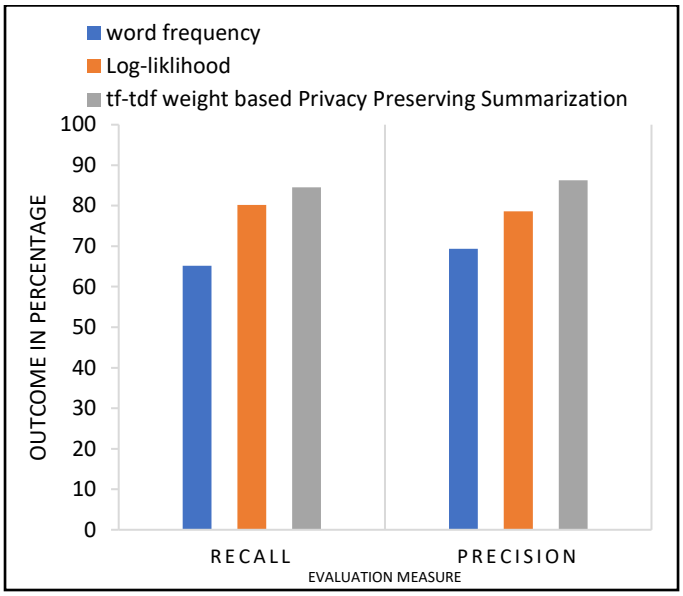

Figure 4: Evaluation of privacy preserving text document Summarization

The proposed approach uses tf-idf-based feature selection to generate the privacy preserved summary. It achieves better results when compared with word frequency where how many times a required word appears in the document sections without referring to sensitivity context. It performs well when compared to log-likelihood topic selection approaches where it requires a sample summary that contains the required terms and relevant terms probability distributions without referring to sensitivity context. The proposed approach compared against word frequency-based and log likelihood-based summarization approaches are described in Figure 4.

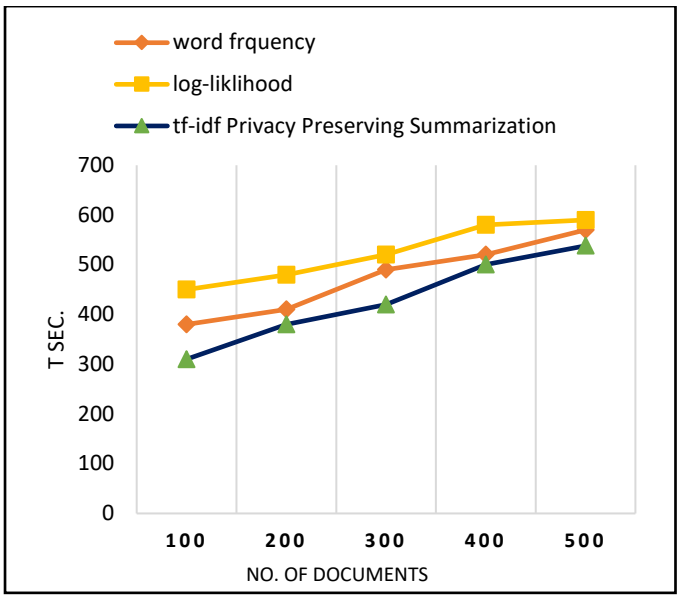

Figure 5: tf-idf based summarization computational analysis.

The proposed approach also generates a summary in a lower computational time since it has already classified the documents based on sensitivity context which is an automated process and not sensitive documents are directly subjected to summarization. When compared to word frequency and log-likelihood topic selection approaches where sensitivity context based selection and classification is not used. The computational evaluation of the proposed approach is described in Figure 5.

5. Conclusion

The proposed approach preserves the privacy of patients whose details are available in the discharge summary as unstructured text data. Extensive domain-specific text pre-processing is required prior to the privacy preserved summary generation. The proposed approach uses a classification technique to initially categorize the discharge summary document based on sensitivity context. The results indicate that the proposed tf-idf-based summarization computationally performed well when compared with other summarization techniques. It also preserves patient privacy without defying privacy constraints.

Acknowledgments

The corpus used in the proposed research work is from the Informatics for Integrating Biology and Bedside Organization, Harward University, USA. Each Discharge summary is unique to a patient, and it consists of patient disease-specific treatment procedures It contains cardiac, diabetes, and tuberculosis patient discharge summaries.

References

[1] K. P. Ramya Shree A N, RNSIT, "Privacy preserving data mining on unstructured data," *International Conference on Science, Technology, Engineering and Management (ICSTEM'17)*, vol. 2, no. 2, 2017.

[2] A. N. R. Shree, P. Kiran, "Sensitivity Context Aware Privacy Preserving Text Document Summarization," *Proceedings of the 4th International Conference on Electronics, Communication and Aerospace Technology, ICECA 2020*, pp. 1517–1523, 2020, doi:10.1109/ICECA49313.2020.9297415.

[3] K. P. Ramya Shree A N, "Privacy Preserving Unstructured Data Publishing (PPUDP) Approach for Big Data," *International Journal of Computer Applications*, vol. 178, no. 28, pp. 4–9, 2019, doi:10.5120/ijca2019919091.

[4] A. N. R. Shree, P. Kiran, "Quasi Attribute Utility Enhancement (QAUE) - A Hybrid Method for PPDP," *International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075*, vol. 9, no. 2S, pp. 330–335, 2019, doi:10.35940/ijitee.B1087.1292S19.

[5] F. Dernoncourt et al., "De-identification of patient notes with recurrent neural networks," *Journal of the American Medical Informatics Association*, vol. 24, no. 3, pp. 596–606, 2017, doi:10.1093/jamia/ocw156.

[6] V. T. Chakaravarthy et al., "Efficient techniques for document sanitization," *International Conference on Information and Knowledge Management, Proceedings*, pp. 843–852, 2008, doi:10.1145/1458082.1458194.

[7] B. Gedik, L. Liu, "A Customizable k-Anonymity Model for Protecting Location Privacy," (Springer, 2004), 620–629, doi:https://doi.org/10.1007/978-981-16-9012-9_49.

[8] K. LeFevre, D. J. DeWitt, R. Ramakrishnan, "Incognito: Efficient full-domain K-anonymity," *Proceedings of the ACM SIGMOD*

International Conference on Management of Data, pp. 49–60, 2005, doi:10.1145/1066157.1066164.

[9] T. Christensen, A. Grimsmo, "Instant availability of patient records, but diminished availability of patient information: A multi-method study of GP's use of electronic patient records," *BMC Medical Informatics and Decision Making*, vol. 8, pp. 1–8, 2008, doi:10.1186/1472-6947-8-12.

[10] R. S. K, "A New Efficient Cloud Model for Data Intensive Application," *Global Journal of Computer Science and Technology*, vol.15,no.1,pp.19–30,2015, doi:https://computerresearch.org/index.php/computer/article/view/1135.

[11] A. N. Ramya Shree, P. Kiran, S. Chhibber, "Sensitivity Context-Aware PrivacyPreserving Sentiment Analysis," *Smart Innovation, Systems and Technologies*, vol. 213 SIST, pp. 407–416, 2021, doi:10.1007/978-981-33-4443-3_39.

[12] A. Majeed, S. O. Hwang, "A Comprehensive Analysis of Privacy Protection Techniques Developed for COVID-19 Pandemic," *IEEE Access*, vol. 9, pp. 164159–164187, 2021, doi:10.1109/ACCESS.2021.3130610.

[13] E. K. Lee, K. Uppal, "CERC: an interactive content extraction, recognition, and construction tool for clinical and biomedical text," *BMC Medical Informatics and Decision Making*, vol. 20, no. Suppl 14, pp. 1–14, 2020, doi:10.1186/s12911-020-01330-8.

[14] M. R. Naqvi et al., "Importance of Big Data in Precision and Personalized Medicine," *HORA 2020 - 2nd International Congress on Human-Computer Interaction, Optimization and Robotic Applications, Proceedings*, pp. 2–7, 2020, doi:10.1109/HORA49412.2020.9152842.

[15] N. K. Anuar, M. Uniten R&D Sdn. Bhd., Kajang, Selangor, ; Asmidar Abu Bakar; Aishah Abu Bakar, "No Title," *2021 IEEE 6th International Conference on Signal and Image Processing (ICSIP)*, no. 6, pp. 1048–1052, 2021, doi:https://doi.org/10.1109/ICSIP52628.2021.9688624.

[16] C. C. Aggarwal, P. S. Yu, "A General Survey of Privacy-Preserving Data Mining Models and Algorithms," pp. 11–52, 2008, doi:10.1007/978-0-387-70992-5_2.

[17] A. N. R. Shree, P. Kiran, "Sensitivity Context Awareness based Privacy Preserving Recommender System," *SSRN Electronic Journal*, no. Icicc, pp. 1–5, 2021, doi:10.2139/ssrn.3835011.

[18] Kiran P A N Ramya Shree, "SCAA—Sensitivity Context Aware Anonymization—An Automated Hybrid PPUDP Technique for Big Data," in *Sustainable Advanced Computing*, ed S.K Aurelia, S., Hiremath, S.S., Subramanian, K., Biswas (Singapore: Springer Nature, 2022), 615–626, doi:https://doi.org/10.1007/978-981-16-9012-9_49.

[19] B. B. Mehta, U. P. Rao, "Privacy Preserving Unstructured Big Data Analytics: Issues and Challenges," *Physics Procedia*, vol. 78, pp. 120–124, 2016, doi:10.1016/j.procs.2016.02.020.

[20] A. El Haddadi et al., "Mining unstructured data for a competitive intelligence system XEW," *SIIE 2015 - 6th International Conference on "Information Systems and Economic Intelligence,"* pp. 146–149, 2015, doi:10.1109/ISEI.2015.7358737.

[21] A. Bafna, J. Wiens, "Automated feature learning: Mining unstructured data for useful abstractions," *Proceedings - IEEE International Conference on Data Mining, ICDM*, vol. 2016-Janua, pp. 703–708, 2016, doi:10.1109/ICDM.2015.115.

[22] P. Jain, M. Gyanchandani, N. Khare, "Big data privacy: a technological perspective and review," *Journal of Big Data*, vol. 3, no. 1, 2016, doi:10.1186/s40537-016-0059-y.

[23] X. Wu et al., "Privacy preserving data mining research: Current status and key issues," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 4489 LNCS, no. PART 3, pp. 762–772, 2007, doi:10.1007/978-3-540-72588-6_125.

[24] C. Zhang et al., "Automatic keyword extraction from documents using conditional random fields," *Journal of Computational Information Systems*, vol. 4, no. 3, pp. 1169–1180, 2008.

Ms. A N Ramya Shree has done her bachelor's degree from KVGCE institution in 2005. She has done her master's degree from SJBIT institution in 2010. The author has a total of 16 + Years of Academic and Research experience. Her research interests include Privacy Preserving Data Publishing and Natural Language Processing. She received the "Predictive Analytics Modeler - Explorer Award 2020" badge by IBM Bengaluru in 2020. She has presented and published a total of 12 research papers in reputed international conferences and journals.

Dr. P Kiran has done his bachelor's degree from AIT institution in 2000. He had done his master's degree from SJCE institution in 2003. He completed his Ph.D. degree in Computer Science from Visvesvaraya Technological University in 2014. The author has a total of 20 + Years of Academic and Research experience. His research interests include Cryptography, Randomization, Anonymization methods in Generalization, Indexing techniques, and Design Patterns. He has presented and published 40+ research papers in reputed international conferences and journals.