Selective Text Encryption Using RSA for E-governance Applications for Pdf Document

Subhajit Adhikari and Sunil Karforma

1 Introduction

The exchange of data or information is now quite frequent in e-governance applications. Textual data, like legal data and the personal data of citizens, flows from different departments in e-governance. If there is any form of leakage during transit, security properties like confidentiality will not be preserved. The confidentiality of sensitive data is to be checked during transmission from the sender to the receiver. To remove threats to confidentiality and other security parameters, the technique of encryption is widely used. Traditional encryption systems can be divided into two subcategories: symmetric and asymmetric methods [1]. But in recent studies, there have been various proofs available to disqualify the applicability of the symmetric key concept in terms of textual information encoding. So, as a consequence, the asymmetric key concept is a good choice for encryption of textual data. With a different view point, it can also be stated that the encoding methods can be of two types: encoding with a selective portion and encoding with the whole portion of the original text. Both the two methods have its benefits and drawbacks. Full encryption methods are not suitable for resource constrained environment [2]. Considering the method of whole text encoding, it is obvious that it must consume the more

S. Adhikari (✉)
Assistant Professor, BSH Department, Institute of Engineering and Management, University of Engineering and Management, Kolkata, India
e-mail: Subhajit.adhikari@iem.edu.in

Research Scholar, Department of Computer Science, University of Burdwan, Burdwan, India

S. Karforma
Dean(Science) Faculty, Department of Computer Science, The University of Burdwan, Burdwan, India
e-mail: skarforma@cs.buruniv.ac.in

J. K. Mandal et al. (eds.), *Proceedings of International Conference on Network Security and Blockchain Technology*, Lecture Notes in Networks and Systems 738,
https://doi.org/10.1007/978-981-99-4433-0_22

computation time than selective encoding, but the speedup factor is also a major factor [3]. In selective encoding, the speed of encryption is much higher for huge amounts of data produced from different sources maintaining same level of security of whole text encryption method. In our proposed method, we consider the benefits of both the asymmetric key method and the selective encoding approach to design a robust and secure encryption system. So, regular expressions are used to select the segment of textual data, given a text as user input, and then RSA cryptography is implemented to encrypt the selected segment of text. In our research study, 1024 bit RSA is used for strongest encryption process. The cryptosystem RSA is very famous for its class of algorithms in asymmetric key cryptography [4]. The steps of RSA algorithm has already defined in [5]. In our research study, the predefined function $rsa.encrypt(Org\_msg, Pub\_key)$ of 1024 bits in Python-RSA module [6] as pure Python-RSA implementation for encryption is taken for the experiment. In decoding step, $rsa.decrypt(Enc\_msg, Priv\_key)$ is used to decode the orginal text, where Org_msg depicts original message, Pub_Key depicts public key of the receiver and $Priv\_Key$ depicts the private key of the receiver. The message is encoded and decoded with the 'utf8' format before encrypting process and after decrypting process respectively.

2 Our Contribution

Selective encryption in the context of text encryption is very rare. Our main contribution is that some portion of the data must be untraceable, even if the attacker manages to extract the rest of the unencrypted data. Assume the PAN or the Aadhaar number is important information of citizen that must be kept private. Whenever an Income Tax Return form is generated by the authority, the PAN number is added to it. If the attacker can obtain the PAN number, he or she can obtain all the legal information pertaining to a particular citizen. Our aim is to encode only the PAN, while the rest of the document will not be encoded. So, RSA with a 1024-bit encoding technique is implemented. We combine the benefits of selective encryption and an asymmetric key algorithm to design our new encoding technique. We chose the selective encoding method by search to reduce the time required by traditional whole text encryption. The asymmetric key encoding scheme is then used to achieve the highest level of security while maintaining the data's confidentiality. Our method can be extended and applied to secure medical records and sensitive data generated by wireless and IOT devices.

3 Literature Review

The purpose of the research study [1] is to introduce a novel selective significant data encryption algorithm, where a significant amount of uncertainty is added to data as it is encrypted. This algorithm takes help of the concept of natural language processing and extracts the data from the whole text. There are four steps to the selective encryption technique studied in this study. First step is to removing special characters, secondly tokenization fetches all words available in the message l, after that the words signifies termination have been removed. Lastly, encryption process is applied to the keywords to leaving the common words as it is. Both encrypted keywords and plain common words are sent to the network. In recent times, a research [2] is carried out considering selective encryption for image and audio data in resource constrained environment in terms of low memory, low computation capacity and low power requirements. Also, selective encoding technique is evaluated in association with metrics like tenability, degradation of visual effect, cryptographic security, encryption ratio, compression friendliness, format compliance and error tolerance. The categorization of selective encoding is also done based on pre-compression, in-compression and post-compression approaches. The selective significant data encryption [3] approach for text data encryption was introduced in the previous study. This method chooses just relevant data from the entire message in terms of the whole message's keywords, which gives the data encryption procedure enough uncertainty. This improves speed and cuts down on the overhead associated with encryption. The symmetric key encoding technique is used to carry out the encryption process. The Blowfish algorithm is employed for this. A comparative study of the proposed technique, the full encoding scheme, and the toss of a coin method is also included in terms of proportion of encoded text and computation time. In this study of a selective encoding scheme[7], they provide an innovative AES-Rijndael-based encryption technique for medical data. Firstly, a selector component is depicted that allows the method to be implemented on a variety of platforms, with the required size of input, count of rounds. In the second phase, the compression process of original picture is done with the Huffman algorithm to decrease the size of the picture and encryption time of AES method by more than half. And thirdly, the simulation time of AES algorithm is kept minimum with the concept of loop unrolling and methods of merging in proposed algorithm. Experimental study proves that this novel selective encoding scheme cut down the average execution time by 35% comparing to traditional AES scheme. Previously, a modified RSA [8] method has been presented with improved security for message encryption. By identifying three factors of n instead of two, makes the proposed encrypting model more difficult for an attacker to guess by the process of factorization. Thus the security is raised by two levels. Finding a public key and a private key as a result of the second modulus x being used in place of the modulus n being passed is challenging since only using these keys makes it feasible to encrypt or decode messages while maintaining message secrecy. The time to produce the keys of the encoding system is less than the traditional RSA cryptographic method. In this article, a new selective encryption technique[9]

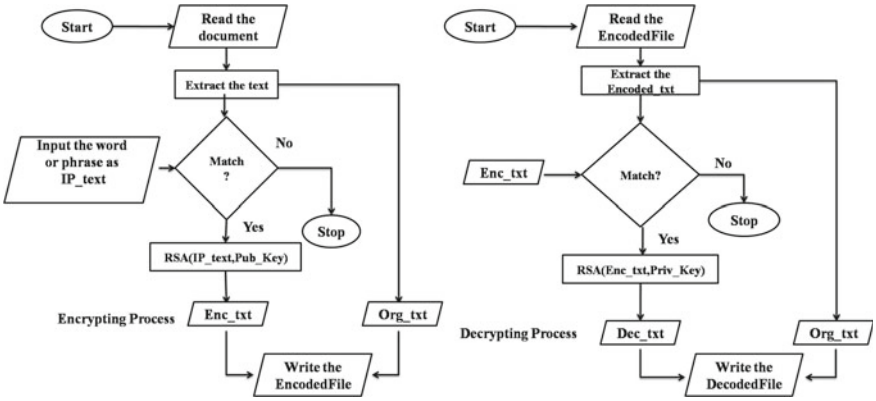

Fig. 1 Block diagram of encryption and decryption process

is demonstrated that employs a safe, index-based chaotic sequence to encrypt only the chosen compressed video frames from each set of images. Simulation results and statistical analysis have done based on quality analysis, keyspace metric, psnr analysis, mean-square-error analysis and computation time analysis and it is found effective and efficient rather than traditional AES and RC5 encoding algorithms. The concept of the CMYK color model [10] has already been used to create a unique encoding and decoding approach with four keys for conversion from text to image. This approach encrypts data faster in terms of text characters. In order to prevent the mathematical factorization of n from leading to the factors p and q, the modified RSA algorithm [11] incorporates the removal of the large prime number n from the key. A one-digit number serves as the initial message in this experiment. According to the analytical report, the suggested approach encrypts and decrypts faster than a conventional RSA strategy. To address the issue of slow key decryption or slow key transmission, an improved method of homomorphic encryption based on Chinese remainder theorem with a Rivest-Shamir-Adleman [12] method was developed, utilizing multiple keys. It performs the cipher text decoding better than standard RSA for documents.

4 Proposed Algorithm

The proposed algorithm is depicted in a block diagram in the Fig. 1.

4.1 Encrypting and Decrypting Procedure

The process of encrypting and decrypting schemes are given below.

Algorithm 1: Encryption Procedure

Input: OriginalPDF, Input text as IP_txt
Output: PDF as encodedFile
1.Read the text lines from the document in Org_txt.
2.Take input the word or phrase to be searched and saved into IP_txt.
3.Loop
4. If IP_txt== Org_txt then
5. Compute rsa.encrypt(IP_txt, R_pubKey) and save it to Enc_txt.
6. Add a special symbol "??" To the end of Enc_txt and save it to $FinEnc\_txt$.
7. Write $FinEnc\_txt$ as string to a encoded file as "encodedFile".
8. Else
9. Org_txt as string to a encoded file "encodedFile".
10. EndIf
11.Untill End of File.
12.Stop.

Algorithm 2: Decryption Procedure

Input: PDF as encodedFile
Output: OrginalPDF as DecodedFile
1. Read the text from the encodedFile.
2. Separate the encoded string in "EncSting" using special symbol "??" from the original text Org_txt.
3. Compute rsa.decrypt($EncSting$, $R\_privKey$) and save it to Dec_txt.
4. Write Dec_txt to the file DecodedFile .
5. Write Org_txt to the file DecodedFile.
6. Stop.

5 Implementation Example

The experiment has been conducted in Intel 3rd gen processor computer having 1.70 GHz cpu speed, 500GB HDD and RAM of 4GB capacity. The software Pycharm of version 2020.2 is utilized for the experiment along with Matlab R2016b for statistical analysis. Different standard pdf documents are collected from the web sources [13–15]. In the following example, the content of the pdf document is considerd for analysis irrespective of the position and layout and font of the pdf document. The content "July 4, 1776" is selected from second line of text the for encrypting and decrypting process. The process of selective encoding mechanism is applied to the selected part "July 4, 1776" and the encrypted form of the text is written to the encoded pdf file. The content of encoded pdf file is shown in Fig. 2 in the middle .The decrypting process converts the encoded content back to the original text "July 4, 1776" and written to a new decoded pdf file. The content of decoded pdf file is shown in Fig. 2 in bottom part.

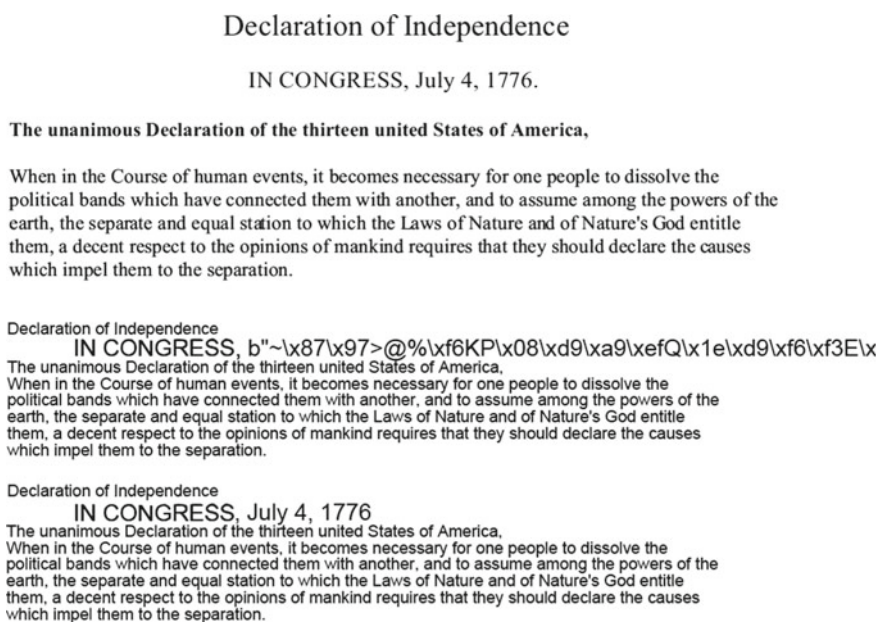

Fig. 2 Original text, encrypted text and decrypted text

6 Analysis of Security Parameters

The dataset is composed of three standard pdf documents. The extracted portion of the text is named "Data1","Data2" and "Data3", respectively. As for example the "Data1" consists of the text "July 4, 1776". As for example the "Data2" consists of the text "SEMPRONIO". As for example the "Data3" consists of the text "Contents".

6.1 Study of Key Space

Study of keyspace considers the number of changing variables used for computation. The high value of this metric discards any type of attacks that are bruteforce in nature. The standardization made with IEEE floating-point value consideration, is that the accuracy of double variables is approximately 10^{-15} with the bit capacity 64. We have four double variables as p,q,e and d. So, the keyspace value is about $10^{60} \approx 2^{199.31569}$. So, our scheme of encrypting and decrypting text is constituted to give protection about all attakcs made in bruteforce approach considering this large keyspace.

Table 1 Study of entropy

Content	Original	Encrypted
Data1	3.251629	4.3741
Data2	2.947702	3.16992
Data3	2.5	4.54205

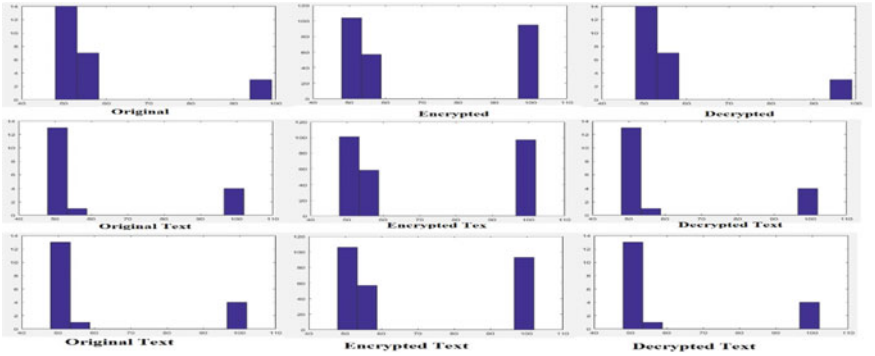

Fig. 3 Study of histogram of Data1, Data2 and Data3

6.2 Entropy

The term is first uttered by the famous mathematician Shannon as a metric to measure uncertainty. It has been applied in the domain of information processing [16]. The value of a text with a lower probability of the occurrence of an event retains more information, and thus it has a higher information entropy [17]. As a consequent, suppose "Data security" has less probability of appearance than the sentence "Data security is applicable to different fields". The metric entropy of a sentence represents indicates how much information it contains [18]. The study of entropy can be depicted as the Eq. 1 given below [19]

$$H(P) = \sum_{i=0}^{255} [\text{Prob}(X_i) \times \log(\frac{1}{\text{Prob}(X_i)})] \tag{1}$$

In the above equation $Prob(X_i)$ represents the probability of existence of symbol X_i

From the above Table 1, the encrypted text has more entropy value than original text. The higher value of entropy makes the encrypting and decrypting scheme very hard to crack.

Table 2 Study of avalanche effect

Content	Avalanche_Effect
Data1	0.51956947
Data2	0.5112414467
Data3	0.5341796875

6.3 Histogram Analysis

Each letter or symbol that appears in the message “Msg" is shown by a histogram. If the spread of the letters or symbols is uniform, the encrypting technique is also insurmountable in the face of statistical assaults [20]. The histogram plot of the ciphered text should differ drastically from the histogram of the plain text and should be as evenly distributed as is humanly feasible, meaning that the likelihood of any value existing is the same [10]. In the above Fig. 3, the histogram of original, encoded and decoded text is depicted taking conversion to ASCII format. For the encrypted text, the histogram representation is uniform in terms of vertical bars than the histogram of original text.

6.4 Avalanche Effect

A feature of an encryption method known as the ”avalanche effect” causes changes in multiple bits of the encoded text when one bit of the original text is changed [21]. Avalanche impact should be 0.5 under ideal circumstances [22]. The Eq. 2 of avalanche effect is depicted below. In the equation ”CTEXT” represents cipher text.

$$\text{Avalanche Effect} = \frac{\text{Number of Bits Flipped in Ctext}}{\text{Number of Bits in Ctext}} \tag{2}$$

From the above Table 2, the conclusion can be made easily that our proposed technique crossed the ideal range of the avalanche effect value, depcting a good encrypting system property.

6.5 Plaintext Sensitivity

The study of plaintext sensitivity depicts that a small moderation in the original content in terms of a bit can create a rapid change in the encoded content. The original text is ”July 4, 1776” is changed to ”July 4, 1777” to compute the plaintext sensitivity and the result is given in the above Fig. 4. As a consequent, the above

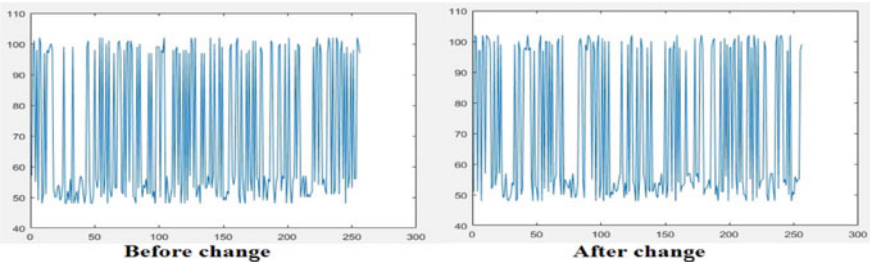

Fig. 4 Study of plaintext sensitivity

Table 3 Study of encryption and decryption time

Content	Enc_Time(Sec)	Dec_Time(Sec)
Data1	0.0004920	0.0133413
Data2	0.000353	0.027816
Data3	0.00031	0.009625

Table 4 Comparison result of proposed text encoding with others

	Entropy	Avalanche	Enc_Time(Sec)
Proposed method	4.3741	0.51956947	0.0133413
[3]	NA	NA	1500
[8]	NA	NA	0.0050
[10]	6.92	NA	0.000001
[11]	NA	NA	21980
[12]	NA	NA	0.412

two encoded images are totally different before and after the encoding process. So, only one-digit change in the original string make a huge change in cipher text.The correlation between two cipher files is -0.0276. This low value of correlation means there is no relation between two encoded files.

6.6 Computation Time Analysis

In the Table 3, the computation time for encoding and decoding text file is given in seconds. The time analysis satisfies that our method consumes less cpu time and can be incorporated not only in e-governance application but also in resource limited environment.

From the above Table 4, it is very clear that existing methods of text encryption lack in detailed statistical anlysis in terms of metrics like entropy and avalanche effects

and only present required encryption time. Our method has high value of entropy, ideal value of avalanche effect with low encryption time. Also, our propsed method of encoding text constists of detailed study of statistical metrics which proves the robustness against different attaks. The important metrics like plaintext sensitivity and histogram study have also been included in our research study to qualify as a good cryptosystem.

7 Conclusion and Future Scope

Our research study provides the text data security in e-governance applications. The asymmetric approach of encoding text is discussed in this paper using 1024 bit RSA cryptographic algorithm. The confidentiality property of data is guaranteed by our proposed method along with high security features. Government documents and Legal documents can be secured using our proposed encoding scheme. Important selected data like account number, PAN and Aadhaar of any citizen can be encrypted using proposed method and added in the government documents. Attacker may find the document but unable to decrypt the selected part of the content which leads to an unsuccessful attempt of data theft. The security analysis report proves the robustness of our method against different attacks causing security threats. Also, the proposed model of encrypting and decrypting specific part of the content fetched from pdf document takes less time than whole text encoding. As a consequence, the applicability of our encrypting method rises for resource limited environment. As of now, the method is implemented for text in pdf document but can also be applied for multimedia content like image and video. In future, chaotic functions may be incorporated to introduce more randomness in the encoding and decoding technique. The encoding scheme can also be extended with the elliptic curve cryptography. The proposed method of encryption can be done with any length and in any position, but in the context of "Selective Encryption", a small portion of the whole text is taken for experiment.

References

1. Kushwaha A, Sharma HR, Ambhaikar A (2018) Selective encryption using natural language processing for text data in mobile ad hoc network. In: Modeling, simulation, and optimization. Springer, Cham, pp 15–26
2. Massoudi A, Lefebvre F, De Vleeschouwer C, Macq B, Quisquater JJ (2008) Overview on selective encryption of image and video: challenges and perspectives. Eurasip J Inf Secur 2008(1):179290
3. Kushwaha A, Sharma HR, Ambhaikar A (2016) A novel selective encryption method for securing text over mobile ad hoc network. Procedia Comput Sci 79:16–23
4. Kota CM, Aissi C (2022) Implementation of the RSA algorithm and its cryptanalysis. In: 2002 GSW

5. Shawkat SA (2007) Enhancing steganography techniques in digital images. Faculty of Computers and Information, Mansoura University Egypt-2016
6. https://stuvel.eu/python-rsa-doc/usage.html , Accessed 06 Dec 2022
7. Oh JY, Yang DI, Chon KH (2010) A selective encryption algorithm based on AES for medical information. Healthc Inf Res 16(1):22–29
8. Jaju SA, Chowhan SS (2015) A modified RSA algorithm to enhance security for digital signature. In: 2015 international conference and workshop on computing and communication (IEMCON). IEEE, pp 1–5
9. Batham S, Yadav VK, Mallik AK (2014) ICSECV: an efficient approach of video encryption. In: 2014 seventh international conference on contemporary computing (IC3). IEEE, pp 425–430
10. Noor NS, Hammood DA, Al-Naji A, Chahl J (2022) A fast text-to-image encryption-decryption algorithm for secure network communication. Computers 11(3):39
11. Minni R, Sultania K, Mishra S, Vincent DR (2013) An algorithm to enhance security in RSA. In: 2013 fourth international conference on computing, communications and networking technologies (ICCCNT). IEEE, pp 1–4
12. Abid R, Iwendi C, Javed AR, Rizwan M, Jalil Z, Anajemba JH, Biamba C (2021) An optimised homomorphic CRT-RSA algorithm for secure and efficient communication. Pers Ubiquitous Comput 1–14
13. https://www.kaggle.com/code/gauravduttakiit/working-with-pdf-files/data , Accessed 06 Dec 2022
14. https://www.kaggle.com/datasets/paretogp/examples-exams-pdf , Accessed 06 Dec 2022
15. https://www.bl.uk/collection-metadata/downloads , Accessed 06 Dec 2022
16. Xu W, Pan Y, Chen X, Ding W, Qian Y (2022) A novel dynamic fusion approach using information entropy for interval-valued ordered datasets. IEEE Trans Big Data
17. Xu H, Lv Y (2022) Mining and application of tourism online review text based on natural language processing and text classification technology. Wireless Commun Mob Comput
18. Khurana A, Bhatnagar V (2022) Investigating entropy for extractive document summarization. Expert Syst Appl 187:115820
19. Lin H, Wang C, Cui L, Sun Y, Zhang X, Yao W (2022) Hyperchaotic memristive ring neural network and application in medical image encryption. Nonlinear Dyn 110(1):841–855
20. Hagras T, Salama D, Youness H (2022) Anti-attacks encryption algorithm based on DNA computing and data encryption standard. Alexandria Eng J 61(12):11651–11662
21. Gamido HV, Sison AM, Medina RP (2018) Modified AES for text and image encryption. Indonesian J Electr Eng Comput Sci 11(3):942–948
22. Ghadirli HM, Nodehi A, Enayatifar R (2019) An overview of encryption algorithms in color images. Sig Process 164:163–185