Aggrandize Text Security and Hiding Data through Text Steganography

Sunita Chaudhary
Department of Computer Science and IT
Jagannath University ,Jaipur
er.sunita03@gmail.com

Meenu Dave
Department of Computer Science and IT
Jagan Nath University, Jaipur
meenu.s.dave@gmail.com

Amit Sanghi
Dept. of Computer Science and Engg.
Marudhar Engineering College, Bikaner
dr.amitsanghi@gmail.com

***Abstract*—Steganography is the technique of cover up information in another media and protects it from prying eyes. Modern Steganography intends to hide the data in a cover media such as text, digital images, video, audio, to exchange secret message. Communication parties rely on the change in the structure and features of the cover media in such a manner as is not identifiable by prying eyes. However, using the text as the cover medium is relatively difficult as compared to the other cover media. This difficulty is observed because of the lack of redundant information in a text file, as compared to an image, video or a sound clip which contains much redundancy that is exploited by the steganography algorithms. In this paper, we present and evaluate our contribution to design two new approaches for text Steganography and named them as CASE (Capital Alphabet Shape Encoding) and ISET (Indian Script Encoding Technique). These methods are combination of the random character sequence and feature coding method. Here we take two processes in these text steganography approaches. Firstly encode all the characters of the secret message with a new encoding technique base on the classification of the Hindi characters or English characters. Second hide the message in the randomly generated cover text. CASE and ISET reduces the memory consumption and size of cover text used for steganography. In these methods, one letter can hide maximum of eight bits which improves time overhead and memory overhead.**

Keywords—Text; Steganography; CASE; ISET; Random; Cover Text.

I. INTRODUCTION

Steganography is the practice of hiding a confidential message in another non-secret message such that it conceals communication [1]. The word Steganography comes from Greek terminology and that stands for "covered writing"[2]. The goal of Steganography is to hide a message inside other media in a way that does not allow any third party to detect even the existence of the communication taken place. It uses a cover message such as text, image, audio, video file format, etc. to hide a secret message [3]. The secret message embeds in the cover message by applying some mathematic logic. Text Steganography is a process to cover up the secret information within text (i.e. in character based) messages [4].

978-1-4673-8962-4/16/$31.00 ©2016 IEEE

Steganography includes a various technique for hiding a message in a variety of media. Digital Steganography uses a cover media such as text, images, audio and more[5]. The advantage of using steganography is that it conceals the discovery of communication taken place between intended parties. Mathematically a Stenography process is:-

$$Secret\_Object + Cover\_Medium\_Object + Steno\_Key = Steno\_Media\_Object$$

Where Secret_Object is a secret object such as confidential text message, image and copyright music file. Cover_Medium Object is an object of cover media such as any text file, audio clip, digital image, etc. Steno_Key is logic behind deceit process. Steno_Media_Object is a resultant Camouflage objects[6].

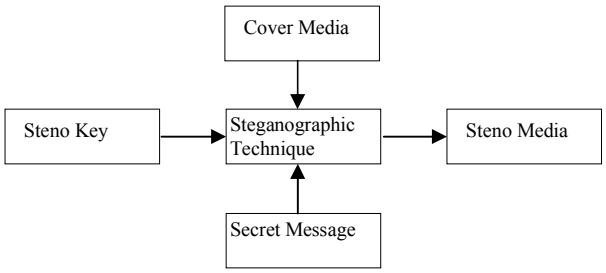

Fig. 1. A Typical Steganography Process

II. GENERIC CLASSIFICATION OF TEXT STEGANOGRAPHY

Three basic categories are provided in Text Steganography: Format-based methods, Random and statistical generation methods, and Linguistic methods[7]. Brief description of all these three categories is given here [8].

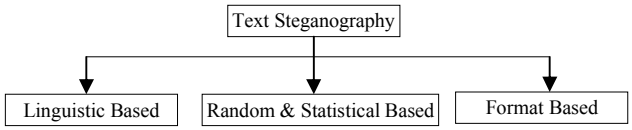

Fig. 2. Generic classification of text steganography

III. PROPOSED APPROACH

A. Capital Alphabet Shape Encoding(CASE)

The initial step in the approach of text steganography is to encode the secrete message according to the shape of English alphabets, as we are using only capital alphabet. In the next step we hide this message with the random cover text by mixing it with the contents of cover text, i.e. first we encode the secret character into a 8-bit binary string which is according to the grouping given in table I and II and then we perform process of finding the ASCII equivalent of the 8-bit format of the secret character and hide it with the contents of random cover text [15]. We will get 8-bit format according to the table I.

TABLE I. BIT FORMAT OF CASE

0^{th}bit	1^{st}bit	2^{nd}bit	3^{rd}bit	4^{th} bit	5^{th}bit	6^{th}bit	7^{th}bit
Alphabet(0)	Group no.			Case	Group position in alphabet Group.		

1) Grouping for CASE: As shown in Table II, we divide the English letters into 8 groups viz; G1, G2, G3, G4, G5, G6, G7 & G8. G1 group contain Letters with no curve, horizontal & vertical line like V, W, X, Y and group bits CHV will be '000' here, CHV stands for curve, horizontal and vertical lines.G2 group contain Letters with vertical straight line like K, M and N and group bits CHV will be '001'.G3 group contain Letters with horizontal straight line like A and Z and group bits CHV will be '010'. G4 group contain Letters with only horizontal & vertical line like E, F, H, I, L and T and group bits CHV will be '011' .G5 group contain Curved Letters like C, O, Q, S and U and group bits CHV will be '100. G6 group contain Letters with curve & vertical line like B, D, P and R and group bits CHV will be '101'.G7 group contain Letters with curve & horizontal line like G and group bits CHV will be '110'.G8 group contain Letters with curve, horizontal & vertical line like J and group bits CHV will be '111'.

To hide it or embed it with the cover text, we made a new technique. In this technique we encode the first three letter of the cover text by using CASE approach and then count the bits having value 1. This count value is the key value for hiding the data. After calculating the key value message is mixed up, one character of message comes after key number character of cover text and this process of embedding are repeated until whole message is hidden in the cover text.

For example, we have generated secret message "is" and take the random character sequence of cover text like "This chapter describes Layer 2". From this we will encode first three letters "Thi" and by getting their 8-bit sequence format we count the number of "1" bit in the byte sequence. Then the key value or number of one will be "13".Now,every 13th letter will be replaced by the encoded secret message character, say if it is "3" and "F" then it will be hidden as "This chapter3 describes LF". Here we are appending letters on 13th bit position to hide the first letter and next 13th bit position for next character. In this method, we are grouping English letters based on their features.

TABLE II. GROUPING IN CASE

Group Name	Bits (CHV)	Letters in Group					
Letters with no curve, horizontal & vertical line (G1)	000	000	001	010	011		
		V	W	X	Y		
Letters with vertical straight line (G2)	001	000	001	010			
		K	M	N			
Letters with horizontal straight line (G3)	010	000	001				
		A	Z				
Letters with only horizontal & vertical line (G4)	011	000	001	010	011	100	101
		E	F	H	I	L	T
Curved Letter (G5)	100	000	001	010	011	100	
		C	O	Q	S	U	
Letters with curve and vertical line (G6)	101	000	001	010	011		
		B	D	P	R		
Letters with curve and horizontal line (G7)	110	000					
		G					
Letters with Curve, Vertical and Horizontal Line (G8)	111	000					
		J					

B. Indian Script Encoding Technique (ISET)

The proposed method uses an encoding scheme to hide the confidential information of hindi language [11] in randomly generated cover text. The acronym ISET stands for Indian Script Encoding Technique. In the ISET, every letter is chosen for hiding firstly convert into a binary string which is 8-bit long according to ISET tables. This 8-bit long string is now converted into the equivalent ASCII character. For Hindi Alphabets, the First two bits of 8-bit encoding ISET remain 0.The bit at position III, IV and V represent the alphabet group. The bit VI, VII and VIII are used to mark the position of letter in the group. The 8-bit ISET format for Alphabet is shown in Table III.

TABLE III. 8- BIT FORMAT OF ISET

I bit	II bit	III bit	IV bit	V bit	VI bit	VII bit	VIII bit
Always '0'	Always '0'	ISET Alphabet Group			Group position in alphabet		

1) ISET grouping: We have made eight group of Hindi letter [10] based on how they pronounced (based on articulation of letters). The bit at position II, III, IV represents the group number. All alphabets are divided into group and every alphabet has its position in corresponding group. The position of character is represents using last three bit at V, VI and VII bit position of 8-bit binary string.

For Example, To encode the letter द then using ISET approach it falls in the fourth group at the third position, it can be coded as a 00011010 and its ASCII equivalent is ':' which is in decimal equivalent to 58. After encoding, now letter ':' is scrambled with the cover text. And letter हिडेस all the 8 bits of original letter द in the cover text. The Group structure for ISET alphabet illustrates in Table IV.

TABLE IV. GROUPING IN ISET

Group Name	Bits	Letters in group						
G1: Velar (कण्ठ)	000	000	001	010	011	100	101	110
		अ	आ	क	ख	ग	घ	ङ
G2: Palatal (तालु)	001	000	001	010	011	100	101	110	111
		इ	ई	च	छ	ज	झ	य	श
G3: Retroflex (मूर्धा)	010	000	001	010	011	100	101	110
		ऋ	ट	ठ	ड	ढ	र	ष
G4: Dental (दंत)	011	000	001	010	011	100	101
		त	थ	द	घ	ल	स
G5: Labial (होठ)	100	000	001	010	011	100	101	110
		उ	ऊ	प	फ	ब	भ	व
G6: Nasal (नाक)	101	000	001	010	100	101
		ङ	ञ	ण	न	म
G7: Conjucts (संयुक्ताक्षर)	110	000	001	010	011	100	101	110
		ए	ऐ	ओ	औ	क्ष	त्र	ज्ञ
G8: Ayogwah (अयोग्वाह)	111	000	001	010
		ॅ	-	:

a) Velar group G1: In the first group, we include those letters which are pronounced with the velar. Candidate for this group are :

अ	आ	क	ख	ग	घ	ङ

The letter chosen from this group can hide '001' bits.

b) Palatal group G2: For the group second, those letters are including which are pronounced using palatal. Letters chosen form this group can hide '001' bits. The members of this group are:

इ	ई	च	छ	ज	झ	य	श

c) Retroflex group G3: For the group third, those letters which are spoken with the help of retroflex. Letters from this group can hide '010' bits. Candidates for this group are:

ऋ	ट	ठ	ड	ढ	र	ष

d) Dental group G4: In the fourth group, those letters are included which are spoken with the help of teeth. Any letter from this group is used to hide '011' bits. Candidates for this group are :

त	थ	द	घ	ल	स

e) Labial group G5: Fifth group includes those letters which are spoken by using the lips. A letter from this group can use to hide '100' bits. Candidates for this group are

उ	ऊ	प	फ	ब	भ	व

f) Nasal group G6: In the sixth group, those letters which are pronounced from the nose. A letter from this group used to hide '101' bits. Candidates for this group are :

ङ	ञ	ण	न	म

g) Conjuncts group G7: Conjuncts are those word which are the combination of two vowels or constant word. In seventh group, conjuncts words are included. These word in this group can hide '110' bits. The candidates for seventh group are:

ए	ऐ	ओ	औ	क्ष	त्र	ज्ञ

h) Agyovaha group G8: Letters from this group can hide '111' bits. In last Eighth group, letter such as . and : are candidate for this group.

ॅ	-	:

IV. PERFORMANCE RESULTS

We have taken five different secret message sizes and cover text size viz; 200 bytes, 400, 600,800 & 1000 Bytes. We have checked the number of bytes can hide by some existing methods and both proposed approaches using random character sequence. We have measured time overhead and number of bytes can be hidden for various methods. Table 5 shows results of our proposed approaches for 200 bytes.

TABLE V . OVERHEAD IN VARIOUS TEXT STEGANOGRAPHY APPROACHES

Text Steganography Approach	Message Text Size (Bytes)	Cover Text Size (Bytes)	No. of bytes can hide (Bytes)	Time Overhead (ms)
Feature Coding Method	200	660	13	18,158
Inter Sentence space Method	200	660	1	19,276
Inter Word space Method	200	660	14	20,906
Random Character Sequence Method	200	660	14	28,100
CASE based Text steganography Method	**200**	**660**	**66**	**1,672**
ISET Hindi Text Steganography Method	**200**	**660**	**66**	**1,973**

A. *Bytes Hidden*

Fig. 3 shows the numbers of bytes hide by some existing methods and both proposed approaches. Using Indian Script Encoding Technique (ISET) and Capital Alphabet Shape Encoding (CASE), we can cover up more numbers of bytes. The least number of bytes can cover up by Inter-sentence space method because space between two sentences is used to hide secret bits.

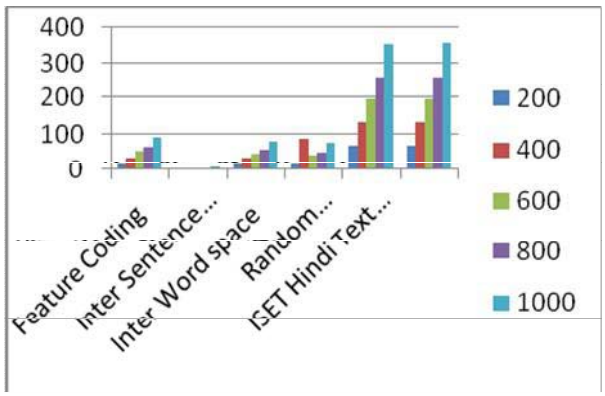

Fig.3. Bytes hidden

B. *Required Cover Text*

Fig. 4 shows maximum cover text size requires when text message of size is 200 bytes and 1000 bytes. Inter-Sentence space method required maximum cover text size to hide secret bits, while a minimum number of cover text size requires, to hide secret bits is the ISET and CASE approach.

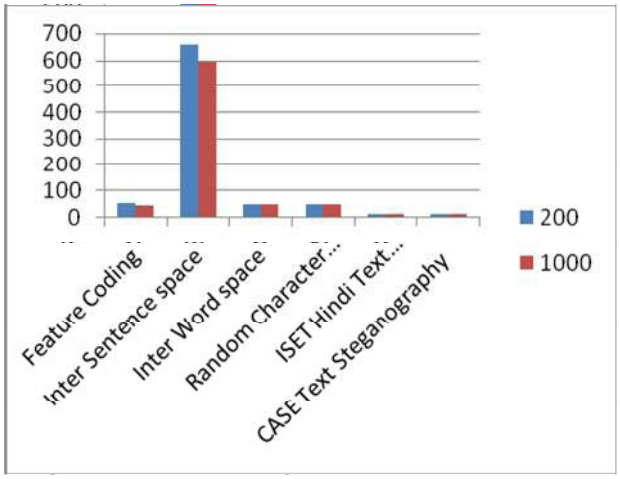

Fig.4. Required Cover text

C. *Formal Analysis*

The computer programs easy recognize the use of deliberate misspelling and space method and it can also render it to display the message. The proposed approaches in this paper are based on features of English and Hindi language, the encoded text is also very similar to random cover text [13]. Even the text after embedding it in the cover text is also random. We can see in table V that the proposed techniques reduce the memory uses and over head along with size of cover text used for Steganography.

In comparison with previous approaches in term of ease of cracking, then open space method is well known, very popular and easily noticeable hence cracked easily then feature encoding method. Security of Random character sequence is its randomness [9]. Text in this approach may seem as garbage text [14], even if someone notice it then also they can't be decode it until they known hiding key technique. The proposed technique uses feature coding to group the English and Hindi characters and random character sequence for hiding them. So, it is hard to crack as it has goodness of both the properties of feature coding and random character methods [12].

V. CONCLUSION

In this paper, we have proposed new approaches for text-based steganography for English and Hindi language texts. In the CASE method, we exploit the shapes of the English characters to hide secret bits so this technique can be used as encryption for data [16]. Based on our survey of the existing Text Steganography approaches, we show that our proposed approach can hide more number of bytes; it has very small cover text and required very less time overhead as compare to other techniques as shown in fig. 3 and fig. 4, that it can hide 66 bytes in 660 bytes cover text means we can say it require only 2000 byte of cover text to hide 200 byte of data while other techniques require more than this, also the time overhead is minimum as compared to other technique as shown in tables and graphs.

Our analysis reveals that our approach imparts increased randomness in encoding because of that the same cannot be attacked easily. This approach is applicable to the soft-copy texts as well as hard-copy texts. In addition, the proposed approach is also immune to retyping and reformatting of text. However, one of the weaknesses of the proposed approach is that once known about their applicability, they can easily be attacked. Hence, it is essential to keep the application of a particular approach to a particular data set secret, while using them.

VI. FUTURE WORK

Because of randomness, the use of random character sequence to hide secret data is noticeable. So in future we can also generate other ways to hide secret data like we can use sentence case, in which we can skip sentences if sentences are not according to secret data but because of that paragraph can be meaningless. Another way to hide secret data could be that we can use incorrect grammar approach. In which, we can randomly place word as starting word of a sentence. The randomly selected word would depend on secret bits and that is why this sentence will also become meaningless sometimes. So both these ways are again noticeable.

In future, we intend to carry out the formal security analysis of the method proposed as well as to extend this work to Linguistic Steganography [11] in which the syntax of a sentence and sequence of a sentence can also be taken care of, So that sentence and paragraph both will be correct grammatically.

The presented approach is best for hiding the text. But, in this method if we try to cover up the binary string directly then it increases the memory overhead as it requires 8-bits to represent 1-bit of the binary string. So in future, we will try to remove this limitation of our approach. In future, we can implement this text steganography technique to hide the image in the text.

REFERENCES

[1] M.Shirali-Shahreza, “Text steganography by changing words spelling”, 10th International Conference on Advanced Communication Technology, Korea, 2008.

[2] M. H. Shirali-Shahreza and M. Shirali-Shahreza, "A New Approach to Persian/Arabic Text Steganography", 5th IEEE/ACIS International Conference on Computer and information Science (ICIS COMSAK’06), 2006, pp. 310-315.

[3] W. Bender, D. Gruhl, N. Morimoto, and A. Lu, “Techniques for data hiding,” IBM System's Journal, vol. 35 (Issues 3 &4), 1996, p.p.313-336.

[4] K. Alla, and Dr R. Shivramprasad, "An evolution of Hindi text steganography", 6th International Conference on Information Technology, 2009.

[5] B. Dunbar, “A Detailed look at Steganographic techniques and their use in an open-system environment”, SANS Institute, 2002.

[6] K. Bennett, “Linguistic Steganography: Survey, Analysis, and Robustness Concerns for hiding Information in Text”, Purdue University, CERIAS Tech. Report, 2004.

[7] A. Gutub and M. Fattani, “A Novel Arabic Text Steganography Method Using Letter Points and Extensions”, World Academy of Science, Engineering and Technology, 2007. .

[8] M Shirali-Shahreza, and S. Shirali-Shahreza, “Steganography in TeX Documents”, Proceedings of 3rd International Conference on Intelligent System and Knowledge Engineering, 2008.

[9] Shraddha Dulera et.al. “Experimenting with the Novel Approaches in Text Steganography”, published on International Journal of Network Security & its application (IJNSA), Vol.3, No.6, November 2011, pp 213-225.

[10] K. Alla and R. S. R. Prasad, "An Evolution of Hindi Text Steganography," Information Technology: New Generations, Sixth International Conference on, Las Vegas, NV, 2009, pp. 1577-1578.

[11] S. Changder, D. Ghosh and N. C. Debnath, "Linguistic approach for text steganography through Indian text," Computer Technology and Development (ICCTD), 2010 2nd International Conference on, Cairo, 2010, pp. 318-322.

[12] A Novel Approach to Hindi Text Steganography - Mayank Srivastava, Mohd. Qasim Rafiq, and Rajesh Kumar Tiwari Page 295 Second International Conference on Advances in Communication, Network, and Computing, India, 2011.

[13] Kalavathi. Alla and R. Siva Ram Prasad, “A Novel Hindi Text Steganography Using Letter Diacritics and Its compound Words”, IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.12, December 2008

[14] Rajesh Shah and Yashwant Singh Chouhan, “Encoding of Hindi Text Using Steganography Technique”, International Journal of Scientifi International Journal of Scientific Research in Comp c Research in Computer Science and Engineering, vol-2 issue-1, 2014

[15] Sunita Chaudhary, P. Mathur, T. Kumar and R. Sharma., “A Capital alphabet shape encoding(CASE) based text steganography” Conference on Advances in Communication and Control Systems, Atlantis press,pp.120-124,2013

[16] Sunita Chaudhary, Dr. Meenu Dave and Dr. Amit Sanghi, “An Elucidation on Steganography and Cryptography”, Second International Conference on Information and Communication Technology for Competitive Strategies, ACM, March 2016, Udaipur, India