Information Hiding in Hindi and English Language using Text Steganography

Samarth Saxena[1], Shivani Dubey[2]

[1]*MCA Student, Department of Computer Applications, J.S.S. Academy of Technical Education, Noida, India*
[2]*Assistant Professor, Department of Computer Applications, J.S.S. Academy of Technical Education, Noida, India*

Abstract - *With the unceasing use of web and other online resources, it turned out that copying, distributing, and transmitting digital content over the Internet was extremely easy. Because the text is one of the key data sources available and the most commonly used digital media on the web, only the plain text is the significant part of blogs, books, posts, daily paper and so forth. Therefore, transmission of data without any disruption or illegal modifications or exploitation, so that the information could be made available to the right people and at the right time maintaining its integrity and confidentiality is need of the hour. This paper proposes DLES (Dual Language Encryption using Steganography) system with two frames of an algorithm, one for the cover text along with secret message generation from the sentence and another that will extract the original message from its hidden form for the purpose of information hiding of the texts based on Hindi and English language using the concept of text steganography. This study will help in maintaining integrity of information in such a way that information transmitted in a secret manner would indeed be indistinguishable from white noise and there will be no signs of its presence, even though the message is suspected by the potential intruder.*

Key Words: Information Hiding, Steganography, Integrity, Text Security, DLES

1. INTRODUCTION

Hindi and English are commonly used and of great importance as a means of communication among people who use it for their daily lives. It has not only been commonly used as a form of communication language but also recognized by many people across the globe. So, when someone uses it for knowledge sharing, it comes in a position to make it accessible only to the right people.

Following the rapid development of the Internet and the proliferation of electronic services, digital publishing has become an important subject, and offices (e.g. agencies and publishers) tend to be paperless in the next generation of organizations. There are currently numerous studies underway to implement and coordinate ideas such as e-commerce, e-government, and online libraries. Digital publishing has many advantages, but it does have some fundamental challenges, such as the unauthorized use of any content, the manipulation of data and the dissemination of such information [7, 8]. In this case, some security solutions consisting of protection of reputation, honesty, and Confidentiality is necessary to avoid problems of forgery. Manipulating and reusing a text without any control these days is easy; therefore protection of this knowledge from processes of alteration and reproduction is of utmost concern [3-5]. Text data has a lower capacity for text retention (i.e. relative to other digital media such as image, audio, and video). In addition, in the present era, the text is still a major type of widely accessible digital media and is a significant part of people's contact with other media.

The different categories of information security systems are depicted in Figure 1. The cryptography and information hiding is a security system used to secure data from intruders. Many malicious users also choose to create signs of cuts, manipulations, and infections [10].

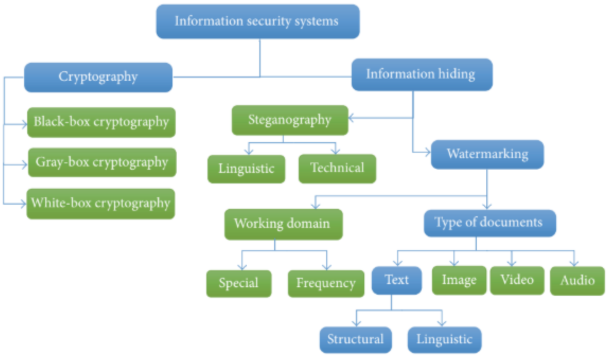

Figure-1 Different categories of information security systems

Cryptography scrambles the plain text into a cipher text that is reversible without data loss. The purpose of cryptography is to prevent unauthorized access to secret information by scrambling the content of the information. At the other hand, data hiding is an effective protection strategy that hides secret data in cover media (e.g. text, image, audio, or video) so that the evidence of embedding hidden data is completely unnoticeable. Data hiding is about trying to cover up data intentionally. Generally, data hiding can be further classified as steganography and watermarking. The purpose of the steganography is to hide a secret message in the cover-up media in order to transmit the secret information; therefore, the main concern is how to conceal the secret information without raising suspicion. That is, the steganography needs to mask the fact that the message is concealed.

Steganography has played a major role since the pre-World War and will continue to function until this human being is alive and is intended to preserve its data and messages from unknown persons. People have been using it for a lot of days in various forms. Steganography can be interpreted as the art and science of embedding secret messages into cover messages in such a way that no one, apart from the sender and recipient, knows the presence of a message. This research provides a solution through which steganography can be implemented on Hindi and English texts with a similar algorithm and more precisely for a mixed sentence that too without much change in a file size as well.

2. BACKGROUND

The first description about the use of steganography traced back to the Greeks period. Herodotus tells how a message or instruction was passed to the Greeks authority [1] about Xerses' hostile intentions and plans underneath the wax of a writing tablet, and describes a technique of dotting successive letters in a cover text having a secret ink, due to Aeneas of the Tactician.[14]

Pirate legends tell of the practice of tattooing secret information message, such as a map, on the head of someone bald, so that the hair would conceal it.

Kahn tells[16] of a trick used in China of embedding a code ideogram at a prearranged position in a dispatch for transfer, a similar idea led to the grille system and was used in medieval Europe, where a wooden template or block would be placed over a seemingly innocuous text, highlighting an embedded secret message. These were some of the available and mostly made passed practices for any language message traversal.

During World War-II the grille method or some variants were used by spies and was later revealed by some of their members only. In the same period, the Germans developed microdot technology, which prints a clear, good quality photograph shrinking it to the size of a dot so as to make it very difficult to see in general.

There are rumors [15-16] about the 1980's Margareth Thatcher, Prime Minister in UK of that time, became so irritated about press leaks of cabinet documents, that she had to implement the word processors programmed so to encode the identity of the writer in the word spacing, thus being able to trace the disloyal ministers or if any other member of her cabinet.

During the "Cold War" period, U.S. and USSR wanted and was in need to hide their sensors in the enemy's facilities. These tools had to send data to their nations, without being spotted. These might some of the noticed practices of them sending. Exchanging information, but this has been in great use throughout the journey and life of mankind.

3. TEXT STEGANOGRAPHY

Text steganography can be majorly divided or classified into three basic categories –

1. Format-based. 2. Random and statistical. 3. Linguistic.

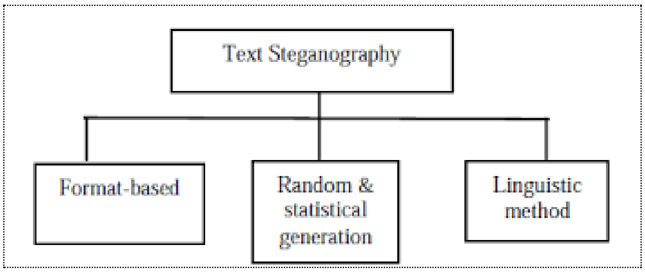

Figure-2 Classes of text steganography

3(A). Format-based: This technique uses to modify the cover-text formatting to conceal data. They change no word or phrase, so it doesn't damage the cover-text's 'value.' Open space method is a format-based Text steganography method [9, 11]. To hide content, in this method extra white spaces are added to the text. You should add some white spaces afterwards. Every word, sentence or paragraph ends. A single space is defined as '0' and two successive spaces as '1.' While in a document a small amount of data can be hidden, this approach can be extended to almost all forms [2] of text without exposing the presence of the hidden data. Another two approaches that are based on format are word shifting and line shifting. In word shifting process, the horizontal alignments of some words are shifted by adjusting distances to embed information between words [10]. These adjustments are difficult to understand because the different distances between words in documents are very common.

3(B). Random and Statistical: Random and statistical method of generation are used to automatically produce cover-text according to the language's statistical properties. These methods use grammars for example to generate cover-text in a particular natural language. A probabilistic context-free grammar (PCFG) is a widely used language model with a probability [11] associated with each transformation rule of a context-free grammar has a possibility associated or joint with it. APCFG can be used to generate word sequences by beginning with the root node and applying randomly selected rules recursively. The sentences are constructed according to the secret message to be hidden inside. The quality of the stego-message produced is directly dependent upon the grammar standard used. Another alternative to this type of method is to produce terms that have the same statistical properties in the original message such as word length and letter frequency of a word. The created words are often devoid of any lexical meaning.

3(C). Linguistic Method: The linguistic method takes into account the linguistic properties [12] of the text in order to modify it. The method uses message language structure as a place to hide information. Syntactic method is a linguistic steganography method where certain punctuation signs such as comma and full-stop are put in the document in the proper places to embed a data. This approach involves careful identification of the positions in which the sign scan is inserted. Another form of linguistic steganography is Semantic. In this system the synonym for certain preselected words is used. To cover details in it, the words are replaced by their synonyms.

4. METHODOLOGY

The proposed DLES (Dual Language Encryption using Steganography) system makes use of two frames of algorithm, one for the cover text along with secret message generation from the sentence and another that will get the hidden message to its original form. Here we make use of 2 different tables that has been devised for both Hindi and English language depending upon vowels and consonants in each language.

English alphabet consists of 26 symbols, including 21 consonants and 5 vowels. On the other hand Hindi language is of 49 different symbols, out of which 36 are consonants and 13 are vowels. Besides this, there are around 9 Diacritics in Hindi language which are also taken into consideration.

4(A). Proposed formation of tables:

There are 2 tables that will form a part of encryption for hiding secret message through steganography in DLES. The first table will be having all the vowels and diacritics of Hindi language along with consonants of English language. The second on the converse will be containing all consonants of Hindi language along with the vowels of those of English language. There will be a special arrangement for every element (referred as key-value pair) in a table having a key and a value assigned to it. The key will be simply a natural number, and both tables will be having a Parity Code as 0 or 1 respectively, which will be used to know as to which table the encrypted message is referring.

4(B). Tables along with their codes:

The first table is consisting of total 43 elements out of which there are 13 vowels and 9 diacritics of Hindi within which 21 consonants from English is added on. The key in which is a continuous set of number starting from 1.

The first table is so designed that no two elements are adjacent from a similar language. Without making any changes in their order of occurrence, the table is formed with values in actual order of presence in their respective languages.

Table – 01 (Parity Code-0)							
Element		Element		Element		Element	
Key	***Value***	***Key***	***Value***	***Key***	***Value***	***Key***	***Value***
1	अ	12	H	23	अं	34	V
2	B	13	ऋ	24	P	35	(द्व)
3	आ	14	J	25	अः	36	W
4	C	15	ए	26	Q	37	(ज़)
5	इ	16	K	27	(◌ं)	38	X
6	D	17	ऐ	28	R	39	(र्य)
7	ई	18	L	29	(◌ँ)	40	Y
8	F	19	ओ	30	S	41	(क्)
9	उ	20	M	31	(:)	42	Z
10	G	21	औ	32	T	43	(प्र)
11	ऊ	22	N	33	(प्त)		

Figure-3 Table 01 (Hindi Vowels and diacritics + English Consonants)

Next we have a second table that will be having total of 41 elements (key-value pair). Out of which there will be 36 consonants from Hindi language and 5 vowels of English language arranged at equal distance as every 8th element in a table.

Table – 02 (Parity Code-1)							
Element		Element		Element		Element	
Key	***Value***	***Key***	***Value***	***Key***	***Value***	***Key***	***Value***
1	क	12	ट	23	प	34	श
2	ख	13	ठ	24	**I**	35	ष
3	ग	14	ड	25	फ	36	स
4	घ	15	ढ	26	ब	37	ह
5	ङ	16	**E**	27	भ	38	क्ष
6	च	17	ण	28	म	39	त्र
7	छ	18	त	29	य	40	**U**
8	**A**	19	थ	30	र	41	ज्ञ
9	ज	20	द	31	ल		
10	झ	21	ध	32	**O**		
11	ञ	22	न	33	व		

Figure-4 Table 02 (Hindi Consonants + English Vowels)

4(C). Proposed Algorithm:

For implementing a cover text for either Hindi, English or mixed sentences. We consider 8-bit binary number for containing a value for each character in a word. This value will be binary equivalent for respective key value. Both tables have 43 and 41 different keys, so for their representation in binary a maximum of 7 bit will be used.

The remaining last and the most significant bit (MSB) will be that of Parity Code (0 for 1st table, 1 for 2nd table). This will signify for which value the respective key refers. Thus, it will help in getting a steganography of text both for our English and Hindi language.

Parity Code							

8 7 6 5 4 3 2 1

4(C).1 Algorithm for Encoding:

Starting with a key for each value from a sentence, we took 8 bit representation of a key into its binary format, which is now brought to its hexadecimal format containing maximum 2 bits for each character.
Thus, just replacing another bit at most for data in memory or for storage of a particular character.

Step 01: Convert all English characters to Upper Case in a sentence.

Step 02: Fetch a value for each character one by one in a sentence.

Step 03: If a character is of English language go to step 05.

Step 04: A character of Hindi language has to be checked if it is a compound word, follow following steps for each part.

Step 05: Check its respective key and parity code from the table.

Step 06: Convert key into its at most 7-bit equivalent binary number.

Step 07: Add most significant bit with value of Parity code.

Step 08: Convert its value into hexadecimal equivalent.

Step 09: Write it down, 1 or 2-digit hexadecimal value.

Step 10: Now send the cover text value as message.

Let us take an example,

Actual: Good Day रमा,

After 1st step the resultant will be,

GOOD	DAY	रमा

This step checked for each character in a text of a sentence and then converted it into an upper case if it is a English language alphabet and left as it is for a Hindi language. This was simply to get all the characters in their desired form, so as to be referred in the table. Following different steps given, the resultant will be as shown after each step in a table below:

Steps					
Value	Key	Parity Code	7-bit Code	8-bit Binary Code	Hexadecimal Code
G	10	0	0001010	00001010	A
O	32	1	0100000	10100000	A0
O	32	1	0100000	10100000	A0
D	6	0	0000110	00000110	6
D	6	0	0000110	00000110	6
A	8	1	0001000	10001000	88
Y	40	0	0101000	00101000	28
र	30	1	0011110	10011110	9E
म	28	1	0011100	10011100	9C
आ	3	0	0000011	00000011	3
Resultant: A A0 A0 6 6 88 28 9E 9C 3					

Figure-5 Tabular Representation of Data for Encoding

4(C).2 Algorithm for Decoding:

Moving forward with the received text that has been altered by conversion process is now put to the reverse process so as to get its actual meaning in the text.

Step 01: Break received set of characters (separated by space) into different set.

Step 02: Covert the 1 or 2 digit word that is actually a hexadecimal code into its binary equivalent.

Step 03: Separate the parity code and remaining 7-bit Code.

Step 04: Convert a 7-bit binary number into its decimal equivalent.

Step 05: Move to the table based on respective parity code.

Step 06: Fetch a word of respective language and write it down.

Step 07: Repeat step 02 to 06 until no more character is left.

Step 08: Finally, try to combine if any Hindi language word possible.

Let us take an example,

In consideration we will take the previously fetched resultant as an end product of encoding algorithm:

A	A0	A0	6	6	88	28	9E	9C	3

Following the various steps of decoding algorithm (as stated) on the achieved encoded result, the step wise result is in tabular format is as follows:

Steps					
Hexadecimal Code	8-bit Binary Code	7-bit Code	Parity Code	Key	Value
A	00001010	0001010	0	10	G
A0	10100000	0100000	1	32	O
A0	10100000	0100000	1	32	O
6	00000110	0000110	0	6	D
6	00000110	0000110	0	6	D
88	10001000	0001000	1	8	A
28	00101000	0101000	0	40	Y
9E	10011110	0011110	1	30	र
9C	10011100	0011100	1	28	म
3	00000011	0000011	0	3	आ
Output: G O O D D A Y र म आ					

Figure-6 Tabular Representation of Data for Decoding

Thus, the Output at receiver end for cover text is received as:

G	**O**	**O**	**D**	**D**	**A**	**Y**	**र**	**म**	**आ**

With some human intellect, this message can be read as the actual or intended meaning: "GOOD DAY रमा"

5. RESULT ANALYSIS

The proposed DLES system of algorithm follows a steganography process that took it applicable to both the required English as well as Hindi Text language. The resultant for Hindi text with simple words were easy to read, while that with a combined alphabet with vowel + consonant or consonant + diacritics has to be read with human intellectual as it is forming a a distributed result from extraction which has just to be combined to form a complete word.

The study of steganography through this DLES system can be implemented for both Hindi, English as well as combined text language sentences also. The most important part is a formation of a data hiding technique that could make it possible for both the different languages at the same time.

Further, most of the bits were reserved in size as a hexadecimal itself was of 1 bit as that of a value for any element, in simple words alphabet of any language.

Otherwise, it was at most 2-bit in every condition due to availability of only 8-bit max of binary number format.

Using a binary conversion technique along with a hexadecimal format is quite easy to achieve with any existing tools or techniques that has been serving since so long. Further it won't be a big challenge for both the parties, or namely sender and receiver to do so. The most important and needful requirement is availability of tables.

The table has been proposed as a combination of vowel on one language with the consonant of other so that an approx of equal number of elements in both the units can be populated. Further it will be a good way to even change the parity code on the will of both the parties so as to make it changed to almost half of the possibility.

6. CONCLUSION

This paper provides two frames of a novel algorithm in its DLES system which uses 2 different tables that has been devised for both Hindi and English language depending upon vowels and consonants in each language applying the key concept of text steganography for both the languages at a same time. The concept of data hiding with least possible change in actual size of a file is formed, with use of only 1 or 2 bit hexadecimal code for each character has also been worked upon.

The add-on to it is an availability of both the different languages into one algorithm itself. Not only for processing a similar table is used, but also the receiver and sender has to go through a similar set of steps for any sort of language and sentence. It also provides an add on scope for a tool or techniques that might get involved for various different set of languages to be brought up under same head for their evaluation or conversion into a secret message.

The DLES system on a whole has a capability to work on both Hindi and English language at a same time. As a result, it is fount that this system is capable of working on mixed sentences as well. The algorithm works all fine with direct Hindi and English language of any combinations. Just has to have an intellect during more complex Hindi words. This DLES proposed system is found to be applicable for any set of inputs as well and with just a small or no change in actual size of the file. Therefore, it is a great way to add in creating a cover for any steganography methods for information hiding.

7. FUTURE WORK

The algorithm proposed in DLES system can be used for any sort of words or sentences for both these languages. Not only as an individual but a widely used messaging and informal chats involving combination of both these languages in a single sentence can be evaluated. The future work should be more focused on formulating or involving a set of more instructions especially for compound Hindi words, so that no human intellect may be required instead it can club the different portions into one. With all ease in English words there is left with not much work, but indeed for enabling higher security a cover should be introduced as well to these texts. Thus the carry of text for both these languages can be more effectively carried out.

REFERENCES

[1] Anderson R.J. and Petitcolas F.A.P., "On the Limits of steganography," J. Selected Areas in Comm., vol. 16, no.4, 1998, pp. 474–481.

[2] K. Bennett, "Linguistic Steganography: Survey, Analysis, and Robustness Concerns for Hiding Information in Text", Purdue University, CERIAS Tech. Report, 2004.

[3] F. A. P. Petitcolas, R. J. Anderson, and M. G. Kuhn, "Information hiding—a survey," Proceedings of the IEEE, vol. 87, no.7, pp.1062–1078, 1999.

[4] M. H. Alkawaz, G. Sulong, T. Saba, A. S. Almazyad, and A. Rehman, "Concise analysis of current text automation and watermarking approaches," Security and Communication Networks, vol. 9, no.18, pp. 6365–6378, 2016.

[5] M. A. Qadir and I. Ahmad, "Digital text watermarking: Secure content delivery and data hiding in digital documents," IEEE Aerospace and Electronic Systems Magazine, vol. 21, no.11, pp.18–21, 2006.

[6] A. Mardin, T. Anwar, B. Anwer, "Image Compression: Combination of Discrete Transformation and Matrix Reduction," International Journal of Scientific Research in Network Security and Communication, Vol.5, Issue.1, pp.1-6, 2017.

[7] Z. Jalil and A. M. Mirza, "A review of digital watermarking techniques for text documents," in Proceedings of the Proceeding of the International Conference on Information and Multimedia Technology (ICIMT '09), pp. 230–234, 2009.

[8] A. H. Abdullah, "Data security algorithm using two-way encryption and hiding in multimedia fles," International Journal of Scientifc &amp; Engineering Research, vol.5, no.12, pp.471–475, 2014.

[9] W.Sweldens."The lifting scheme. A construction of second generation wavelets". SIAM J. Math. Anal., 29:511–546, 1997.

[10] T. Moerland, "SteganographyandSteganalysis", May15, 2003, www.liacs.nl /home/ tmoerlan/privtech.pdf

[11] N.F.Johnson. And S. Jajodia. Steganography: seeing the unseen. IEEE Computer, 16:26–34, 1998.

[12] A Novel Approach of Secure Text Based Steganography Model using Word Mapping Method(WMM) Souvik Bhattacharyya, Indradip Banerjee and Gautam Sanyal International Journal of Computer and Information Engineering 4:2 2010.

[13] M. H. Alkawaz, G. Sulong, T. Saba, A. S. Almazyad, and A. Rehman, "Concise analysis of current text automation and watermarking approaches," Security and Communication Networks, vol. 9, no. 18, pp. 6365–6378, 2016.

[14] Arvind Kumar and Km. Pooja "Steganography-A Data Hiding Technique" International Journal of Computer Applications (0975 – 8887) Volume 9– No.7, November 2010.

[15] Khan Farhan Rafat ,"Enhanced Text Steganography By Changing Word's Spelling", FIT'09, December 16–18, 2009, CIIT, 2009, ACM.

[16] Kahn D. (1996) The history of steganography. In: Anderson R. (eds) Information Hiding. IH 1996. Lecture Notes in Computer Science, vol 1174. Springer, Berlin, Heidelberg.

Authors Profile

Mr. Samarth Saxena is a bachelor in computer applications from School of Management Sciences Varanasi, affiliated to Mahatma Gandhi Kashi Vidyapith. He is currently pursuing his masters in computer applications from JSS Academy of Technical Education, a prestigious institute affiliated to Abdul Kalam technical University, Lucknow, UP. As a final year student he is also associated with a cloud based service company for his industrial internship and experience and is working as a developer intern.

Ms. Shivani Dubey is Assistant Professor in JSS Academy of Technical Education, Noida, India. She is pursuing her PhD in Ansal University. Her research areas are distributed system, cloud computing and data mining. She has published multiple research articles in reputed National and International Journal.