VOL. 13, NO. 5, MARCH 2018 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

A COMPARATIVE STUDY ON THE ADVANCEMENT OF TEXT STEGANOGRAPHY TECHNIQUES IN DIGITAL MEDIA

Salwa Shakir Baawi, Mohd Rosmadi Mokhtar and Rossilawati Sulaiman
Center for Cyber Security, Faculty of Information Science and Technology,
National University Malaysia, Bangi, Selangor, Malaysia
E-Mail: salwa_sh2001@yahoo.com

ABSTRACT

One of the common practices applied to ensure secrecy in the modern day is through an information hiding technique known as steganography, which dates back to the ancient Greece. This study investigates digital steganography and its techniques that primarily focus on text steganography. At the same time, it also attempts to present a distinctive classification in dealing with steganography based on each technique. Three types of steganography classifications were discussed, that consist of the type of carrier file, natural key used, and the embedding techniques. Text steganography can then be further separated into three categories: format based methods, random and statistical generation, and linguistics method. Techniques belonging to each category were studied, and comparisons between each technique are introduced by highlighting the findings. This study also confirmed that there are three principal factors that need to be further explored and taken into account in the design of future steganographic systems, which are the capacity, high transparency, and security.

Keywords: information security, information hiding, teganography, text steganography, categories.

1. INTRODUCTION

Currently, information is easily available to anyone because of the development of global communication and computers. Consequently, information security has become particularly important. Information hiding is a common condition that includes various sub-disciplines in the information security field. Cryptography, steganography, and watermarking are the three primary methods of information security, as explained in Figure-1.

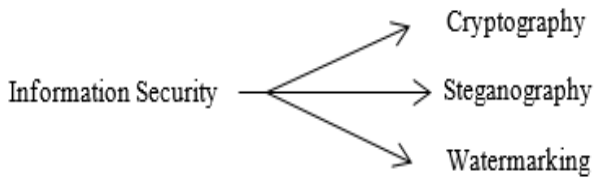

Figure-1. The main sub-disciplines of information security.

Steganography refers to the art of hiding communication. A system that uses steganography hides a secret message within another message without arousing the suspicion of other people such that the information is only detectable by its intended recipient. Steganography includes the mostcommon sub-disciplines of information hiding. The term steganography came from the Greek (στεγανογραφία) and was recorded in (1462–1516) by Johannes Trithemus as "steganographia", which literally means "covered writing"[1].

Steganography has been extensively used in the ancient times. The origins of the conventional method date back to 440 BC. For example, the Greeks hid messages using methods such as shaving the head of a messenger and tattooing a message or image on it. When the hair grows, the message would be undetected until the head is shaved again. Other methods of that time included wax tablet, invisible ink, writing on the stomachs of rabbits, and using birds as [2], [3].

Modern steganography, on the other hand, uses electronic media as opposed to using the ancient physical mediums and texts [1]. The steganography technique only embeds secret messages in carriers without changing the structure of the secret messages.

In steganography, the embedding process generates a stego file by concealing the secret information (message) in the carrier, which may also be encrypted using a stego key. The resultant file is the stego file, which has the same type as the carrier. Figure-2 shows the general steganography model[4],[5].

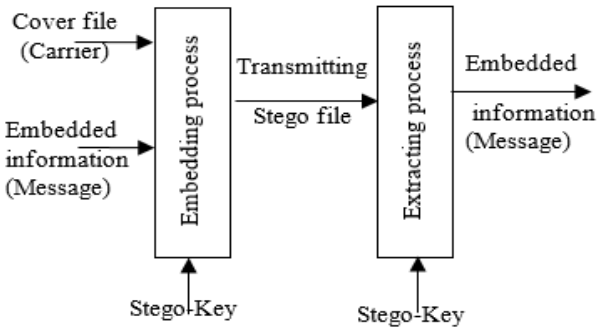

Figure-2. General Model of Steganography.

Meanwhile, the extracting process recovers information from a stego file. The stego file is a combination of the cover object into which a secret message has been embedded. The embedded message can be recovered by using the cover file or object and a decoding stego key (if one was used). In some cases, the

cover file may not be required to extract the embedded message. Therefore, the message is only accessible to a recipient who possesses the decoding stego key.

However, the process of selecting a carrier file is of asensitive nature because this contributes to the protection of the hidden information [6].

2. STEGANOGRAPHY CLASSIFICATION

According to Abbas and Hamza [7], steganography can be classified in numerous ways:

2.1 Carrier file types

Despite, the properties of carrier files vary from one type to another, depending on the redundancy created in the digital representation and the unique characteristics of the carrier file format.However, these properties control how the secret data can be hidden in the digital representation of the carrier files. According to[3],[8], the file types of carriers can be classified as text, image, audio, video, or protocol files, as shown in Figure-3.

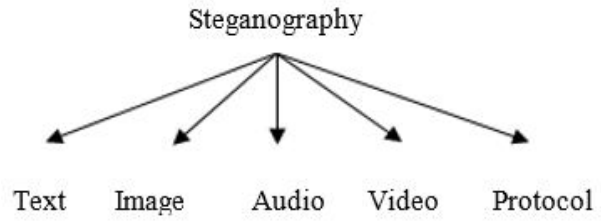

Figure-3. Carrier file types of steganography.

2.2 Types natural keys used in the embedding technique

The kinds of keys used to hide information [7]are explained below.

a) A steganographic system that does not require exchanging the encode is called pure steganography. This system is less secure because the sender and receiver assume that they (sender and receiver) are the only ones aware of the secret message.
b) In secret key steganography, exchanging a secret key is required before communication is established so that only the parties who are aware of the secret key can extract the message. The system becomes unsusceptible to the interference from third parties because of the secret key.
c) The concept of public key steganography is derived from public key cryptography. In this system, the parties who want to communicate with each other use public and private keys. During the encoding process, the sender uses a public key, and a private key is used for the process of extracting the hidden message.

2.3 Embedding techniques

Three techniques are used to embed information in a cover object: insertion-based, substitution-based, and generation-based techniques [9], which are presented as follows:

a) **Insertion (Injection):** This technique specifies a few areas that will be ignored in cover files by processing applications that read this cover file and then select a suitable area within the cover file in which to embed the secret message. Given that this technique functions by adding the secret message to the cover file, it provides the advantage of retaining the cover file's contents. The drawback of this technique is the enlarged resulting stego file which may arouse suspicion. Therefore, the main aim of current algorithms is the addition of hidden messages without raising the suspicion of potential attackers. For example, most files contain a mark in the EOF or an HTML tag, which considers the embedding characteristic.
b) **Substitution:** This technique is based on the interchange of a carrier file's component with the secret message in a way that is undetectable by an attacker. This technique works by substituting some bits of information or deliberately modifying the cover file with the least amount of distortion to the cover file. Consequently, the sizes of both the stego file and the carrier file are similar. In order to avoid suspicion, it is important that a suitable replacement process is selected, which makes it necessary to select and replace only insignificant components of the carrier file.
c) **Generation:** This technique disregards the use of a cover file and instead uses a generation engine which uses the secret message as input to generate a file which appears to be regular and may be in text, music or graphic format.

3. TEXT STEGANOGRAPHY CATEGORIES

Text steganography is the most challenging because of the insufficient redundant data in textual documents compared to other digital media, such as image, audio, or video file [6], [10], [11].

Text steganography can be broadly divided into three categories as illustrated in Fig. 4: format-based method, random and statistical generation, and linguisticmethod[12][13].

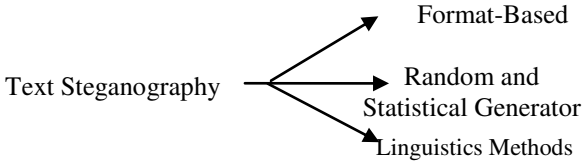

Figure-4. Categories of text steganography

3.1 Format-based method

Format based methods are achieved through the use of physical text formatting to hide information. These methods do not alter any word or sentence; therefore, retaining the value of the cover text. Several studies used format-based methods to improve the text steganography capacity by altering the physical form of the text format.

For example, Roy and Manasmita[14] attempted to develop a text steganography algorithm which uses special characters, line-shifting and word-shifting to code the text. The goal of this algorithm is to ensure a copy protection technique with high cover object capacity. The method converts the secret data from its original characters to binary data which is then hidden, making it possible to hide multiple bits in a single line of cover text which makes the changes made to the original document undetectable. However, due to the large volume of text required to encode a small number of bits and the high computational complexity, this algorithm is considered to be inefficient. Furthermore, it has low robustness given that the encoded private data is lost once the spaces are deleted in a word processing software.

Another method that employs open space coding is the format-based method presented by Por, Wong, and Chee[15]. This approach is based on a space character manipulation referred to as UniSpaCh. UniSpaCh is recommended for text steganography using Microsoft Word documents which use Unicode space characters.

UniSpaCh makes use of the white spaces which are present in any document by hiding the payload in the inter-sentence, inter-word, end-of-line, and inter-paragraph spacings using Unicode space characters. The manipulation of white spaces is a suitable technique because its effect is insignificant in the overall appearance of the document. In comparison to conventional methods, UniSpaCh is undetectable, has higher embedding efficiency and is robust when subjected to the DASH attack. Moreover, the contents of the cover document have minimal influence on the embedding efficiency of UniSpaCh.

Another format-based method is the vertical point shifting method. According to Bensaad and Yagoubi[16], the vertical point shifting method is fragile when subjected to attacks such as the use of OCR(optical character readers) and retyping using different fonts. In response to this, Odeh, *et al.*[9]developed the vertical point shifting method which examines the shifting and the point-to-point distance and then uses this to pass two bits in each multipoint letter. The stego object is then converted to an image which improves its hiding capacity and robustness thus eliminating the retyping issues. The algorithm has the added advantage of applicability to other languages, such as Pashto and Urdu. Table-1 displays possible cases using multipoint letters. This method falls in the category of feature coding using format-based methods (see Table-1).

Table-1. Various possible cases of using multipoint letters.

Secret bits	Shifting points	Distance between points
00	No shift	Normal
01	No shift	Points are separated
10	Shift up	Normal
11	Shift up	Points are separated

However, few studies on feature coding (character coding) belong to format-based methods, which deal with font attributes, such as type, case, size, color, and style (Italic, Bold, and underline). For instance, in a format-based steganography. Bhaya, Rahma, and AL-Nasrawi[17] introduced a new method for text steganography called (SEFT) using a Microsoft Word document as the carrier by changing the font with similar fonts. The secret message is embedded in the capital letters of the carrier. The benefits of this method include a very high capacity and good perceptual transparency. However, stego document increased by approximately 0.766% of the original size.

Bhaya [18] presented another text steganography method which uses an SMS message as the carrier. The font of the characters in the messageis altered such that they take either the proportional font or system font, which hold the bit values “1” and ”0” respectively.

However, this SMS service method is complex and has a low capacity because of the fixed size of SMS in mobile.

To obtain a better visual stego text quality, this method oversees the invisible space characters between words area for embedding data. Mahato et al. [19] Used one bit to hide secret information on each invisible space character by changing the font size in the Microsoft word files. The idea of the authors is that slight variation in font size of invisible letter space from other letters is not reflected in the file and in the required disk space for the file. Therefore, steganography can be intelligently achieved. Moreover, the hiding capacity will be very high. This method can be used in all languages and it can resist attacks, such as copying, cutting, changing text color, font style, and inserting a word or phrase in the text. The drawback of this method is its susceptibility to attacks of changing font size.

In 2014, text steganography in feature coding was categorized as a format-based method. Stojanov and his fellows [20] introduced another approach for embedding data called property coding using Microsoft Word files as carriers. Property coding exploited the properties of different document objects, such as characters scaling or style underline and border for both paragraphs or sentences, for embedding data. However, test results showed that this approach was not applicable to copyright protection applications which require robustness. Moreover, despite its high accuracy and transparency, the

approach introduced a slight overhead on the stego file size of approximately 1%.

Mohamed [6] presented a promising steganography algorithm which is applicable to text written in Arabic characters, which focuses on providing a high-capacity and more secure algorithm for the cover media. This method searches for these letters in the word displayed in the carrier text. Therefore, data can be hidden in the carrier text using this type of letters without any noticeable change in the target word. The single (isolated) letters at the beginning and at the end of the word and not the single letters of the word should are considered in simplifying the algorithm. The results of this algorithm showed a high carrier media capacity and high embedding capacity rate ratio. Moreover, this method is resistant to traditional attacks due to the minimal changes made to the carrier text.

Apart from the previous studies on text steganographic methods, several studies have been proposed for text steganography based on glyph (shape of characters). For example, another study by Roslan and his fellows[12] attempted to increase the capacity of text steganography by using the primitive structure of Arabic characters. This method hides the secret bits in the primitive structure (sharp edges, dots, and typographical proportion) of an Arabic letter. The results obtained from the use of this technique showed a higher capacity and perceptual transparency. However, the drawback of this approach originates from its low-level security.

3.2 Random and statistical generation

The statistical properties of a language are extracted and then used to generate cover texts. This is done with the help of the Vedic Numeric Code in the English language features proposed by Roy and Venkateswaran[21]. The frequency of the characters in the English alphabet in conjunction with the Vedic Numeric Code is used for the steganography technique. The zero distortion technique works in most cases. If there is a match in the bit values, these positions are saved in a matrix of spots. The matrix is made up of an array of information from which secret texts can be generated. In order to ensure the security of these array positions, the positions are encrypted using an indexed-based chaotic sequence. The results of the proposed text steganography method can provide two layers; one for authentication and another security layer which is applicable for physical security and applications such as online banking and online shopping. However, flaws of this method are that the size of the file is increased significantly and the larger the message, exceeding number of words required. The ratio of the secret message to original message is 1: 2 which determines that this method needed to double the required number of words to hide the secret message.

In Ryabko and Ryabko [22], the authors presented steganographic systems for the case when carriers (containers) are made via a finite-memory source with possibly unknown statistics. The probability distributions of carriers with and without embedded information are the same; this mean that the proposed stego systems are absolutely safe, i.e. an observer cannot define whether embedded information is being sent. The velocity of transmission of hidden information can be made arbitrary close to the theoretical limit to the Shannon entropy of the source of the carriers. At that place is an interesting property of the proposed stegosystem is that do not need to any secret key or public key. Moreover, This proposed stegosystem does not change the probabilistic characteristics of the source, provided the stego text file consists of independent and identically distributed (i.i.d.) equiprobable bits. Therefore, an observer cannot tell whether a stego text file is being passed. Nevertheless, the authors outline two disadvantages on steganography: one, it uses only two independent and identically distributed (i.i.d.) carriers and second, the rate of sending of secret message is not optimal.

A novel approach to text steganography is based on the curves of the English language characters. According to Dulera, Jinwala, and Dasgupta[23], the approach hides a character using a combination of the random character sequence and feature coding methods. The first step involves classifying the characters into two groups, with Group A containing letters which have curves while Group B containsletters without curves, as shown in Table-2.

Table-2. Groups based on round shape/curve.

Group	Group description	Hidden bit	Group description
A	Curved characters	0	Curved characters
B	Characters without curves	1	Characters without curves

Table-2 shows the groups based on the round shape/curve method. The next step is encoding whereby, a character which has complete or partial curves is used to represent a “0” bit, whereas a character which has no curves is used to represent a “1” bit.

Another glyph based text steganography approach is based on the vertical straight line, whereby groups are constituted based on the presence or absence of straight vertical lines. The secret message is embedded in the shape of the letter. The characters are classified into two groups with Group A containing letters with a single vertical line which represent a “0” bit, while Group B contains letters with multiple or no vertical lines which are used to represent a “0” bit (see Table-3).

Table-3. Approach based on straight vertical line.

Group	Group description	Hidden bit	Alphabetical characters
A	Multiple or no vertical lines	0	A, C, G, H, M, N, O,Q, S, U, V, W, X, Y, Z
B	Single vertical line	1	B, D, E, F, I, J K, L, P, R, T

There also exists a quadruple categorization method, whereby characters are used to represent two bits of data depending on classification based on the presence of curves, a middle horizontal straight line, a single straight vertical line, or diagonal lines as shown in Table-4.

Table-4. Quadruple categorization.

Group	Group description	Hidden bit	Alphabetical characters
A	Curves letters	00	C, D, G, O, Q, S, U
B	Letters middle horizontal straight line	01	A, B, E, F, H, P, R
C	Letters with one vertical straight line	10	I, J, K, L, T
D	Letters with diagonal line	11	M, N, V,W, X, Z

Table-4 shows the categorizations whereby alphabetical characters mostly made up of curves are used to represent “00” data bits, letters middle horizontal straight lines represent “01”, letters with a single vertical straight line represent “10” and finally letters containing diagonal lines represent the data bits “11”.

In this paper [10], Majumder and Changder presented a novel method for text steganography, which works by generating a summary of a text file whose contents are in English. The proposed approach obtains the secret message and inputs a publicly available text. The secret message is concealed in the summary depending on the reflection symmetry features of the letters of the English alphabet along the axis of reflection. For example, a cover text is generated by summarizing the selected input text, which is sent to the receiver. Upon receiving the cover text, the secret bits are extracted in order to recover the original message, based on the similarity of features in the English alphabet. The experimental results showed that this method can hide a large volume of data. However, it has weak security.

In addition, Satir and Isik[24] presented a new steganographic scheme that aims to solve capacity and security issues. This approach employed the LZW data compression algorithm for information hiding in the text document. The authors chose email as their communication medium while the stego cover was a mail forwarding platform. The embedding process used two secret keys: a global stego key made up of a collection of email addresses, and the second key is a set of selected email addresses which have been modified. Moreover, the final second key value in the embedding process is built based on the original message’s bits and the secret message bits is generatedin accordance with the algorithm. In the extraction phase, each element of the first key is compared with the second key value in order to decode the message. The pixel’s embedding capacity is determined by the cover image’s complexity. The experimental results show that an algorithm using a confidential message whose length is 300 characters, has 7.042% improved performance in comparison to existing methods.

Another research [13], discussed an embedding method which is a mixing up of lossless compression techniques and Vigenere cipher, co-authored by Tutuncu and Hassan. This study used an email environment to cover the hidden message. After choosing the cover text that has a highest repetition pattern related to the secret message the distance matrix is organized. Nonetheless, The size of the secret message was reduced by making a hybrid of Run Length Encoding, Burrows-Wheeler Transform, Move to Front (MTF), Run Length Encoding, and Arithmetic Encoding lossless compression algorithms sequence. Furthermore, a Vigenere cipher provided another layer of security or complexity to the system to obtain stego key (K1). Thus, the experimental results proved that this approach achieved the hiding capacity and also the security or complexity of the system is too improved by employing Vigener cipher.

In Kumar *et al*. [25] the authors proposed an email based text steganography scheme by using Huffman compression. This system aims to improve cover object capacity so that the cost of communication can be minimized. This scheme is achieved by using the number of characters used in email id to indicate the hidden secret data bits. Therefore, to make optimal utilization of characters number in email ids. Experimental results show that the proposed method performs better than some relevant existing methods regarding hiding capacity.

3.3 Linguisticmethod

The linguistic approach of text steganography considers the use of the language properties in text modification. Using this approach, it would not be necessary to alter the text's physical format. Mir and Hussain[26] discussed a few text steganography methods for secret communication of messages specifically in textual format. They found that although the use of synonyms and acronyms was suitable for embedding secret steganography messages in digital content, it was only secure if the lists are in the users' possession. If the lists were obtained by a third party, then this breaches the security of this method.

In addition, Shu, Liu, Tian, and Miao [27] attempted to introduce a new embedding method for text steganography, using clear text. The hidden information could be efficiently distributed by repeated extraction and thereafter embedded using any of the available carrier texts. This method could be effectively improved. Given that this method is based on semantics, it eliminates the need to change the layout, modify and delete elements of the carrier text. Therefore, this approach has higher robustness.

Another study attempted to ensure greater safety of the information exchange through a new linguistic steganography method presented by Vidhya and Paul[28]. These authors rely on the use of the Malayalam language as cover text because of the limited awareness of the local language. This method was made possible through the use of Unicode as well as embedding algorithms. In this method, two matrices are used to index the characters in the common language and in the local languages in ascending order. The method used for Unicode extraction works by selecting the Malayalam text which corresponds to the English text. A diagonal encoding scheme is used. Based on an evaluation of the efficiency of the suggested method, the method provides security, a more precise encoding process, and a balanced decoding process.

However, Shivani, Yadav, and Batham [11] proposed a new approach to increase security and improve the data hiding capacity without distorting the cover object. This technique is a hybrid of the abbreviation method with zero distortion technique used to overcome the limitation of the abbreviation method, whereby only a small amount of data can be hidden in a text file at a time. The abbreviation method provides for a reduced secret text size. Therefore, in the case of a large secret text, the text will be abbreviated thus reducing its size. This is followed by the application of the zero distortion technique so as to ensure that the cover image is not distorted and making it possible to hide large volumes of data with minimal changes to the cover image.

A new linguistic steganographic scheme for hiding data in the text file, which is based on data compression technique to protect the hidden message, co-authored by Lee and Chen [29]. This scheme aims to give a high embedding capacity and increase the security of the embedded secret message as well as reduce transmission cost. This work used a lossless compression coding which termed variable Huffman coding. In the suggested scheme, each leaf of variable Huffman tree can be employed to convey a secret bit at least. Furthermore, experimental results indicated that this proposed system accomplished high embedding capacity and reduced transmission cost. Besides the security issue are also increased and addressed.

Chang and Clark [2] have developed a novel lexicalsubstitution-based stegosystem by vertex coding that enhances the data concealing capacity compared to previous systems. The idea behind vertex coding method describes synonym replacement as a synonym graph which would be basically used for English text. Hence, the relations between words can be clearly recognized. In this manner works by replacing selected words with the same part of speech synonyms. Therefore, the modification is likely to be grammatical because it does not involve operating on the sentence structure. That is, text paraphrasing is a multi-word substitution. The evaluation results indicated that this stegosystem had achieved a small capacity by reaching the payload capacity of around two bits per sentence within a reasonable level of security. The work of Qi [30] used synonym substitution and aimed to improve the security of steganographic scheme by two novel methods to employ the advantage of abandoned synonym in traditional steganography based synonym substitution. However, experimental results showed that proposed approach obtains higher capacity compared to traditional approaches and robustness, as well as achieved a minimal creating syntax error in English text using context-based.

In Table-5, summarizes the preceding related works on text steganography and their corresponding findings. The Table shows the type of carrier and the text steganography categories used in each technique.

Table-5. The related works on text steganography.

Ref.	Techniques	Category	Carrier	Finding
[14]	Special Character, Line Shifting, and Word shifting coding techniques.	Format based method	Text file	Provides good hiding capacity and no change in stego file, but with low robustness.
[15]	Modifying inter-sentence, inter-word, end of the line and inter-paragraph spacing.		Ms-Word document	Higher embedding efficiency. Robust on DASH attack. Also, the contents of the cover file have minimal influence on the embedding efficiency.
[9]	Multi-point letters		Text file	Enhance the capacity and robustness.
[17]	Capitalizing selected letters in the cover media and changed font type.		Ms-Word document	The capacity is very high and has good transparency. However, stego document increased by approximately 0.766% from the original size.
[18]	Changing font type.		Text file	Low capacity, provide high transparency and complexity.
[19]	Changing font size of invisible letter space.		Ms-Word document	Very high capacity, but low robustness when the attacker changing the font size.
[20]	Property Coding: character scale or underline and the border of paragraph and sentence.		Ms-Word document	High capacity, but low robustness and slight increase in the size of stego-file of approximately 0.1%
[6]	Using single letters without any noticeable change in the target word.		Text file	High capacity; the embedding capacity rate ratio is also high. Resist traditional attacking methods.
[12]	Primitive structure; Sharp edges, dots, the typographical proportion of the Arabic letter.		Text file	Higher capacity and higher transparency, but low security.
[21]	Vedic Numeric Code	Random and statistical generator	Text file	Provides two layers of authentication and safety system.
[22]	Carriers (containers) are made via a finite-memory source with possibly unknown statistics.		Text file	High security, but the rate of sending of secret message is not optimal and it uses only two (i.i.d) carrier.
[23]	Random character sequencing and feature coding methods		Text file	Increased randomness, thus aiding in higher security at lower overhead.
[10]	Generating the summary of a textual file.		Text file	Higher capacity and low security.
[24]	LZW data compression algorithm.		Text file	Provides a significant increment with regard to capacity.
[13]	A hybrid of Run Length Encoding, Burrows-Wheeler Transform, Move to Front (MTF), Run Length Encoding, and Arithmetic Encoding lossless compression algorithms sequence.		Email	Achieved the hiding capacity and also improved the security.
[25]	Using the number of characters used in the email id to indicate the hidden secret data bits.		Email	Improved the hiding capacity compared to some relevant existing methods.
[26]	AES algorithm and synonyms	Linguistic method	XML file	Improves the security.
[27]	An algorithm based on multi-text.		Text file	Improves the security and higher robustness.
[28]	Unicode extraction and diagonal encoding indexing.		Text file	Achieved greater security.
[11]	Abbreviation method and Zero Distortion Technique		Text file	Increase security and improves data hiding capacity.
[29]	Used a lossless compression coding, which termed variable Huffman coding.		Text file	High embedding capacity, reduced transmission cost and also increasing the security.
[2]	Replacing selected words with the same part of speech synonyms.		Text file	Achieved a small capacity within a reasonable level of security.
[30]	Employ the benefit of abandoned synonym in traditional steganography based synonym substitution.		Text file	Achieved higher capacity, robustness and a minimal creating syntax error in English text.

4. EVALUATION CRITERIA

Steganography techniques conceal a message within a cover. Criteria (factors) for measuring the strength and weakness of the data hiding algorithm are used. These factors are used to develop a steganography system. Therefore, the relative importance of each factor depends on its application. These factors are described below:

4.1 Capacity factor

The capacity factor refers to the amount of data that can hide in the cover object in comparison to the size of the cover. This is a numerical quantity which is measured in units of bit-per-bit. Two capacity measurement factors are used, namely, hiding ratio (HR) and saving space ratio (SSR) [31];

a) Hiding Ratio (HR) is used to determine the suitability of the selected cover text in hiding the desired hidden text. This ratio is given by Equation (1):

$$HR = \frac{Total\ bits\ of\ stego\ text - Total\ bits\ of\ cover\ text}{Total\ bits\ of\ cover\ text + Total\ bits\ of\ hidden\ text} \times 100\% \quad (1)$$

b) Saving Space Ratio (SSR) is used to determine the bits of text that can be saved during the embedding process. This measure helps the steganographer establish the maximum key space that can be used in the cover text to embed the hidden text. This ratio is given by Equation (2).

$$SSR = \frac{Total\ bits\ of\ expacted\ stego\ text - Total\ bits\ of\ stego\ text}{Total\ bits\ of\ expacted\ stego\ text} \times 100\% \quad (2)$$

4.2 Invisibility factor

Transparency or invisibility refers to the non-suspicious appearance of the stego text. This is measured using Jaro-Winkler distance which gives a quantitative measure of the similarity between the cover text and the stego file. The two strings are compared and a Jaro distance, d_j is calculated using Equation (3). The result is then normalized to a value that lies in the range 0 to 1, where 1 indicates a high similarity[32].

$$The\ Jaro\ distance\ is: d_j = \frac{1}{3}\left(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{|m|}\right) \quad (3)$$

Where $|s_i|$ refers to the string length, m is the number of matched characters, and t is the number of transpositions.

4.3 Security factor

Security refers to the property in which the third party should be incapable of distinguishing the original and the stego-text.

4.4 Robustness factor

Robustness refers to system resistance, which means that confidential information has significant resistance against intentional and unintentional changes, such as scaling, rotation, change color or copy and paste, retyping, and OCR. This factor is important for copyright protection and watermarking application[33].

5. DISCUSSIONS

This study reveals that the ultimate challenge in text steganography is the hiding capacity that usualy caused by the insufficient redundant data in textual documents as compared to other digital media such as image, audio, and video files.

Based to Table-5, numerous text steganography techniques have been developed to enhance or improve the capacity of the carrier and provide high security such as described by Tutuncu and Hassan [13], Lee and Chen [29], Dulera *et al.*[23], Satir and Isik [24], and Shivani *et al.*[11]. Meanwhile, the work by Majumder and Changder [10] achieved high capacity and but low security. While the work by [2] achieved low capacity within a reasonable level of security. However, a few techniques can be classified as having low capacity, although they provide high transparency [18]. Besides, Kumar and his follows [25] improved the hiding capacity compared to some relevant existing method.

As mentioned in Table-5, several algorithms for text steganography can be categorized as methods which have high transparency in producing high-quality stego-file, as presented by Bhaya *et al*. [17], Roy and Venkateswaran [21], although they claimed to provide high security. While, other methods provided higher capacity and higher, with low security such as introduced by Roslan *etal.*[12] and [10], Majumder and Changder. Nevertheless, other techniques achieved high transparency and high capacity, with increased overhead on stego-file such as those techniques introduced by Bhaya *et al*.[17] and Stojanov *et al.*[20].

In Table-5, a few techniques can be classified as improving the security, such as that introduced by Mir and Hussain [26] and Vidhya and Paul[28]. Meanwhile, the work by Ryabko and Ryabko [22] achieved high security and but the rate of sending of secret message is not optimal.

Other methods are designed to improve robustness, despite being capable of providing a high capacity, such as that presented by Por *et al*.[15], Odeh *et al.*[9], Qi [30], Mohamed [6]. bisedes, the work by Shu *et al.*[27] provided a high robustness although they claimed to provide high security. Whereas other methods have low robustness and high capacity, such as Roy and Manasmita [14] and Mathato *et al*. [19]. Besides, other techniques achieved low robustness and high capacity, with increased overhead on stego-file such as the work introduced by Stojanov *et al*.[20].

The findings also provide evidence that high capacity, high transparency and security are the main

factors of a text steganographic system, which are attributed to the main goal of steganography applications in embedding substantial amount of information while preserving the quality of the carrier.

6. CONCLUSIONS

This research provides an extensive review on the latest techniques in text steganography. It also provides a classification of steganography based on each technique. Three calssification of steganography are presented, which based on carrier file types, keys used, and the embedding techniques. On other hand, text steganography methods are also divided into three categories: format based methods, random and statistical generation, and linguistics. Techniques belonging to each category were discussed. Moreover, comparisons between those techniques and their findings are highlighted. The review evidently shows that ensuring information privacy, increasing the capacity, robustness, and information transparency remain a hot topic in text steganography.

REFERENCES

[1] M. Agarwal. 2013. An Efficient Dual Text Steganographic Approach: Hiding Data in a List of Words. Computer Networks & Communications (NetCom), Lecture Notes in Electrical Engineering.vol. 131. Springer, New York, pp. 477-488.

[2] C.-Y. Chang and S. Clark. 2014. Practical Linguistic Steganography using Contextual Synonym Substitution and a Novel Vertex Coding Method. Comput. Linguist. 40(2): 403-448.

[3] S. Malik and W. Mitra. 2015. Hiding Information- A Survey. J. Inf. Sci. Comput. Technol. 3(3): 232-240.

[4] A. Majumder and S. Changder. 2013. A Novel Approach for Text Steganography : Generating Text Summary using Reflection Symmetry. Int. Conf. Comput. Intell. Model. Tech. Appl. 2013, Procedia Technol. 10: 112-120.

[5] C. Patel and N. Patel. 2015. A Survey Paper on Information Hiding on Web Pages. Int. J. Smart Device Appl. 3(1): 1-10.

[6] A. A. Mohamed. 2014. An improved algorithm for information hiding based on features of Arabic text: A Unicode approach. Egypt. Informatics J. 15: 79-87.

[7] T. A. Abbas and H. K. Hamza. 2014. Steganography Using Fractal Images Technique. IOSR J. Eng. 4(2): 52-61.

[8] R. Amirtharaj and J. B. B. Rayappan. 2013. Steganography-Time to Time: A Review. Res. J. Inf. Technol. 5(2): 53-66.

[9] A. Odeh, A. Alzubi, Q. B. Hani and K. Elleithy. 2012.Steganography by multipoint Arabic letters.in IEEE Long Island Systems, Applications and Technology Conference (LISAT).pp. 1-7.

[10] A. Majumder and S. Changder. 2013. A Novel Approach for Text Steganography: Generating Text Summary using Reflection Symmetry. Int. Conf. Comput. Intell. Model. Tech. Appl. 2013, Procedia Technol. 10: 112-120.

[11] Shivani, V. K. Yadav and S. Batham. 2015. A Novel Approach of Bulk Data Hiding using Text Steganography. 3rdInt. Conf. Recent Trends Comput. 2015 A, Procedia Comput. Sci. 57: 1401-1410.

[12] N. A. Roslan, R. Mahmod, N. I. U. R. I. Udzir, and Z. A. Zurkarnain. 2014. Primitive Structural Method for High Capacity Text Steganography. J. Theor. Appl. Inf. Technol. 67(2): 373-383.

[13] K. Tutuncu and A. A. Hassan. 2015. New Approach in E-mail Based Text Steganography. Int. J. Intell. Syst. Appl. Eng. 3(2): 54-56.

[14] S. Roy and M. Manasmita. 2011. A Novel Approach to Format Based Text Steganography. in Proc. of Int. Conf. on Communication, Computing & Security (ICCCS).ACM, New York, NY, USA. pp. 511-516.

[15] L. Y. Por, K. Wong and K. O. Chee. 2012. UniSpaCh: A Text-Based Data Hiding Method Using Unicode Space Characters. J. Syst. Softw. 85: 1075-1082.

[16] M. L. Bensaad and M. B. Yagoubi. 2011. High Capacity Diacritics-based Method For Information Hiding in Arabic Text.in International Conference on Innovations in Information Technology.pp. 433-436.

[17] W. Bhaya, A. M. Rahma, and D. AL-Nasrawi. 2013. Text Steganography Based on Font Type in MS-Word Documents. J. Comput. Sci. 9(7): 898-904.

[18] W. S. Bhaya. 2011.Text hiding in mobile phone simple message service using fonts. J. Comput. Sci. 7(11): 1626-1628.

[19] S. Mahato, D. K. Yadavand D. A. Khan. 2014. A novel approach to text steganography using font size

of invisible space characters in microsoft word document.in Intelligent Computing, Networking, and Informatics, Springer, New Delhi.pp. 1047-1054.

[20] I. Stojanov, A. Mileva and I. Stojanovi. 2014. A New Property Coding in Text Steganography of Microsoft Word Documents. in The Eighth International Conference on Emerging Security Information, Systems and Technologies.pp. 25-30.

[21] S. Roy and P. Venkateswaran. 2013. A Text based Steganography Technique with Indian Root. Int. Conf. Comput. Intell. Model. Tech. Appl. 2013, Procedia Technol. 10: 167-171.

[22] B. Ryabko and D. Ryabko. 2011.Constructing Perfect Steganographic Systems. Inf. Comput. 209: 1223-1230.

[23] S. Dulera, D. Jinwala and A. Dasgupta. 2011. Experimenting With The Novel Approaches in Text Steganography. Int. J. Netw. Secur. Its Appl. 3(6): 213-225.

[24] E. Satir and H. Isik. 2012. A compression based text steganography method. J. Syst. Softw. 85: 2385-2394.

[25] R. Kumar, A. Malik, S. Singh and S. Chand. 2016. A High Capacity Email Based Text Steganography Scheme Using Huffman Compression. in Signal Processing and Integrated Networks (SPIN), 2016 3rd International Conference on.pp. 53-56.

[26] N. Mir and S. A. Hussain. 2011. Secure Web-Based Communication. Procedia Comput. Sci. 3: 556-562.

[27] Y. Shu, L. Liu, W. Tian, and X. Miao. 2011. Algorithm for information hiding in optional multi-text. Procedia Eng. 15: 3936-3941.

[28] P. M. Vidhya and V. Paul. 2015. A Method for Text Steganography using MalayalamText. Procedia Comput. Sci. 46(Icict 2014): 524-531.

[29] C. F. Lee and H. L. Chen. 2013. Lossless Text Steganography in Compression Coding.in Recent Advances in Information Hiding and Applications, ISRL 40.pp. 155-179.

[30] C. Qi, S. Xingming and X. Lingyun. 2014. A SecureText Steganography Based on Synonym Substitution. in Conference Anthology, IEEE. p. 3.

[31] B. G. Banik and S. K. Bandyopadhyay. 2015. Review on Steganography in Digital Media. Int. J. Sci. Res. 4(2): 265-274.

[32] I. Banerjee, S. Bhattacharyya and G. Sanyal. 2011. Text Steganography using Article Mapping Technique (AMT) and SSCE. J. Glob. Res. Comput. Sci. 2(4): 69-75.

[33] A. S. Alfagi, A. A. Manaf, B. A. Hamida, and M. G. Hamza. A Systematic Literature Review on Necessity, Challenges, Applications and Attacks of Watermarking Relational Database. J. Telecommun. Electron. Comput. Eng. 9(1-3): 101-108, 201AD.