MJoSHT
Malaysian Journal of Science, Health & Technology
[mjosht.usim.edu.my]

USIM

Article

A Review Analysis for Text Steganography

Anes.A.Shaker [1,a], Farida Hazwani Mohd Ridzuan[1,b, 2] Sakinah Ali Pitchay[1,c, 2]

[1] *Faculty of Science and Technology (FST), Universiti Sains Islam Malaysia, Nilai, Negeri Sembilan, Malaysia*

Email: [a]anes_a_shaker@yahoo.com, [b]farida@usim.edu.my, [c]sakinah.ali@usim.edu.my

[2] *Islamic Science Institute (ISI), Universiti Sains Islam Malaysia, Nilai, Negeri Sembilan, Malaysia*

***Abstract*—Securing data is considered as a very challenging issue. The data that travels over the Internet could be modified, altered or stolen by hackers and spies. Steganography thus plays the role to secure the modern communication. Steganography hides the existence of the message. Therefore, this research paper presents a systematic review analysis on the existing techniques in text steganography. Based on the review made, the weaknesses and strengths of the methods have been identified. This paper can be used as a good reference and guidance for further studies on steganography.**

***Keywords*— Steganography**

I. INTRODUCTION

Steganography can be described as the concealment of confidential messages through implanting these messages into other apparently regular messages, graphics or sounds [1]. Steganography can be described as the study of imperceptible interaction. For the most part, it has to do with the means of concealing the presence of data to ensure its confidentiality is maintained. Confidentiality through text steganography is realized by implanting data into the cover text and the generation of a stego-text. The steganography methods come in a variety of forms, each with its own benefits and setbacks. Five distinct security and data concealing techniques are utilized for the implementation of steganography [2].

This paper is an extended version of the work published in [3]. The structure of the paper is as follows: the related works are presented in Section II, the comparison between methods are presented in Section III and finally, the conclusions are presented in Section IV.

II. EXISTING TEXT STEGANOGRAPHY WORK

Steganography methods come in a variety of forms, such as text, audio, and image. However, this paper only focuses on text steganography. Existing methods on text steganography are explained in the next subsection.

A. Word spelling method

The author presents a new text steganography method for hiding data in English texts. This method is based on substituting US and UK spellings of words. English words have different spelling in US and UK. For example, “program” has different spellings in UK (programme) and US (program). In this method, the data is hidden in the text by substituting such words [4].

B. Semantic Method

Semantic methods are comparable to syntactic methods. Instead of encoding binary data by taking advantage of the vagueness in appearance, these methods allocate two synonyms: primary value or secondary value. As an example, the word ‘big’ may be deemed primary and the word ‘large’ secondary. Whether a word comes with a primary or secondary value is of no consequence to the frequency of its usage. However, during the decoding process, primary words will be taken as ones, while secondary words will be taken as zeros [5].

C. Line-shift Coding Method

This method involves the barely perceptible upward or downward shift of each even line in accordance to the value of a particular bit. In the event, the bit is one, the shifting of the corresponding line is upwards, or else, the shift is downwards. As the odd lines are deemed control lines, they remain stationary [6].

D. Word Shifting Method

In this method, the word is relocated to the left or right, while immediate adjoining words are left stationary. These stationary words can then be used as reference locations during the decoding process. Structured documents with justified text more often than not use alterable spacing between words to disseminate white space in a manner that is pleasing to the eye. Readers are

acceptable to a broad disparity in text setting within a line, and apparently, horizontal word displacements of 1/150 are more likely to be overlooked. As the word spacing in the primary document is irregular, the detection of a word displacement calls for information on the initial word spacing [7].

E. Syntactic Methods

This method entails the exploitation of punctuation marks such as (.) and (;) to denote concealed transcripts. For instance, "NY, CT, and NJ" are comparable to "NY, CT and NJ" where the comma prior to 'and' denotes 1, and the other denotes 0. From the perspective of steganalysis, the inconsistent use of punctuation marks will not go unnoticed [8].

F. The Utilization of Letter Points and Extensions

The Arabic language come with dots. While these dotted letters are loaded with confidential bit 'one', the letters without dots are loaded with the confidential bit 'zero'. As the confidential information needs to be in conformity with the cover-text letters, not every letter is loaded with confidential bits [9].

G. Vertical Displacement of the Points

This technique, which makes use of dotted letters, has proven to be outstandingly effective. While texts in languages such as English come with merely the two dotted letters of 'i' and 'j'. With this algorithm, '1' is encoded to move up the point, or else '0' is encoded. This process is replicated for the following dotted characters in the text as well as the following bits of information [10].

H. Steganography Based on Arabic Diacritics

This technique employs an entirely diacritized Arabic text as the cover media. Subsequent to the reading of the initial bit of the implanted data by a computer program, it is compared to the initial diacritic in the cover media. If, for instance, the first bit to be implanted is a '1' and the first diacritic is a fatha, then the diacritic is maintained on the cover media and an index for the implanted text and the cover media are raised. However, in the event the first diacritic is not a fatha, it is taken off the cover media and the index is raised to scrutinize the following diacritic. This process is replicated up to the point when a fatha is detected. A similar process is employed in the implementation of zeros with the only difference being the search by zero will not be for the fatha, but for the other seven diacritics. The entire process is replicated until there are no bits left for concealment [11].

I. The Procedures for Inter-word and Inter-paragraph Spacing

This approach involves the concealment of data through the supplementation of added white spaces in the text. These white spaces can be positioned at the close of each line, the close of each paragraph, or in the midst of the words. This procedure can be applied to any random text and it does not alert the reading party to the presence of the concealed data [12].

J. Mixed-case Font

The concept for this procedure was formed during an Internet search for popular fonts used for chatting and presentations. The authors came across an innovative kind of font that can type capital and small letters in sequence. For instance, if one typed the word 'software', this word would appear as 'SoFtWaRe'. Sometimes the size of the letters would differ, and at other times they would be similarly sized. Armed with this newly-discovered font, the authors proceeded to develop an innovative text steganography technique for the transmission of confidential information [13].

K. Two-extension 'Kashida' Character

This process entails the transformation of secret object letters into secret bits by way of the corresponding code for each letter present in the mapping table. A single extension letter is installed after a letter is able to keep it away from the cover object in the event, whereby the secret bit is 'zero'. These secret bits are represented as follows: one extension letter will be inserted after a letter can hold it from the cover object if the secret bit is 'zero'. This process is repeated if the secret bit is 'one', but in this circumstance, the insertion involves two consecutive extension letters instead of one [14].

L. Move the Diacritic Up

The Arabic language comes with diacritics. More often than not, the inclusion of these diacritics in most Arabic texts is not obligatory. This study's procedure emphasizes on the employment of the non-obligatory characteristics of the Arabian language, which are the diacritics. The vertical shifting of the diacritic is in accordance to the character. 'Zero' denotes no change, and 'one' denotes the increased distance between the letter and its diacritics [15].

M. The Utilization of Multiple Diacritics in Arabic Text Steganography

By hitting (generating) several extra-diacritic keystrokes equivalent to the binary number denoting the message, the entire message can be concealed in a solitary diacritic mark. In this situation, take the example (110001)b as a confidential message. The first diacritic is replicated 3 additional times (3 = (11)b), the second 0 additional times (0 = (00)b), and the third 1 extra time (1=(01)b) [16].

N. High *Capacity Diacritics-based*

In this procedure, excluded diacritics are used for the concealment of secret bits. In a circumstance where the secret bit is '1', the diacritic remains in place. However, if it is '0', then the diacritic is taken out [17].

O. Reverse Fatha

The study reverses the original manner of the fatha from a small line inclining left above the letter to the right by installing new font properties. The regular fatha is used

to encode one and the reverse fatha is used to encode zero [18].

P. Enhanced Kashida

The authors encoded the initial text document with kashida in keeping with a specific key. Kashida is slotted ahead of a particular list of characters {ذ _ د_ؤ _و_أ _ا} up to the point where the key ends. Kashida is included for bit '1', and excluded for bit '0'. This process is replicated in a round robin manner until the close of the document [19].

Q. Utilizing Similar Letters with Different Codes

This innovative steganography method for Persian and Arabic texts takes into account two similarly-shaped letters ("Ya" « ى» and "Kaf" « ک»), with dissimilar unicode. The authors utilized the Persian characters « ک» or « ى» to conceal bit '0' and the Arabic characters «ك '» or « ي» to conceal bit '1' [20].

R. Recurrence Frequency of Characters

In accordance with the feature character repetition, the Arabic letters are separated into two sets. Set A holds the 14 high frequency letters, while Set B holds the remaining letters. The insertion of the kashida is implemented in two distinct situations: (a) if the key bit is '0' and the character is in Set A; and (b) if the key bit is '1' and the character is in Set B [21].

S. Utilization of the'La' Word

This procedure is based on feature coding utilizing the 'La' word. This word derives from the combination of 'Lam' and 'Alef' letters into a single word. The concealment technique is founded on the existence of two modes of these letters: special form 'La' (" ﻻ"), which comes with a unique code, and normal form 'La' ("لا"). Concealment is realized through the insertion of the Arabic extension character between the 'Lam' and 'Alef' letters. The concealment of bit '0' is achieved through the use of the normal form 'La', while bit '1' is concealed through the use of the special word [22].

T. Improved 'La' word

The authors recommended an enhanced procedure for the utilization of the "La" word. This involved the use of a different unicode of 'Lam" and 'Alef' to fashion the 'La' word into both special and normal forms. This recommendation takes into account the fact that each letter comes with four dissimilar outlines depending on its location in the word [23].

U. Sharp-edges Method

This technique exploits the sharp-edged Arabic characters for the concealment of confidential information. It is particularly efficient for bit concealment. Keys are introduced to facilitate the positioning of the secret bit. The diverse number of sharp edges in Arabic characters enhances the concealment effectiveness of bits '1' and '0'. The character with one sharp edge can conceal either secret bit '1' or '2'. At the same time, if the number of sharp edges is two, the possible bit location is 11, 10, 00 or 01 [24].

V. Text Abbreviation

This procedure is comparable to the short forms associated with the short message service (SMS). A dictionary is conceived comprising the abbreviation and meaning for each word. The accessibility of this dictionary is restricted to the interacting parties. The text abbreviation procedure works in this manner: if one sends the word 'see', for instance, it could be interpreted as 'do you understand'. Of late, the bulk of electronic communications use abbreviations for effortless and protected interactions. These communication avenues include Internet chats, email, and mobile messaging [25].

III. Comparison between Several Steganography Methods

Steganography methods come in many forms, each with its own benefit(s) and setback(s). Table 1 presents a comparison between all these methods.

TABLE1

COMPARISON BETWEEN STEGANOGRAPHY METHODS.

Method	Benefit	Setback
Word spelling	1-Concealed data is not obliterated. 2-The fact that it is newly created reduces the likelihood of its infiltration.	1-Its capacity to conceal data in the text is minimal.
Semantic method	1-It cannot be broken through retyping or the use of OCR programs.	1-Smart reader with its considerable data on synonyms and antonyms can be used to break it.
Line-shift coding method	1-It is only appropriate for printed texts. Thus, OCR (character recognition) is never used.	1-The use of OCR programmes results in the obliteration of the concealed information.
Word shifting	1-The possibility of detection is low due to the frequent alterations in the distance between the words and the fill line.	1-Awareness of the algorithm of distances can be exploited through a comparison of the current text with the algorithm and using the disparity to obtain the concealed data.
Syntactic	1-It cannot be broken through retyping or the use of OCR programs.	1-The capacity for concealed data is minimal compared to that of cover media.
Using letter points and extensions	1-It comes with security, high capacity, and potency. 2-It can be	1-Not every letter can be extended. This is due to their location in words and the form of

	applied to languages with similar texts as Arabic (including Persian and Urdu).	Arabic writing.
Vertical displacement of the points	1-It can encode a sizeable number of bits and it takes a robust OCR to detect the alterations.	1-The concealed information can go missing during any retyping or scanning process.
Arabic diacritics-based steganography	1- It is speedy. 2- It is uncomplicated and can be manually executed if the need arises.	1-Its use may give rise to suspicions as the delivery of diacritized text is currently rare.
Inter-word spacing and inter paragraph spacing approach	1-It comes with the capacity to conceal a great quantity of data bits in the cover text.	1-Its decoding algorithm is weak as the concealed data is obliterated upon the deletion of spaces by a word processing software.
Mixed-case font	1-It conceals data in 7 letters (not 7 words). This represents a huge amount of data when compared to other methods.	1-Retyping removes the whole message and this may rouse suspicions as the sending of messages in mixed-case is currently rare.
Two-extension 'kashida' character	1-It comes with a high level of security, capacity, and potency.	1-Retyping leads to the loss of all information.
Move the diacritic up	1-It is uncomplicated and can be manually applied if necessary. 2-It does not enlarge the size of the cover object.	1-Upon the detection of a similar message with dissimilar diacritics by OCR, it might suspect the presence of concealed data. 2- Retyping will result in the removal of the implanted message
Arabic text steganography using multiple diacritics	1-It presents a range of situations in which random capacities can be realized. Its operating cost is decreased if more than one diacritical bit is utilized simultaneously.	1-If OCR detects a similar message with disparate diacritics, it may come to the conclusion that the message holds concealed data. 2- Retyping leads to the removal of the implanted message.
High capacity diacritics-based	1-The amount of hidden bits is approximately double the number of visible diacritics.	1- Upon the discovery of the same message with dissimilar diacritics by OCR, it may suspect the presence of concealed data. 2- Retyping will result in the removal
		of the implanted message.
Reverse fatha	1-It can be effectively implemented for printed documents.	1-It is easy for attackers to perceive the presence of a concealed message.
An enhanced kashida	1-The recommended procedure proved to be effective for the protection of documents. 2- It is more robust than other kashida methods.	1-Its capacity is low.
Utilizing similar letters with different codes	1-Its level of imperceptibility is elevated as no visible alterations are detectable in the text.	1-The utilization of two letters of the text renders it low in terms of capacity.
Frequency recurrence of characters	1-It offers higher capacity and better imperceptibility when compared to the other kashida-based methods.	1-Retyping results in the loss of the information.
Using 'La' word	1-It is not restricted to electronic documents and can also be employed for printed documents.	1-It is hampered by low capacity as the 'La' word is limited. Its use also swells the file size and causes the text to take on a peculiar look.
Improve 'La' word	1-It does not alter the file size and the text appears natural.	1-The capacity ratio is low for above 'La'.
Sharp-edges method	1-It has a superior capacity for concealing secret bits.	1-Its security is threatened by the fact that the random position for the sharp-edges method is restricted to solely odd and even inputs of keys.
Text abbreviation	1-It reduces writing time and the space required for message writing. 2- It can control the keyboard limitation character.	1-In a situation where these abbreviations are not used in standard applications, suspicions may be roused by steganalysis systems.

As a result, the main drawback for all the above methods lies in hiding a small amount of bits, but the methods that use the extension kashida provide good capability to conceal more capacity compared to other existing methods.

IV. CONCLUSION

Steganography is a very suitable technique to achieve secrecy in communication. The comparison table that summarizes these methods with their advantages and disadvantages are presented. The diacritics-based methods are easy to implement and give good capacity, but cannot be applied in texts that use the appearance of diacritic. Kashida-based methods give good capacity, however, they can be easily detected.

REFERENCES

[1] Siper, A., Farley, R., & Lombardo, C. (2005). The rise of steganography. Proceedings of Student/Faculty Research Day, CSIS, Pace University.

[2] Kour, J., & Verma, D. (2014). Steganography techniques–A review paper. International Journal of Emerging Research in Management &Technology ISSN, 2278-9359.

[3] Shaker, A. A., Ridzuan, F., & Pitchay, S. A,"A Review Analysis for Text Steganography,"in KOSIST16,2016.

[4] Shirali-Shahreza, M. (2008, February). Text steganography by changing words spelling. In Advanced Communication Technology, 2008. ICACT 2008. 10th International Conference on (Vol. 3, pp. 1912-1913). IEEE.

[5] Sarmah, D. K., & Bajpai, N. (2010). Proposed System for data hiding using Cryptography and Steganography. International Journal of Computer Applications, 8(9), 7-10.

[6] Alattar, A. M., & Alattar, O. M. (2004, June). Watermarking electronic text documents containing justified paragraphs and irregular line spacing. In Security, Steganography, and Watermarking of Multimedia Contents VI (Vol. 5306, pp. 685-696). International Society for Optics and Photonics.

[7] Brassil, J. T., Low, S., & Maxemchuk, N. F. (1999). Copyright protection for the electronic distribution of text documents. Proceedings of the IEEE, 87(7), 1181-1196.

[8] Sarmah, D. K., & Bajpai, N. (2010). Proposed System for data hiding using Cryptography and Steganography. International Journal of Computer Applications, 8(9), 7-10.

[9] Gutub, A., & Fattani, M. (2007). A novel Arabic text steganography method using letter points and extensions.

[10] Shirali-Shahreza, M. H., & Shirali-Shahreza, M. (2006, July). A new approach to Persian/Arabic text steganography. In Computer and Information Science, 2006 and 2006 1st IEEE/ACIS International Workshop on Component-Based Software Engineering, Software Architecture and Reuse. ICIS-COMSAR 2006. 5th IEEE/ACIS International Conference on (pp. 310-315). IEEE.

[11] Aabed, M. A., Awaideh, S. M., Elshafei, A. R. M., & Gutub, A. A. (2007, November). Arabic diacritics based steganography. In Signal Processing and Communications, 2007. ICSPC 2007. IEEE International Conference on (pp. 756-759). IEEE

[12] Al-Haidari, F., Gutub, A., Al-Kahsah, K., & Hamodi, J. (2009, May). Improving security and capacity for arabic text steganography using 'Kashida'extensions. In Computer Systems and Applications, 2009. AICCSA 2009. IEEE/ACS International Conference on (pp. 396-399). IEEE.

[13] Ali, A. A. (2013). New Text Steganography Technique by using Mixed-Case Font. International Journal of Computer Applications, 62(3).

[14] Gutub, A. A. A., Al-Alwani, W., & Mahfoodh, A. B. (2010). Improved method of Arabic text steganography using the extension 'Kashida'character. Bahria University Journal of Information & Communication Technology, 3(1), 68-72.

[15] Odeh, A., & Elleithy, K. M. (2012). Steganography in Arabic Text Using Full Diacritics Text.

[16] Gutub, A., Elarian, Y., Awaideh, S., & Alvi, A. (2008). Arabic text steganography using multiple diacritics.

[17] Bensaad, M. L., & Yagoubi, M. B. (2011, April). High capacity diacritics-based method for information hiding in Arabic text. In Innovations in Information Technology (IIT), 2011 International Conference on (pp. 433-436). IEEE

[18] Memon, M. S., & Asadullah, S. (2011). A novel text steganography technique to Arabic language using reverse Fatha. Pak. j. eng. technol. sci, 1, 106-113.

[19] Gutub, A. A. A., Al-Haidari, F., Al-Kahsah, K. M., & Hamodi, J. (2010). e-Text watermarking: utilizing'Kashida'extensions in Arabic language electronic writing. Journal of Emerging Technologies in Web Intelligence, 2(1), 48-55.

[20] Shirali-Shahreza, M. H., & Shirali-Shahreza, M. (2010). Arabic/Persian text steganography utilizing similar letters with different codes. The Arabian Journal For Science And Engineering, 35(1b).

[21] Kabir, M. N., Alginahi, Y. M., & Tayan, O. (2014). An enhanced Kashida-based watermarking approach for increased protection in Arabic text-documents based on frequency recurrence of characters.

[22] Alotaibi, R. A., & Elrefaei, L. A. (2015). Arabic Text Watermarking: A Review. arXiv preprint arXiv:1508.01485.

[23] Shirali-Shahreza, M., & Shirali-Shahreza, M. H. (2008, August). An Improved Version of Persian/Arabic Text Steganography Using" La" Word. In Telecommunication Technologies 2008 and 2008 2nd Malaysia Conference on Photonics. NCTT-MCP 2008. 6th National Conference on (pp. 372-376). IEEE.

[24] Roslan, N. A., Mahmod, R., & Udzir, N. I. (2011). Sharp-Edges Method In Arabic Text Steganography.

[25] Shirali-Shahreza, M., & Shirali-Shahreza, M. H. (2007, November). Text steganography in SMS. In Convergence Information Technology, 2007. International Conference on (pp. 2260-2265). IEEE.