

A review of arabic text steganography: past and present

Suhaibah Jusoh¹, Aida Mustapha², Azizan Ismail³, Roshidi Din⁴

^{1,2,3}Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Malaysia

⁴School of Computing, College Arts and Sciences, Universiti Utara Malaysia, Malaysia

Article Info

Article history:

Received May 10, 2019

Revised Aug 1, 2019

Accepted Aug 15, 2019

Keywords:

Arabic text

Hidden message

Steganography

Text steganography

ABSTRACT

Steganography is a strategy for hiding secret information in a cover document in order to avoid attacker from predict about hidden information. Steganography exploit cover message, for instance text, audio, picture and video to hide the secret message. Before this, linguistic text steganographic techniques are implemented just for the English language. But nowadays different languages are used to hide the data like Arabic language. This language is still new in the steganography and still need practices for empowerment. This paper will present the text steganographic method for Arabic language, scholar paper within 5 year will be analyze and compared. The main objective of this paper is to give the comparative analysis in the Arabic steganography method that has been applied by previous researchers. Finally, the disadvantage and advantage of the method also will be presented in this paper.

Copyright © 2020 Institute of Advanced Engineering and Science.
All rights reserved.

Corresponding Author:

Aida Mustapha,

Faculty of Computer Science and Information Technology,

Universiti Tun Hussein Onn Malaysia,

86400 Parit Raja, Batu Pahat, Johor, Malaysia.

Email: aidam@uthm.edu.my

1. INTRODUCTION

The issue of information security has acquired unique importance by developing computers and expanding their use in distinct fields of life and work. The concept of the hidden exchange of information is one of the concerns in the field of information security. Different techniques have been used for this purpose including cryptography, steganography and coding. Steganography is one of the methods which have attracted more attention during the recent years [1]. In steganography, the data is embedded in a cover media, so the existence of the secret information will be difficult to see [2]. The steganography's primary objective is to make the concealed information undetectable by humans or machines [3]. Steganography can be assumed as fail if there is hidden data that can be detected, even the hidden data is not recovered.

The implementation of steganography can be divided to a number of mediums. Text, images [4], audio [5] and even video [6] are all commonly used as carrier for secret messages. Text steganography has received less attention in recent years, this is due to lower capacity to hide information that is associated with it. Text steganography embed the secret message in the medium of the plain text files so that the third party is unable to detect the existence of message [7]. However, steganography in the text domain is the hardest techniques compared to image and audio file due to lack of redundant data in the text file [8].

Text steganography can be divided into three categories. The first is format-based steganography; whereby this type of steganography hides the secret text by changing format of the carrier file such as word, space [8], line [9], and any other characters in the sentence of the text. The second is random and statistical generation steganography that will generate a cover text depending on the character or word sequence at random and then a random sequence of word or character will be used to hide the data. The third is linguistic steganography which covers the secret text by altering the data that have been encoded based on the linguistic order [10]. Specifically on linguistic steganography as it deals at the level of semantic [11] which is

beyond lexical level of the text. Substitution-based method is used to hide the secret message by embedding them in the text based on the replacement of another text in the original text.

Initially linguistic text steganographic techniques are implemented just for the English language such as technique in [12] that used English antonym to hiding the true intention behind the text format. But now days different language are likewise used to hide the data like Arabic, which increase the security level in order to get the secret information, since not everyone aware about all the language. The language that have attention for text steganography such as Urdu, Chinese, Arabic and Persian since they have many letters with the dot symbols [13]. Besides that Arabic language also language which became a common since its the language of holy Quran and beautiful way to communicate. The Arabic language consists of 28 characters [14]. It has a few highlights for example, the Arabic text is written from right to left and has no equivalent to capital letters as various English writings. The Arabic word could be comprising of completely associated letters or a single word may contain more than one segments [15].

Our research is based on the reality that in distinct locations of words some letters in Arabic languages have distinct forms [16]. This Arabic language function is endorsed in Unicode format and is used to hide information in documents using Unicode standard. Some work on hiding data in English texts has been performed by the previous researcher. The steganography in random character and word sequences by [17], specific information may be concealed in this sequence by creating a random sequence of characters or phrases. The sequence of characters or phrases is random in this technique, it is therefore irrelevant and draws too much attention. This technique does not seem to be steganography, but it is some sort of encryption. Next method is word shifting [18], information is concealed in the text by moving words horizontally and altering the distance between words. This method is appropriate for documents where the distance between words varies but the distance of the algorithm is easily to detected by the attacker and the attacker will be use the difference to extract the hidden information.

This paper present the different types of Arabic steganography method that already done by past researcher based on advantage and disadvantage of each method. This paper will be reviewed on which method is suitable to use in Arabic text language.

2. MATERIAL AND METHODS

In Arabic character there is no different between upper case or lower case and that is one of special features of Arabic character for text steganography. The method of text Arabic steganography is related with dots and connectivity between each character [19]. Table 1 shows review method in Arabic text steganography.

Table 1. Review Methods

No.	Authors	Methods
1	Odeh,A. / (2012) [20]	Multipoint Arabic letter
2	Odeh, A., Elleithy, K. & Faezipour, M. / (2013) [21]	Kashida
3	M. H. Shahreza, M. S. Shahreza / 2006 [22]	Pointed letter
4	Gutub, A.A., Ghouti, L.M, Elarian, Y.S., Awaideh, S.M., Alvi, A.K / 2010 [23]	Arabic diacritics
5	M.S. Shahreza, M. H. Shahreza / 2008 [24]	La steganography
6	M.H.Shirali-Shahreza and M. Shirali-Shahreza / 2008 [25]	Pseudo-space and Pseudo-connection character

2.1. Multipoint Arabic Letter

Define Odeh [20] has suggested an Arabic multipoint letter algorithm that allows us to conceal more than two bits per letter. In this research, a vertical shift point algorithm will combine to raise the size of the hidden file. The carrier's size will be continuous without alteration. After adding the information, the file will be converted to picture to prevent the issue of retyping. The primary problem with this strategy is to retype process all the hidden information, as the hidden information depends on the format of the file. The coherent format used in the scheme could increase an attacker's amount of suspicion [26].

2.2. Kashida

Existing research on Arabic language that using zero width and kashida letters by [21] which is new steganography algorithm for Arabic text. The algorithm uses some letters that can be added to other letters such as kashida and zero width character. If kashida expanded letter is added to other letters, the significance of the phrase does not alter. Kashida only can be added between letters in words, at the beginning and at the end of the word kashida can not be added, as shown in Figure 1. In other words, if a non-pointed letter with an expansion to conceal zero, pointed letter will holds one. Not all characters will conceal a bit in Kashida.

The embedded information may influence the type, size and format of the document. The possibility of being found using steganalysis tools will therefore lead to the disclosure of the concealed information.

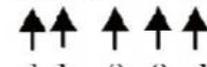
Watermarking bits	1:0010
Cover-text	من حسن اسلام المرء تركه مالا يعنيه
Output text	من حسن اسلام المرء تركه مالا يعنيه  1 1 0 0 1 0

Figure 1. Example of Kashida method [27]

2.3. Pointed Letter

This method is propose by Shirali-Shahreza [22] to hide the secret text into dots of letters as shown in Figure 2, the size of the secret text will be compress. The cover medium content is examined character by character and line by line. Location of dots may be impacted by the hidden information bit if pointed letter is discovered. This method has an advantage in secrecy and capacity of secret text. The unpointed character is remain unchanged in order to create secrecy for this method.



Figure 2. Pointed letter

2.4. Arabic Diacritics

The next method presents the Arabic diacritic [23]. As shown in Figure 3, there are eight diacritics in Arabic text. Figure 4 demonstrates some marked letters and their pronunciation. The cover text is presumed to be a fully diacritical text in this technique. The bit ‘1’ is held in the diacritics and in the non-diacritic for the bit ‘0’ is retained and the other diacritics are not used. This technique has a large ability but low invisibility because it draws the reader's attention.

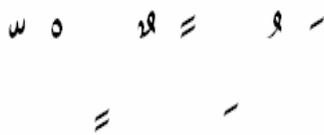


Figure 3. Arabic diacritic marks [28]

Haraka	Letter with Haraka	Pronunciation
Dama	دُ	Do
Kasra	دِ	De
Fatha	دَ	Da

Figure 4. Arabic diacritics pronunciation

2.5. La Steganography Method

The La Steganography method [24] using La word which is combination of Lam and Alef character to hide the secret information. They will hide the secret information by embedded an Arabic extension character between Alef and Lam. If they want to hide bit 0, the normal form of La will be used but if they want to hide bit 1 then special word La will be used. This method can be used in printed and not only limited to electronic document.

2.6. Pseudo-space and Pseudo-connection Character Method

The pseudo-space and pseudo-connection character method [25] will conceal one bits in each character. This technique has recognized whether or not the character is linked to the next letter. If the letter is attached to the next letter, a ZWJ letter between two letters will be inserted to conceal part 1 without changing the text structure.

3. COMPARATIVE ANALYSIS

Most of the techniques in Arabic steganography is using kashida, diacritic and multipoint. The scholarly papers that have researched in Arabic steganography in the last decade are presented in Table 1. Table 2 identified the advantages and disadvantages of Arabic text steganography in the last decade.

Table 2. Comparative Analysis

No.	Authors	Methods	Advantages	Disadvantages
1	Odeh,A. [20]	Multipoint Arabic letter	The carrier's size will be continuous without alteration.	The primary problem with this technique is the process of retyping removes all the hidden information as the hidden information depends on the format of the file. The coherent format used in the scheme could increase an attacker's amount of suspicion.
2	Odeh, A., Elleithy, K. & Faezipour, M. [21]	Kashida	Kashida method won't alter the word's significance	At the beginning and at the end of the phares, Kashida cannot be added, it can only be placed in phrases between letters. The information embedded may influence the type, file size and format of the document. The risk of being found using stegoanalysis tools will therefore lead to the disclosure of the concealed information
3	M. H. Shahreza, M. S. Shahreza [22]	Pointed letter	This method has an advantage in secrecy and capacity of secret text. The unpointed character is remain unchanged in order to create secrecy for this method.	The hidden data bit can affect the location of dots if a pointed letter is found. In the event of retyping, information will be lost.
4	Gutub, A. A., Ghouti, L. M, Elarian, Y.S., Awaideh, S.M., Alvi, A.K [23]	Arabic diacritics	This method has high capacity	This technique has a large ability but low invisibility as it draws the reader's attention to attack.
5	M.S. Shahreza, M. H. Shahreza [24]	La steganography	This method can be used in printed and not only limited to electronic document.	The main issue for this method is only suitable for Arabic language and La word is very limited in Arabic sentence. This will affect the capacity of hidden data and raise suspicion.
6	M. H. Shirali-Shahreza and M. Shirali-Shahreza [25]	Pseudo-space and Pseudo-connection character	Do not have any effect on the sentence.	The main issue with this method is when removing the Unicode of ZWJ its will impact the secret data.

Based on Table 2 and Figure 5, the advantages of techniques are seen through the high capacity to embed the secret message [22, 23] and also the effect on word that will not alter the sentence [21-25]. Besides that, the advantage of technique which introduced by [20] using method of multipoint Arabic letter is the size of carrier file will not change after secret message is embedded. Lastly the technique used by [24] can be used in printed and not only limited to the electronic document.

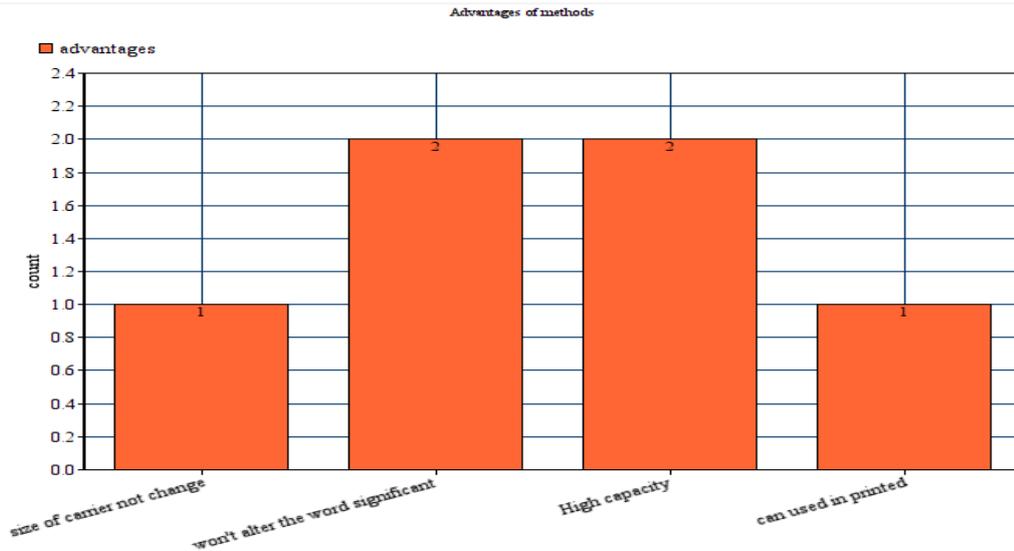


Figure 5. Advantages of methods

Based on Table 2 and Figure 6, the disadvantage of each method has been discovered. Researcher on [20] suggested an Arabic multipoint letter algorithm that allows us to conceal more than two bits per letter. An algorithm will combine with a vertical moving point algorithm in this study to boost the size of the hidden file. The main issue with this approach is retyping process removes all the hidden data [29], which is the researcher on [22] and [25] also have the issue regarding to the retyping process. Arabic language research using zero width and kashida letters by the latest steganography algorithm for Arabic text in [21]. In kashida, not all characters will hide a bit. At the beginning and at the end of the phares, Kashida can not be added, it can only be placed in phrases between letters. The information embedded may influence the type, file size and format of the document that will be attract the attacker like the technique in [23]. Therefore, choosing the most appropriate evaluation metric for implementation performance is important in order to enhance the application performance [30]. Lastly, low capacity of hidden data in technique [24] because of the La word is very limited in the arabic language.

Figure 7 shows the number of research based on point is much lower than the letter itself. Mostly the previous researcher prefer to use Arabic steganography on letter rather than point.

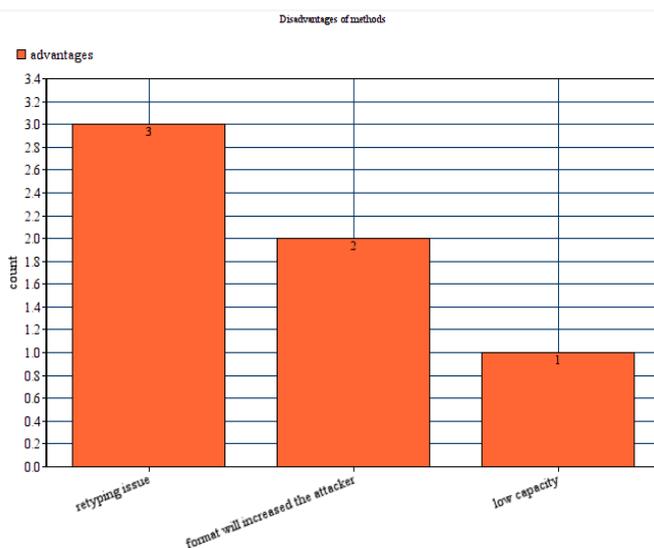


Figure 6. Disadvantages of methods

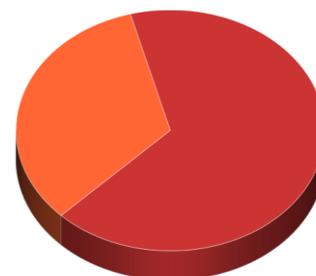


Figure 7. Previous research on pointed letter and letter on Arabic steganography

4. CONCLUSION

In order to observe the growth of these methods beforehand, this paper presented the Arabic text steganography method to reviewed. We have analyzed the list of existing technique for Arabic language text steganography in order to hide secret data. The main issue and advantages of the existing technique have been identified. Comprehensive research has been completed in the field of Arabic text steganography and every method that has been proposed have their own uniqueness and proven to be useful to other researcher. Each method will be recommended for the best secure protection, high capacity, does not affect the sentence or can used in print and electronic document. However, each method also has certain problems that seem to limit these methods such as low capacity, or will impact the size of data. The fundamental obstacle and issue that remaining parts is that in linguistic steganography it is difficult to deal with syntax and sentence structure. It will hide a small measure of information if we use other method than random character as a medium carrier.

ACKNOWLEDGEMENTS

This research is supported by the Postgraduate Research Grant Scheme (GPPS) Vot H336 from Universiti Tun Hussein Onn Malaysia.

REFERENCES

- [1] E. A. P. Petitcolas, R. J. Anderson, M. G. Kuhn, "Information Hiding – A Survey," 1999 *IEEE Proceeding*, vol. 87, no. 7, pp. 1062-1078, 1999.
- [2] J. C. Judge, "Steganography: Past, Present, Future," Technical Report No. UCRL-ID-151879, Lawrence Livermore National Lab., CA (US), 2001.
- [3] M. H. Shirali-Shahreza, M. Shirali-Shahreza, "Arabic/Persian Text Steganography Utilizing Similar Letters with Different Codes," *The Arabian Journal for Science and Engineering*, vol. 35, no. 1B, pp. 213-222, 2010.
- [4] M. Shirali-Shahreza, "An Improved Method for Steganography on Mobile Phone," *WSEAS Transactions on Systems*, vol. 4, no. 7, pp. 955-957, 2005.
- [5] G. Doerr, J. L. Dugelay, "A Guide Tour of Video Watermarking," *Signal Processing: Image Communication*, vol. 18, no. 4, pp. 263-282, 2003.
- [6] K. Gopalan, "Audio Steganography Using Bit Modification," *IEEE International Conference On Acoustics, Speech, and Signal Processing (ICASSP'03)*, pp.421-424, 2003.
- [7] A. Mangarae, "Steganography FAQ," Zone-H.org, 2006.
- [8] W. Bender, D. Gruhl, N. Morimoto, A. Lu, "Techniques for Data Hiding," *IBM Systems Journal*, vol. 35, no. 3-4, pp. 313-336, 1996.
- [9] S. H. Low, N. F. Maxemchuk, J. T. Brasil, L. O'Gorman, "Document Marking and Identification Using Both Line and Word Shifting," 14th Annual Joint Conference of the *IEEE Computer and Communications Societies*, pp. 853-860, 1995.
- [10] C. Chang, S. Clark, "Practical Linguistic Steganography using Contextual Synonym Substitution and a Novel Vertex Coding Method", *Computational Linguistics*, vol. 40, no. 2, pp. 403-448, 2014.
- [11] K. Bennet, "Linguistic Steganography: Survey, Analysis, and Robustness Concerns for Hiding Information in Text," CERIAS Tech. Report, Purdue University, 2004.
- [12] F. Z. Mansor, A. Mustapha, R. Din, A. Ismail, N. A. Samsudin, S. Utama, "Substitution-based Linguistic Steganography based on Antonyms," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 16, no. 1, 2019.
- [13] <http://www.wordribbon.tips.net/t008879-DisplayingNonPrintingcharacters.html>
- [14] M. H. Shirali-Shahreza and M. Shirali-Shahreza, "Persian/Arabic CAPTCHA," *IADIS International Journal on Computer Science and Information Systems*, vol. 2, no. 1, pp. 63-75, 2006.
- [15] Unicode.org, "The Unicode Standard," Unicode.org, 2016.
- [16] M. H. Shirali-Shahreza and M. Shirali-Shahreza, "High Capacity Persian/Arabic Text Steganography", *Journal of Applied Sciences*, vol. 8, pp. 4173-4179, 2008.
- [17] K. Bennett, "Linguistic Steganography: Survey, Analysis and Robustness Concerns for Hiding Information in Text," CERIAS Technical Report, Purdue University, West Lafayette, IN 47907-2086, 2004.
- [18] S. H. Low, N. F. Maxemchuk, J. T. Brassil, L. O'Gorman, "Document Marking and Identification using Both Line and Word Shifting," 14th Annual Joint Conference of the IEEE Computer and Communications Societies, April 2-6, 1995, *IEEE Computer Society*, Washington, USA, pp. 853-860, 1995.
- [19] K. Micha, "Introduction to Unicode," <http://www.Linux.com/archieve/articles/39911>, 2004.
- [20] A. Odeh, "Steganography in Arabic Text using Zero Width and Kashidha Letters", *International Journal of Computer Science and Information Technology*, vol. 4, no. 3, pp. 1-11, 2012.
- [21] A. A. A. Gutub, W. Al-Alwani, A. Mahfoodh, "Improved Method of Arabic Text Steganography Using Extension Kashida," *Bahria University Journal of Information & Communication Technology*, vol. 3, no. 1, 2010.
- [22] A. Odeh, K. Elleithy, M. Faezipour, "Steganography in Arabic Text Using Kashida Variation Algorithm (KVA)", *IEEE Long Island Systems, Applications and Technology Conference (LISAT)*, 2013.

- [23] A. Gutub and M. Fattani, "A Novel Arabic Text Steganography Method Using Letter Points and Extensions," WASET International Conference on Computer, Information and Systems Science and Engineering, Vienna, Austria, vol. 21, pp. 28-31, 2007.
- [24] M. H. Shahreza, and M. S. Shahreza, "A New Approach to Persian/Arabic Text Steganography," 5th *IEEE/ACIS International Conference on Computer and Information Science (ICIS 2006)*, Honolulu, USA, pp. 310-315, 2006.
- [25] A. A. A. Gutub, L. M. Ghouti, Y. S. Elarian, S. M. Awaideh, A. K. Alvi, "Utilizing Diacritic Marks for Arabic Text Steganography," *Kuwait Journal of Science and Engineering*, vol. 37, no. 1B, pp. 89-109, 2010.
- [26] M. A. Aabed, S. M. Awaideh, A. M. Elshafei, A. A.A.Gutub, "Arabic Diacritics-based Steganography," *IEEE International Conference on Signal Processing and communications (ICSPC 2007)*, Dubai, UAE, pp.756-759, 2007.
- [27] M. S. Shahreza, M. H. Shahreza, "An Improved Version of Persian/Arabic Text Steganography Using 'La' Word" *IEEE 2008 6th National Conference on Telecommunication Technologies*, 2008.
- [28] M. H. Shirali-Shahreza, M. Shirali-Shahreza, "Steganography in Persian and Arabic Unicode Texts Using Pseudo-Space and Pseudo-Connection Characters," *Journal of Theoretical and Applied Information Technology (JATIT)*, vol. 4, no. 8, pp. 682-687, 2008.
- [29] T. Moerland, "Steganography and Steganalysis," Leiden Institute of Advanced Computing Science, URL: <http://www.liacs.nl/home/tmoerland/private.pdf>, 2003.
- [30] R. Din, R. Bakar, A. Ismail, A. Mustapha, S. Utama, "Evaluation Review of Effectiveness and Security Metrics Performance on Information Technology Domain," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 16, no. 2, 2019.

BIOGRAPHIES OF AUTHORS



Suhaibah Jusoh is a student at the Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia. She received her Bachelor of Information Technology from Universiti Tun Hussein Onn Malaysia in 2016. Currently, she is currently pursuing for Masters degree in Steganography.



Aida Mustapha received the B.Sc. degree in Computer Science from Michigan Technological University and the M.IT degree in Computer Science from UKM, Malaysia in 1998 and 2004, respectively. She received her Ph.D. in Artificial Intelligence focusing on dialogue systems. She is currently an active researcher in the area of Computational Linguistics, Soft Computing, Data Mining, and Agent-based Systems.



Azizan Ismail is a lecturer at the Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia. His current research interests lie in Information, Computer Security and Communications Technology.



Roshidi Din is an Associate Professor at the School of Computing (SoC), UUM College of Arts and Sciences (CAS), Universiti Utara Malaysia (UUM). His current research interests lie in Information Security, Steganography and Steganalysis.