A New Arabic Text (Diacritics, non Diacritics) Steganography

Aliea Salman Sabir

Computer Science Department / Basrah University /Basrah . Iraq

E_amil:alieea@yahoo.com

Abstract

Steganography is way of concealing the existence of secret communication, it is a form of security through obscurity, in such a way that no one, apart from the sender and intended recipient, suspects the existence of the message.

Many different carrier file formats can be used, but text documents have been widely used and steganography in text more difficult than other media because of a little redundant information in text file. This paper, purposes a new idea for text steganography by using Unicode system characteristics and non-printing characters to hide information in Arabic texts (diacritics, non diacritics). This method can be used for Arabic MS Word documents. Our method has been implemented by Visual Basic 6.0 programming language.

This proposed method has high capacity, it can hide one bit in each Arabic letters in the cover-file , and it is not make any apparent changes in the original text. So ,it satisfies perceptual transparency.

Keywords: Arabic MS Word document, data hiding , Unicode standard, non-printing characters, Text steganography.

1. Introduction

Since the rise of the Internet one of the most important fact of information technology and communication has been the security of information. Cryptography was create data technique for securing the secrecy of communication an many different methods have been developed to encrypt and decrypt data in order to keep the message secret. Unfortunately,it is sometimes not enough to keep the contents of a message secret, also be necessary to keep the exits the message secret. The technique use to implement this, is called *steganography* [M. Rana1, B. Sangwan, J.Jangir ,2012].

Steganography is the art and science of hiding communication, a steganographic system thus embeds hidden content in unremarkable cover media so as not to arouse an eavesdropper’s suspicion. In the past, people used hidden tattoos or invisible ink to convey steganographic content. Today, computer and network technologies provide easy-to-use communication channels for steganography. Modern steganography’s goal is to keep the presence of the message undetectable from an unauthorized access [A.Oluwakemi , A. Kayode , O. Ayotunde ,2012].

All techniques must satisfy a number of requirements so that steganography can be applied correctly. There are three important parameters in designing steganography methods: The integrity of the hidden information after it has been embedded inside the stego object must be correct, the stego object must remain unchanged or almost unchanged to the naked eye, Finally, we always assume that the attacker knows that there is hidden information inside the stego object [S. Magut , 2010].

In steganography the information is hidden in a cover media so nobody notice the existence of the secret is due largely to the relative lack of redundant information in a text file as compared with a picture or a sound file [J.Memon, K. Khowaja, H.Kazi, 2008]. The structure of text document is normally very similar to what is seen, while in all other cover media types (audio, picture, video), the structure is different than what we observe, making the hiding of information in other than text easy without a notable alteration. The advantage of prefer text steganography over other media is its smaller memory occupation and simpler communication, send more information and need less cost for printing as well as some other advantages [H.M Shirali-Shahreza , M. Shirali-Shahreza,2008]. Today ,the computer system have facilitated hiding information in texts. The rang of using hiding information in text has also developed.

The aim of the proposed method it's to build strong method to hide information in Arabic text as cover media by use each Arabic letter in cover text in spite of it was diacritics or not with high capacity and perceptual transparency.

In the next section preview of the related work was introduced . in the third section our proposed method was explained by three sub section , at fist give brief introduction to the Unicode standard , in the second explain the most important feature of Arabic language and in the third we introduced our proposed steganography method . then we explain the experimental result and the conclusions.

2. Related Works

Some research on hiding information in Arabic text had been performed. Different methods are presented , most these works focuses to hide information in non diacritics Arabic text, but little work give interest to diacritics Arabic texts due the further complexity when we deal with it, some of these work are mentioned bellow :

The extensions steganography method [A. Gutub and M. Fattani,2007], the existence of the diacritics in the letters and the redundant Arabic extension character (kashida) to hide secret information bits. It uses the pointed letters with extension to hold the secret bit "1" and the un-pointed letters with extension to hold "0". This steganography method can have the option of adding extension before or after the diacritics letters. The main advantage of this method is its high capacity and also useful to other language having similar text to Arabic such as Persian and Urdu script. But the main disadvantage of this method is that it attracts the attention of the reader. This method also needs a fully diacritical text, but most of Arabic texts have no diacritic.

The pseudo-space and pseudo-connection text steganography method[6], data is hidden in Persian and Arabic unicode texts. In this method the information is hidden in text documents by using pseudo-space (zero width joiner-ZWJ) and pseudo-connection (zero width non joiner-ZWNJ) characters which are respectively prevent Arabic letters from joining or forces them to join together. It insert ZWJ letter between connected letters and ZWNJ letter between not connected letters to hide bit "1" and do not anything for hiding bit "0". This method has a high hiding capacity because it hides one bit in each letter. Also this method is not make any apparent changes in the original text and have a perfect perceptual transparency.

In non-printing Unicode characters [A. Ali,2010] method english secret message is hidden in English cover-text by using unicode standard characters. In this method each English letter had been allocate with binary code, and then, this binary code representation, depending on the assumption that each "1" and "0" is converted into ZWJ (U+200D) and ZWNJ (U+200C), respectively. This method has high capacity, it can hide (K+1) letters in a text with K characters. It does not make any apparent changes in the original text.

The ZWJ and ZWNJ regular expression text steganography method[A.F. Al-Azawi and M.A. Fadhil ,2011] ,its useful for Arabic electronic writing. It works at letter level and uses two regular expression to generate a sequence of special characters that consist of ZWJ and ZWNJ. Regular expression are used to allocate position suitable for inserting a block of non printing characters in the word of the cover text. The ZWJ hides secret bit "1" and ZWNJ hides secret bit "0". This method has high hiding capacity, it may hide any file type in an Arabic Unicode texts and it does not make any apparent changes in the original text.

Isolated characters text steganography method[A.J. Fawzi,2007], it is used to hide secret information in Arabic text by changing the code of isolated characters. The isolated character, any character not connected to other within a word. These characters have the same shape but different codes, secret data is hide in texts by using one of these two codes. To hide secret data, take each word in paragraph, and check if there is an isolated character (range 0600-06FF), then replacing it with the same glyph character (range FE70-FEFF). The main goal of this method is perceptual transparency.

non-printing unicode characters and Unicode system characteristics method [A.S Sabir and W.A Awadh ,2012], data is hidden in text documents (Word, Excel). This method depend on merging two techniques. First technique; use the special Unicode standard characters, ZWNJ and ZWJ characters to embedding the secret bits into English letters, Arabic related letter, Arabic letters separated if they connect, and English/Arabic numbers. Second method; use Unicode system characteristics to embedding the secret bits into separate Arabic letters. This method has high capacity, it can hide one bit in each letter or number in the cover file, and it satisfies perceptual transparency, by does not make any apparent changes in the original text.

3. The Proposed Method

In this paper, we present a new method for text steganography in Arabic texts (diacritics, non diacritics or partial diacritics) . Before explain the method, we mention unicode standard briefly. Then we explain the main characteristics of Arabic language and at last we explain our suggested method in full details.

3.1 Unicode Standard

The unicode standard is the international character-encoding standard used for presenting the texts for computer processing. This standard is compatible to the second version of ISO/IEC 10646-1:2000 and have the same characters and codes as ISO/IEC 10646.

unicode enables us to encode all the characters used in writing the languages of the world. This standard uses the 16-bit encoding, which provides enough space for 65536 characters; that is to say, it is possible to specify and define 65536 characters in different moulds such as, numbers, letters, symbols, and a great number of current characters in all different languages of the world. This standard covers a mathematical and technical symbols, punctuation marks, arrows, and miscellaneous marks. Moreover, because of the wideness of the space dedicated to the characters, this standard also includes most of the symbols necessary for high-quality typesetting. The languages whose writing systems can be supported by this standard are Latin (covering most of the European languages), Cyrillic (Russian and Serbian), Greek, Arabic (including Arabic, Persian, Urdu, Kurdish), Hebrew, Indian, Armenian, Assyrian, Chinese, Katakana, Hiragana (Japanese), and Hangeul (Korean).

An Arabic Unicode table (takes the range 0600-06FF) represents standard forms of all characters used in Arabic language, and another Unicode table (takes the range FE70-FEFF) that has all Arabic characters with isolated form.

unicode table has been developed to cover the characters of the languages which use Arabic writing system. Among these languages we can mention Persian, Urdu, Pashto, Sindhi, and Kurdish. This standard has detailed and careful explanations about the implementation methods including letters-connection method, the exhibition of the right-to-left and bi-direction texts [Unicode Inc ,2012].

3.2 Feature of Arabic Language

In the unicode standard, each Arabic letters has its unique code. Also, all shapes of each letter have their own code. For example, the code of letter (seen س) in the Unicode standard is 069B and the codes of different forms are FEB1 for the isolated form (ﺱ), FEB2 for the final form (ﺲ), FEB3 for the initial form (ﺳ) and FEB4 for the medial form (ﺴ). For saving the documents in the Unicode standard, only the unique code of each character is saved and the program which shows the letter will show the correct shape of letter regarding to its position in the word [J.Memon, K. Khowaja, H.Kazi,2008] Instead of using the four possible shapes of Arabic letters (including the initial form, the medial form, the final form and the isolated form) the Unicode standard provide a unique code which shows the letter in isolated form act as a word representative.

This representative letter can be used with another non-printing code which will give the required shape of the letters, so for each letter in the text, we can save it by using the representative form of letter mixed with the code of correct shape of the letter (regarding to its position in the word) [J.Memon, K. Khowaja, H.Kazi,2008]. These non-printing characters used to connects Arabic language letters in the required shape, they are: **1.** The Zero width Joiner (ZWJ): used to connect two character, (unicode = u+200D) **2.** The Zero width non-Joiner (ZWNJ): used to disconnect characters, (unicode = u+200C).[unicode Inc,2012]

3.3 Proposed Steganography Method

In this paper, a new steganography scheme is introduced for hiding text (Arabic, English, number) in Arabic word documents(diacritics (fully, partial) and non diacritics). The proposed method based on the replace the character's in the original cover text which its Unicode standard table in rage [0600-06FF] with its specific shape from the Unicode standard table from the range [FE70-FEFF] ,by proposed detected shape algorithm which detect the shape of characters(single, left, middle, right) and the way of its concoction with its neighbors with the same character as describe in Table1. Our methods work powerfully because its hide bits (0,1) in each characters in the words in spite of its appearance diacritic or not , that mean its work powerfully by hide in each character in the word in spite of these word fully diacritics (مُحَمِدْ) , partial diarcritics (مُحمِد) or non diarcitics (محمد) .

Table (1) :Describe proposal hiding method

The word	Character detected shape	Character shape in original cover text	Character unicode [0600-06FF] in original cover text	Replaced character shape with original cover text	Character unicode [FE70-FEff] replaced with original cover text
عَلي	Left	عَ	0639	عَـ	FECB
مُعِلم	Middle	ع	0639	ـعـ	FECC
مُجتَمع	Right	ع	0639	ـع	FECA
أجْتِماع	Single	ع	0639	ع	FEC9

In the case of hide bit "1", after detect its shape the original character unicode code in the range [0600-06FF] with its alternative Unicode specific code shape in the range [FE70-FEFF], in the case of diacritics character the diacritics rejoined with the character after the replacement. In the case of hide bit 0, we also make the replacement process but we add non printing Unicode characters ZWJ or ZWNJ, which are used for joining and disjoining the Arabic letters depending of detected shape (single, left , middle, right) to prevent the possible change of character appearance.

.

Table (2) describe the way of hiding bits "1" and "0" with character shape, where the underlining code represent the way of hiding "1" and bolded code represent the way of hiding "0". The example in Table 2 show that we hide bit "0" in two different way the first wit character (ـمَـ) when we fellow its replaced shape and diacritics code Unicode with ZWJ or ZWNJ codes and that deepening on its detected shape and to increase the powerful of the proposed methods because this variation make the chance of detected the hiding method semi possible.

Table (2) Hide bits "1" and "0" with character shape

The word	كمَالُ
The secret message	0101
Character detected shape	Left,middle,right,single
Character Unicode [0600-06FF] in original cover text before hiding	0643,0645,064E,0627,0644,064F
Character Unicode [0600-06FF] in original cover text after hiding	FFDB,**FEE4,064E,200D**,FE8E,**FEDD,200C** 1 0 1 0

Table (3) explaining more details about the hiding process depending on the type of secret bits, detected shape and the form of character (diacritics or not), where:

x1 : represent the current character.

x2 : represent any one of eight Arabic diacritics.

rep(x1): represent the process of replacement of the original character with its shape code .

Table (3) Explaining more details about the hiding process

Secret bits	Character detected Shape	Not Diacritic character	Diacritic character
1	Left	Rep(x1)	Rep(x1)+x2
1	Middle	Rep(x1)	Rep(x1)+x2
1	Right	Rep(x1)	Rep(x1)+x2
1	Single	Rep(x1)	Rep(x1)+x2
0	Left	Rep(x1)+zwj	Rep(x1)+x2+zwj
0	Middle	Rep(x1)+zwj	Rep(x1)+x2+zwj
0	Right	Rep(x1)+zwnj	Rep(x1)+x2+zwnj
0	Single	Rep(x1)+zwnj	Rep(x1)+x2+zwnj

For execration the size and data of secret message the proposed execration algorithm deal with the stego cover text as a stream of characters, we proposed NDFST (Non Deterministic Finite State Automate Translator) to execrate the bits of the hiding secret message, Figure (1) show the proposed NDFST- execrator .

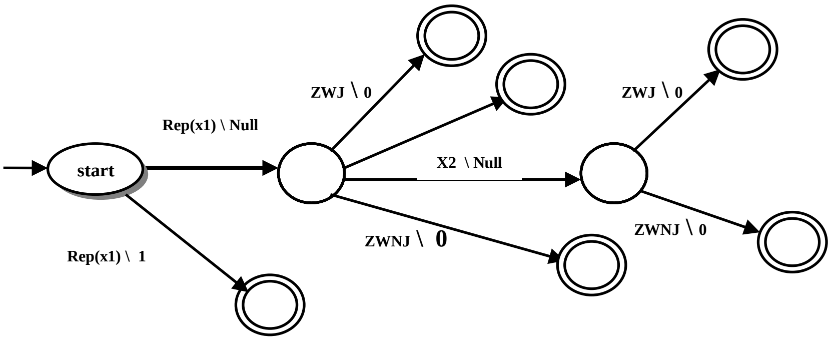
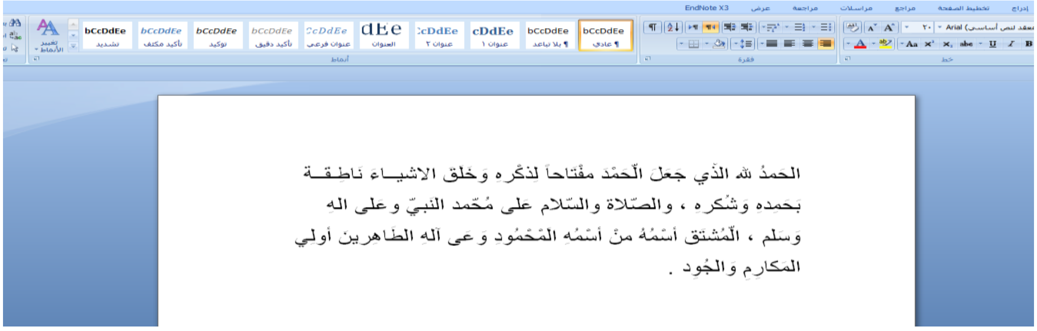


Figure (1) NDFST – execrator

The proposal method work in two stages : the first by the sender which send the secret message after covert its characters to the binary form(0,1) and hide the information in the cover text by use the proposed method , and the second by the receiver which receives the cover text and then extract the embedded binary bits of the secret message and then reconstruct it to the original secret message characters . The following figure show the majors execution steps for hiding the secret message " الاسبوع القادم" in the partial diacritics Arabic cover text. Figure 2 show the cover text before hiding and figure 3 represent the cover after hiding and then the extracted message from the cover text.


الحمدُ لله الذي جعَل الّحمّدَ مفْتاحاً لذكْرِه وَخلق الاشيـــاءَ ناطـقـــة
بَحمدهِ وَشُكره ، والصّلاة والسّلام عَلى مُحّمد النبيّ وعلى الهِ
وَسلم ، الّمُشتق أسْمُهُ منْ أسْمُهِ المْحمُودِ وَعى آلهِ الطَاهِرينَ أولِي
المَكارِمِ وَالجُودِ .

Figure(2) the cover text before hiding

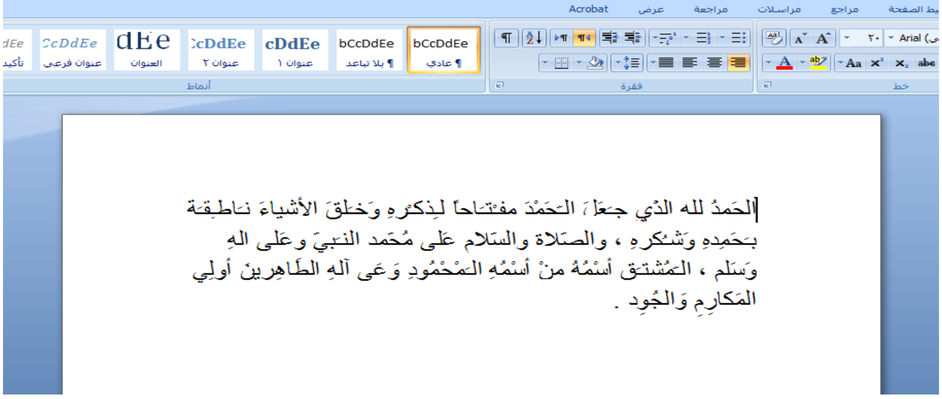
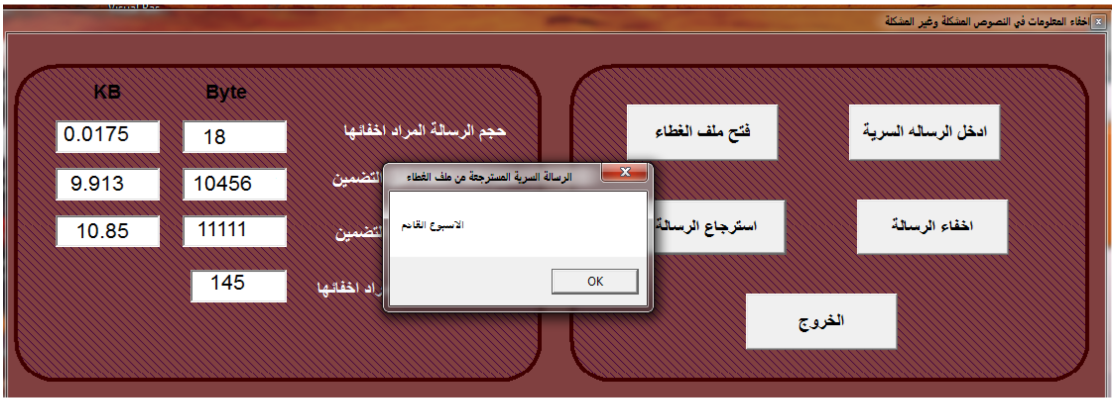

الحَمدُ لله الذي جعَلَا ً الحَمْدَ مفتاحاً لِذكرِه وَخلقَ الأشياءَ ناطِقَة
بَحَمِدهِ وَشكرهِ ، والصّلاة والسّلام عَلى مُحّمد النبيّ وعَلى الهِ
وَسَلم ، المُشتق أسْمُهُ منْ أسْمُهِ المَحْمُودِ وَعَى آلهِ الطَاهِرينَ أولِي
المَكارِمِ وَالجُودِ .

Figure (3) the cover text after hiding



اخفاء المعلومات في النصوص المشكلة وغير المشكلة

KB	Byte	
0.0175	18	حجم الرسالة المراد اخفائها
9.913	10456	التضمين
10.85	11111	التضمين
	145	راد اخفائها

الرسالة السرية المسترجعة من ملف الغطاء

الاسبوع القادم

Figure (4) the extracted message

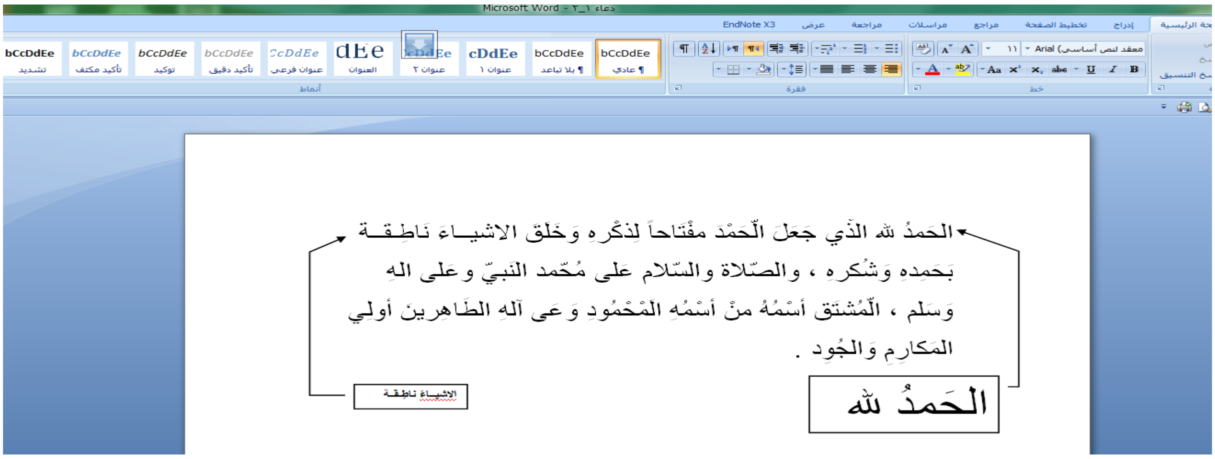
4- Experiment result

The capacity ratio can be measured by divided the amount of hidden secret bits over the number of characters in cover file. Table 4 shows the capacity ratio some of these approaches found in literature. For test the efficiency of our proposed method we test it on more than 50 cover text , and give great result from the perceptual transparency and hiding capacity .

Table (4) Hiding method review

Method type	File name	No. of Characters in cover file	No. of secret bits (bits)	Before hiding (byte)	After hiding (byte)
Proposed method	1.doc	207	128	10661	11072
Hyper Method [A.S Sabir and W.A Awadh,2012]				10661	12709
Unicode [A.J. Fawzi,2007] Method			50	10661	10661
Proposed method	2.doc	1155	1150	12541	16101
Hyper Method			1150	12541	24955
Unicode Method			60	12541	12541

As it seen in the Table (4), in unicode text steganography methods can hide a limited size of information (depended on isolated characters only). But our method capacity is very high. we hide a bit of information in each Arabic letter in cover file. So its does not change any apparent of the text and does not required specific font , its strong against the font size changes figure [5] , so it's give high transparency ,and not dependent on any special format we can save the stego text in numerous formats such as HTML pages or even plain text format. Because the stego unicode text will not change during copy and paste between computer programs, the data hidden in texts remains intact during these operation.



الحَمدُ لله الَّذي جَعَلَ الّحَمْدَ مفْتَاحاً لِذكْرِه وَخَلَقَ الاشيـــاءَ نَاطِـقـــة
بَحَمِدهِ وَشُكرِهِ ، والصّلاة والسّلام عَلى مُحّمد النَبيّ وعَلى الهِ
وَسَلم ، الّمُشتَق أسْمُهُ منْ أسْمُهِ الَمْحْمُودِ وَعى آلهِ الطَاهِرينَ أولِـي
المَكارِمِ وَالجُودِ .

Figure (5) the transparency experiment

5. Conclusion

For text steganography various method have been proposed. We represent a novel approach of hiding information in formal and non formal Arabic text. This method uses Unicode system characteristics and non-printing characters to hide secret information in (formal and non formal) Arabic texts.

Our method satisfies both perceptual transparency and hiding capacity requirement. We did not make any apparent changes in the original text by hiding data. So even if the reader has the original text, it is impossible for him to realize the hiding of the data by merely observing the appearance of the text, and it can hide one bit in each letter in the cover file. In additional to establishing secret communication, this method can be used for preventing illegal duplication and distribution of texts especially electronic texts.

The unicode standard and non-printing characters supports different languages and can be used on different systems ad devices which are supporting the Unicode standard. Moreover, the Arabic is the official language of the Muslims and about two billion Muslims live throughout the word. As a result, a wide range of the users can use our method.

References

A. Ali, "A New Text Steganography Method by Using Non-printing Unicode Characters", Eng & Tech. Journal, Vol. 28, No. 1, 2010.

A. Gutub and M. Fattani, “A Novel Arabic Text Steganography Method Using Letter Points and Extensions,” Proceedings of the WASET Int. Conference on Computer, Information and Systems Science and Engineering(ICCISSE), Vienna, Austria, vol. 21, pp. 28-31, 2007.

A.F. Al-Azawi and M.A. Fadhil, "An RabicText Steganography Technique Using ZWJ and ZWNJ regular expression", International journal of academic research, Vol. 3, No. 3, 2011.

A.J. Fawzi, "*Data hiding in Arabic text*", Ph. D Thesis, University of technology, Baghdad, Iraq, 2007.

A.Oluwakemi , A. Kayode , O. Ayotunde , "Efficient Data Hiding System Using Cryptography and Steganography", International Journal of Applied Information Systems (IJAIS), Vol. 4, ISSN : 2249-0868, No.11, Foundation of Computer Science FCS, New York, USA, 2012.

A.S Sabir and W.A Awadh, " A New Text Steganography Method by Using Non-Printing Unicode Characters and Unicode System Characteristics in English/Arabic documents ", Journal Thi-Qar Science, Vol. 3, 2012.

H.M. 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, Page No. 682-687, 2008.

J.Memon, K. Khowaja, H.Kazi, "Evaluation of Steganography for URDU/ARABIC text", Journal of Theoretical and Applied information Technology, Page No. 232-236, 2008.

M. Rana1, B. Sangwan, J.Jangir, "Art of Hiding: An Introduction to Steganography", International Journal of Engineering And Computer Science, Vol.1 Issue 1, Page No. 11-22, 2012.

N. Hopper,"Toward a theory of steganography", Ph.D. Dissertation, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA, 2004.

S. Magut, "An Overview of Digital Steganography", University of Colorado, 2010.

Unicode Inc. web site ,www.Unicode.com, 2012.

طريقة جديدة للإخفاء المعلومات في النصوص العربية
(المشكلة وغير المشكلة)

علياء سلمان صابر

قسم علوم الحاسبات / جامعة البصرة / العراق / البصرة

E_amil:alieea@yahoo.com

المستخلص

أخفاء المعلومات هي طريقة لحجب وجود الاتصال السري ، ويعد طريقة لانجاز الاتصال بسرية من خلال عملية طمر الرسالة السرية في الغطاء الحامل للرسالة بحيث لا يمكن لكل من المرسل والمستلم للرسالة يكون لديه شك بوجود الرسالة .

في عملية أخفاء المعلومات يمكن استخدام ملفات الغطاء بصيغ مختلفة مثل ملفات الصوت ، الفديو والملفات النصية . وفي الآونة الأخيرة تم استخدام الملفات النصية كملفات غطاء بشكل واسع ، على الرغم من صعوبة إخفاء المعلومات في الملفات النصية عن غيرها من الوسائط بسبب قلة تكرارالمعلومات في الملفات النصية . في بحثنا اقترحنا فكرة جديدة للإخفاء المعلومات في ملفات الغطاء النصية من خلال استخدام خصائص رموز نظام الشفرات الموحد Unicode system characteristics وبعض الرموز الخاصة الغير قابلة للطباعة non printing characters للإخفاء المعلومات في النصوص العربية (المشكلة وغير المشكلة) .والطريقة المقترحة تم تطبيقها على الوثائق النصية ضمن بيئة مايكروسوفت أوفس وقد تم تنفيذها باستخدام الفيجوال بيسك 6. .

امتازت الطريقة المقترحة بسعة أخفاء عالية من خلال أخفاء بت واحد في كل حرف عربي من حروف ملف الغطاء سواء كانت الكلمات مشكله او غير المشكلة كلياً او جزئياً ، كما أنها لا تحدث أي تغيير مرئي على كلمات وشكل ملف الغطاء بعد عملية التضمين . وبذلك فأنها حققت مبدأ الشفافية في عملية الإخفاء.

الكلمات المفتاحية : وثائق مايكروسوف العربية ، اخفاء المعلومات ، نظلم الشفرات الموحدة القياسية ، الرموز الغير مطبوعة ، اخفاء المعلومات في الملفات النصية .