Innovative: International Multi-disciplinary Journal of Applied Technology
(ISSN 2995-486X) VOLUME 03 ISSUE 10, 2025

A Unique Method for Concealing Information

Zaynalov N.R., Safarov R.A., Vafayev M.A., Shakarov A.A., Saidmurodov M.A. Sharipova U.B. Nishonova M.Q.

Samarkand branch of the Tashkent University of Information Technologies named after Muhammad al-Khwarizmi, Uzbekistan
nodirz@mail.ru

Abstract:

The subject of steganography is the means and methods that ensure the concealment of the fact of message transmission. The historical roots of this discipline are very deep. With the development of technology, the methods have also changed. If previously invisible ink was used between the lines of ordinary text, then modern digital steganography consists of embedding data in digital objects with minimal distortion. This is possible due to the use of the redundancy of the container data or the peculiarities of human perception, which is unable to notice these changes. In light of the widespread use of text documents in organizational activities, methods of hiding data in a text container, in particular, in MS Word documents, are becoming especially relevant. Based on this, an MS Word document is considered as an information carrier here. MS Word documents have different parameters, by changing these parameters or properties, it is possible to achieve data embedding. In this article, we present steganography using invisible Unicode characters of the Space type, but with different encoding. /2

Keywords: Information hiding • Steganography • Unicode • MS Word

Introduction

The development of information technology gave rise to modern steganography, which operates with digital data rather than physical media. This is due to the possibility of fully automating the processes of concealing and extracting messages. Thanks to this, it became possible to effectively

conduct experiments using computing technology and specialized algorithms for creating software applications. Protected data transmission and the organization of hidden communication channels have interested mankind since ancient times. Given the widespread use of text documents, text steganography is becoming a key tool for introducing secret information into an open text container. Such a message is transmitted openly, but the very fact of its presence remains invisible to third parties.

The peculiarity of text steganography is low data redundancy and strict linguistic rules, which greatly limits the possibilities of manipulating the text. This creates a complex but interesting task: to imperceptibly hide a message in the text and with the same skill to detect it.

A common feature of these methods and algorithms is that the hidden message is embedded in some harmless, unattractive object, which is transported to the addressee openly [1]. When using cryptography, the presence of an encrypted message itself attracts the attention of an intruder, in the case of steganography, the presence of hidden information remains unnoticed. The plain text, where the information will be hidden by the steganographic algorithm, is called a container.

Capacity, security and robustness, which are the three main factors affecting steganography, are in principle factors that contradict each other. Capacity is the relative amount of bits of secret information that can be hidden in a container. Security is the ability of an adversary to find out the hidden information. Robustness refers to the number of modifications that a stego-environment can withstand before an adversary destroys the hidden information [2]. A proper balance should be sought between the three aspects according to specific requirements.

A number of steganographic methods have been proposed in the last decade, but most of them use a covering medium such as images, video clips and sounds. Despite this, text documents are currently the most common and necessary form of information and are always used as a covering medium [3, 4].

From the review given in the works [3, 4] it can be concluded that most of the text steganography is based on the formats TXT, MS Word, PDF, PPT, etc. However, here is an attempt to improve the method of invisible characters between words with additional spaces for embedding data in an MS Word document. This article also considers the existing algorithmic approaches to steganography in MS Word documents in order to hide additional information in it. As you know, the popular Microsoft Word software is designed to enter and process texts in its own format. One of the reasons for its popularity is its simplicity and a large number of text formatting functions. For example, the font format has various properties, changes in which allow it to be successfully used in steganography. And this approach allows you to embed a large-capacity message that has a good degree of visual invisibility.

Texts are used in a wide range, as numerous text materials are transmitted daily over the global network. Analysis of text steganography methods indicates that the diversity of methods has not yet led to a high-quality text steganography technique that is stable and capacious. Unlike text steganography, which is relatively backward compared to the main methods of hiding that use images, audio and video as cover data, which is due to the lack of redundancy in the text [5, 6].

Despite this, storing text files requires less memory, and its easier compilation and exchange of them makes it preferable compared to other types of steganographic methods

This paper presents a method for hiding data using non-displayable attributes of characters from the Unicode table in MS Word.

This paper presents a new approach to text steganography by hiding a message in a set of Space characters of different Unicode codes, which we denote as UniSpace. This method works with the ASCII value of a character, not bits. The rest of the paper is organized as follows: Section 2 introduces the standard of universal character encoding, which is used to represent the entire character set of all alphabets. Section 3 describes some of the existing approaches to steganography in Word documents. Section 4 describes the proposed approach. Section 5 evaluates the results by comparing them with other methods. Section 6 concludes and discusses the advantages and disadvantages of the proposed steganography method.

II. Unicode Standard

Unicode is a universal character encoding standard used to support characters that are not in the ASCII set. Initially, all text editors were created based on the ASCII encoding, which contains the characters of the English alphabet and consists of only 128 characters. Unicode provides support for all the languages of the world and their unique character sets. In fact, Unicode can support more than 1 million characters. The reason is that Unicode can use more bit positions, which are units of information in computers, to represent a character. ASCII characters require only 7 bits, while Unicode can use 16 bits. This is necessary because some languages, such as Chinese, Arabic, require more bit positions.

At the same time, the Unicode table for characters of a language, such as the Arabic writing system, also includes languages such as Persian, Urdu, Pashto, Sindhi and Kurdish. The standard provides detailed explanations of the implementation methods, including the letter-joining method, right-to-left text insertion, and much more [7].

Table 1 shows the Unicode codes for spaces that interest us and will be used in the following sections, based on the work [8].

Table 1

UniSpace code notation in Unicode.

Code	Name	Code	Name
U+0020	Space	U+2005	Four-Per-Em Space
U+00A0	No-Break Space	U+2006	Six-Per-Em Space
U+1680	Ogham Space Mark	U+2007	Figure Space
U+180E	Mongolian Vowel Separator	U+2008	Punctuation Space
U+2000	En Quad	U+2009	Thin Space
U+2001	Em Quad	U+200A	Hair Space
U+2002	En Space	U+202F	Narrow No-Break Space
U+2003	Em Space	U+205F	Medium Mathematical Space
U+2004	Three-Per-Em Space	U+3000	Ideographic Space

III. EXISTING APPROACHES

In this section, we present some of the known approaches to text steganography in MS Word documents. The considered methods of text steganography are based on invisible characters or based on Unicode encoding, the implementation of which in various ways allows creating sequences of bits of a secret message. Studying scientific literature on this topic allows creating new directions in

methods of hiding information. At the same time, we will not focus on the strengths and weaknesses of these methods. One of the known methods is White Steg, which uses the standard space character Space to hide a secret message. In this case, the encoding of bits is carried out in an understandable way, For example, one space after a word represents bit 0, and two spaces after a word represent bit 1 [9].

The wbStego4open method also uses the space character, together with the null space, which has the code 0x00. In this case, the space between sentences and between words is used to embed the payload. To embed the secret message, the space character is replaced with the code value 0x00 to embed the bit 1 or with the code value 0x20 to embed the bit 0 [10].

A modification of this method is proposed in [11]. In the proposed algorithm, an additional null space will be added if the embedded bit is 1, otherwise the null space will remain unchanged.

But a unique application of the Unicode encoding is given in [12, 13, 14]. In these works, a method based on the Unicode table is proposed, where the composite form of some characters (i.e. the sign consists of two or more Unicode codes) in Unicode is used to hide the bits of the secret code. These characters, defined in Unicode, have both a single form and a composite form. By alternating these forms of writing letters, one bit of information can be represented. The use of this approach to conceal secret data can be seen in Chinese, Bengali, Arabic, and Persian text.

Certain modifications of these algorithms can be observed in other works as well. For example, in [15], the peculiarities of Arabic writing are used and a steganographic algorithm is presented also based on Unicode encoding. The algorithm proposed here is based on processing only connected letters, while the size and shape of the text remains unchanged.

The following articles provide an overview of various steganographic methods for Arabic text, where Arabic letters have many forms according to the Unicode standard [16]. In this method, we use different possible Unicode values of the same letter to hide bits, as explained in [17, 18, 19].

In [16], a steganographic algorithm method based on the peculiarities of Arabic text, taking into account Unicode encoding, is proposed. In this case, the main idea is to process isolated Arabic letters, which use individual letters as data hiding in Arabic texts written in Unicode format. And to simplify the complexity of the algorithm, it is proposed to consider only individual letters at the beginning and at the end of words, and not all isolated letters in words.

In the work [17] a method is proposed which was called UniSpaCh. This method is an improved version of the White Steg method discussed above. Here, additional Space characters from the Unicode encoding are proposed to be inserted between words. For example, such characters as Punctuation, Thin, En Quad, Em Quad, Hair in sentences between words. The advantage of these spaces over a regular space is that the width of these characters is too small. Therefore, more spaces can be introduced, which increases the amount of information that can be hidden in the container document. As an alternative to the text container, a study is conducted in the work [20] to hide bits in an MS Excel document. This work also proposes a steganographic method for effectively hiding information using the Unicode character encoding system. In this case, a unique fact is used here, namely, seven numbers (9, 8, 7, 3, 2, 1, 0) in the Unicode standard have the same form, but different codes in Arabic and Persian. As a result, by alternating these codes, it is possible to hide information in an MS Excel document.

For our research, the method called SEFT technique in the work [21] is useful. In this study, a new text steganography method is proposed, which takes into account the font types. This new method

depends on the similarity of the font types of the English language. It works by replacing the font with more similar fonts. The secret message was encoded and embedded in similar fonts in the capital letters of the cover document by combining different fonts, which are designated as F1, F2, F3. By combining these fonts, 27 characters can be encoded, which is enough for English text. The text steganography method proposed here can work in different cover documents of different font types.

In general, many algorithms are collected in the work [4], which provides a brief overview of scientific research in the field of steganography in MS Word documents. The formation of these methods are given in the works [22-24].

This study proposes hiding information between words by additionally embedding several invisible codes. And instead of the standard Space code, the combination of these invisible UniSpace codes will mean one letter of the Latin alphabet, in accordance with the proposed encoding.

IV. Proposed Approach

As was correctly noted in [19], Unicode-based steganography methods have common disadvantages, which can be characterized as follows:

- Some Unicode-based steganography methods provide high performance, but this requires radically changing the content of the carrier text, while the main idea in steganography is that the method should be statistically undetectable.

But it should be noted that the essence of all Unicode-based steganography methods automatically implies changes in the characters in the empty container text, based on its analogue from the Unicode code table. And this leads to the fact that each letter in the target word will hide data. But, at the same time, the grammatical form of the word or sentence changes, so we need an algorithm that does not spoil the form of words.

The method of changing the interval between words allows embedding a message in a binary format into the text by placing one or two spaces after each word in a sentence. However, this or similar methods have a small volume of embedding. Based on this, it is proposed to embed not binary data, but ASCII characters. The given technology is implemented using the following sequences of codes, which will be basic in this approach, are given in Table 2.

Table 2

Basic codes of spaces in the algorithm.

Пробел	**Unicode**
THIN SPACE	2009
HAIR SPACE	200A
ZERO WIDTH SPACE	200B

Thus, this study proposes a new method using characters that have a common characteristic within the Unicode encoding system (i.e., similar characters with different codes in the Unicode table) to embed a secret message in an MS Word document. In the proposed variant, it is possible to hide a secret message in a Word document using different variants of three basic space codes from Table 2.

Table 3 shows a scheme for comparing one-to-one correspondence of letters of the Latin alphabet (to save space, the word SPACE is omitted from the table).

Table 3

Encoding of Latin letters

Combination of spaces			**Symbol**
THIN	THIN	THIN	A

THIN	THIN	HAIR	B
THIN	THIN	ZERO WIDTH	C
THIN	HAIR	THIN	D
THIN	HAIR	HAIR	E
THIN	HAIR	ZERO WIDTH	F
THIN	ZERO WIDTH	THIN	G
THIN	ZERO WIDTH	HAIR	H
THIN	ZERO WIDTH	ZERO WIDTH	I
HAIR	THIN	THIN	J
HAIR	THIN	HAIR	K
HAIR	THIN	ZERO WIDTH	L
HAIR	HAIR	THIN	M
HAIR	HAIR	HAIR	N
HAIR	HAIR	ZERO WIDTH	O
HAIR	ZERO WIDTH	THIN	P
HAIR	ZERO WIDTH	HAIR	Q
HAIR	ZERO WIDTH	ZERO WIDTH	R
ZERO WIDTH	THIN	THIN	S
ZERO WIDTH	THIN	HAIR	T
ZERO WIDTH	THIN	ZERO WIDTH	U
ZERO WIDTH	HAIR	THIN	V
ZERO WIDTH	HAIR	HAIR	W
ZERO WIDTH	HAIR	ZERO WIDTH	X
ZERO WIDTH	ZERO WIDTH	THIN	Y
ZERO WIDTH	ZERO WIDTH	HAIR	Z
ZERO WIDTH	ZERO WIDTH	ZERO WIDTH	

The last combination of triple ZERO WIDTH can be used as the beginning and end of the hidden text. Table 4 shows the ternary data digitization system: THIN – 0, HAIR – 1, ZERO WIDTH – 2.

Table 4

Numerical encoding of letters of the Latin alphabet

Position			**Numerical value of the code**	**Symbol**
1	2	3		
0	0	0	0	A
0	0	1	1	B
0	0	2	2	C
0	1	0	3	D
0	1	1	4	E
0	1	2	5	F
0	2	0	6	G
0	2	1	7	H
0	2	2	8	I
1	0	0	9	J

1	0	1	10	K
1	0	2	11	L
1	1	0	12	M
1	1	1	13	N
1	1	2	14	O
1	2	0	15	P
1	2	1	16	Q
1	2	2	17	R
2	0	0	18	S
2	0	1	19	T
2	0	2	20	U
2	1	0	21	V
2	1	1	22	W
2	1	2	23	X
2	2	0	24	Y
2	2	1	25	Z
2	2	2	26	

For the convenience of defining a set of spaces by a symbol (and then by its code), we will create an array for 3 types of spaces MySpace(3), where the elements of the MySpace(i) array can take one of the values: THIN, HAIR, ZERO WIDTH. Next, by the numerical value of the code, we will define a set of spaces UniSpace. For example, in the following way, for clarity, we will take the letter 'N' as an example. According to the table given above (see Table 4), this letter has the code icode = 13. Then the set from the MySpace(i) array is defined as follows:

index3 = icode Mod 3 + 1

index2 = (icode \ 3) Mod 3 + 1

index1 = (icode \ 3) \ 3 + 1

Here we have taken into account the fact that in the UniSpace encoding table (see Table 4) we have used the ternary number system. So our UniSpace space set for the letter 'N' will be:

MySpace(index1) , MySpace(index2), MySpace(index3)

Let's consider the basic algorithm in general terms. The proposed concealment algorithm consists of six stages. At the first stage (Step 1), an empty text container is opened, which is prepared in advance and saved in a text file of the .doc or .docx type. At the second stage (Step 2), the hidden text is requested, consisting of a sequence of only Latin letters. At the third stage (Step 3), the starting point for the data injection is taken in the container-document and marked with code 26. At the fourth stage (Step 4), the capacity of the container is checked by the length of the message being inserted, although this may not be done, since text files are usually large. At the fifth stage (Step 5), we sequentially change the standard space characters based on the numeric encoding of the letter with UniSpace characters. And the last sixth stage (Step 6) puts a mark with code 26 at the end of the secret message in the Word document and the document file is saved and the process is completed. To implement this idea, the authors developed a software application in the VBA programming language, which is the basic one in MS Office applications.

To extract data, this process is actually repeated. Namely, first we find the numeric code 26 between words and then each space is analyzed by the value of the space code sequence from the Unicode table (see Table 4). The process will stop if the numeric code 26 is encountered.

V. Results Evaluation

The proposed method was implemented using software developed by the authors. In this case, various documents from the Microsoft Word series were used as a container. The built-in VBA programming language was chosen as the programming language. The choice of programming language is not a fundamental point.

We will demonstrate the program's operation using the following example [25]:

Humanity has always paid special attention to women and girls from time immemorial. In fact, in one of the greatest suras of Holy Quran - "Nisa", women are also glorified.

Fig. 1. Source text, empty container.

If we hide, for example, the word "Nazokat" in this text, then after executing the program we will get the following result:

Humanity has always paid special attention to women and girls from time immemorial. In fact, in one of the greatest suras of Holy Quran - "Nisa", women are also glorified.

Fig.2. Stegotext with the secret word "Nazokat".

To understand the program operation in Fig. 2, after the word "has" an alternation of three UniSpace symbols is additionally shown. Comparing Fig. 1 and 2, we can conclude that it is quite difficult to visually distinguish these two texts. In principle, this text is very difficult to distinguish from the original. Visually, the text of the stegacontainer is practically indistinguishable from the original, i.e. an unprepared reader will most likely not be able to determine the presence of hidden information in the text being read.

When reading the secret message back from this text, we get the word "NAZOKAT". Note that the response contains only uppercase letters, although the input contained both uppercase and lowercase letters. This is due to the fact that in Table 4, the letters of the Latin alphabet are encoded only as uppercase.

Thus, the proposed scheme and algorithm for embedding and reading a secret message in a text document MS Word works. In general, this method has no limitation on the volume of the embedded secret message. However, this algorithm, as well as many text steganography algorithms, has a weakness for changing the text format, which can make the text useless.

VI. Conclusion

Modern steganography deals with information in electronic form, not with physical objects. And therefore, due to the rapid development of digital technologies, steganography has received a strong impetus for development. The reason for this situation is the following circumstance: Embedding and extraction can be automated, since computers can effectively process data. Many studies conducted in this area are based on digital media, such as text, images, audio, video, etc. At the same time, many organizations prefer text documents, so many scientific studies are conducted based on word processors. In general, secret information can be hidden almost anywhere, and at the same time, some container objects are more suitable for hiding information than others. Here is a steganography scheme in an MS Word document based on embedding invisible characters such as Space from the Unicode code set. Since, in the text, the Space character has the highest frequency, and therefore it can be concluded that the amount of embedded information is limited only by the number of this character in the text. The proposed steganography algorithm includes both the embedding and extraction process. Each character of the embedded data is hidden in the cover file without any noticeable deterioration of the cover file itself. As noted, the observed average percentage power of the proposed approach is due to the large volume of the space character in the text. And also the fact that this approach works with ASCII values of characters, and not with their binary value.

Despite the fact that during the embedding in the file the cover is changed, the cover and the stego file exactly match.

The algorithm presented in this work will serve as a basis for further research related to the development of an effective algorithm for embedding a secret message in an MS Word document. This program has a diverse set of attributes that can be used in steganography. Including the attributes of the text itself, which are successfully used in MS Word and for which many scientists have studied the possibilities of hiding data [4].

Thus, steganography created in ancient times received a new impetus for development in connection with the advent of computer technology. Digital steganographic methods using the features of information presentation in computer files are a promising direction of practical science. These methods can be used in such applied areas as copyright protection, preventing counterfeiting of electronic documents, transmitting a secret message and many other applications. In conclusion, I would like to give the following idea: Steganographic messaging is probably more of an art than a conventional method. Therefore, further research is needed in the field of steganography taking into account text, shape, environment and other various attributes.

ACKNOWLEDGMENT

The authors would like to thank colleagues engaged in scientific activities in the field of steganography, in particular, the authors listed in this work. The complexity of studying the possibilities of hiding data in various containers requires the purity of experiments and their reproduction. And this requires a conscientious attitude to scientific work. Enormous successes in the field of steganography have been achieved thanks to such scientists. These scientists are honest before science and really implemented the algorithms for hiding data described by them, and published the results of their work in such an accessible form that it gives other novice researchers the opportunity to repeat their experiments or observations.

REFERENCES

[1] Gutub, A. and M. Fattani. A novel arabic text steganography method using letter points and extensions. Proceedings of the WASET International Conference on Computer, Information and Systems Science and Engineering, May 25-27, 2007, Vienna, Austria, 2007. pp: 28-31.

[2] Chen, B. and G.W. Womell. Quantization index modulation: A class of provably good methods for digital watermarking and information embedding. 2001, IEEE Trans. Inform. Theory, 47: 2001. pp.1423-1443.

[3] R. Bala Krishnan, Prasanth Kumar Thandra, M. Sai Baba. An overview of text steganography. 4th International Conference on Signal Processing, Communications and Networking (ICSCN -2017), March 16 - 18, 2017, Chennai, INDIA

[4] Zaynalov N.R., Narzullaev U.Kh., Muhamadiev A.N., Bekmurodov U.B., Mavlonov O.N. Features of using Invisible Signs in the Word Environment for Hiding Data. 2019. International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8, Issue-9S3, July 2019. pp.1377-1379.

[5] Liu, M., Y. Guo and L. Zhou. Text steganography based on online chat. Proceedings of the 5th International Conference on Intelligent Information Hiding and Multimedia Signal Processing, Sept. 1214, IEEE Xplore Press, Kyoto, 2009. pp: 807-810. DOI: 10.1109/IIH-MSP.2009.

[6] Moraldo, H., An Approach for text steganography based on Markov Chains. Proceedings of the 4th Workshop de Seguridad Informatica, (WSI' 12), 2012.pp: 26-39.

[7] [Online].The Unicode Standard, URL: http://www.unicode.org, last visited: [Last accessed on 16.4.2020] .

[8] Por LY, KosSheik Wong, and Kok Onn Chee, UniSpaChi a text-based data hiding method using unicode space characters. The Journal of Systems and Software.2012. pp. 1075-1082.

[9] Por LY, Delina B. Information hiding—a new approach in text steganography. In: 7th WSEAS international conference on applied computer and applied computational science. Hangzhou China, 2008. pp 689-695.

[10] Murphy, B., Syntactic information hiding in plaintext. Master's Thesis. Trinity College Dublin. 2001.

[11] P. Singh, R. Chaudhary, and A. Agarwal, "A Novel Approach of Text Steganography based on null spaces," IOSR Journal of Computer Engineering, vol. 3, no. 4, 2012. pp. 11-17.

[12] Xinmei, Meng, P., Ye, Y., Hang, L., Steganography in chinese text. In: 2010 International Conference on Computer Application and System Modeling (ICCASM 2010), vol. 8, 2010. pp. V8-651-V8-654.https://doi.org/10.1109/ICCASM. 2010.5620373

[13] Khairullah, M. Steganography in bengali unicode text. SUST J. Sci. Technol. 2018.

[14] Hassan, M. and M. Shirali-Shahreza, Steganography in persian and arabic unicode texts using pseudospace and pseudo connection characters. J. Theoretical Applied Inform. Technol., 4: 2008. pp.682-687.

[15] Obeidat A.A.Arabic text steganography using Unicode of non-joined to right side letters. Journal of Computer Science.13(6), 2017. pp. 184-191.

[16] Mohamed A.A. An improved algorithm for information hiding based on features of Arabic text: A Unicode approach. Egyptian Informatics Journal. 15(2), 2014. pp. 79-87.

[17] Por LY, Wong KokSheik, Chee Kok Onn. UniSpaCh: a text- based data hiding method using Unicode space characters. J Syst Softw 2012;85:1075-82.
[18] Shahreza MS, Shahreza SS. Persian/Arabic Unicode text steganography. In: The fourth international conference on information assurance and security. IEEE; 2008. p. 62-66.
[19] Shahreza MS, Shahreza MH. An improved version of Persian/ Arabic text steganography using ‘‘La” word. In: Proceedings of IEEE 6th national conference on telecommunication technologies; 2008. pp. 372-376..
[20] A. Salman Saber, W. Akeel Awadh. Steganography in MS Excel Document Using Unicode System Characteristics. Journal of Basrah Researches ((Sciences)) Vol.(39). No.(1) .A 2013. pp.10-19.
[21] Wesam Bhaya, Abdul Monem Rahma, Dhamyaa AL-Nasrawi. Text Steganography Based on Font Type in MS-Word Documents. Journal of Computer Science 9 (7): 2013.pp.898-904.
[22] Rabah, K. Steganography-the art of hiding data. Inform// 2004, Technol. J., 3: 245-269.
[23] Low, S.H.,N.F. Maxemchuk, J.T. Brassil and L. O'Gorman. Document marking and identification using both line and word shifting. Proceedings of the 14th Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM'95), April 2-6, 1995, IEEE Computer Society, Washington, DC. USA., pp: 853-860.
[24] Bender, W., D. Gruhl, N. Morimoto and A. Lu. Techniques for data hiding// 1996, IBM Syst. J., 35: 313-336.
[25] [Online]. : http://uza.uz/oz/culture/ayel-mu-addas-m-zhiza-yaratuvchi-15-04-2020. [Last accessed on 16.4.2020]