Combining Invisible Unicode Characters to Hide Information in a Text Document

N.R. Zaynalov#, U.Kh. Narzullaev#, A.N. Muhamadieva#, I.R. Rahmatullaev#, R.K. Buranova#

#Department of information security, Samarkand branch of Tashkent University of Information Technologies named after Muhammad al-Khwarizmi, Samarkand, 140100, Uzbekistan
E-mail: nodirz@mail.ru, nabi8888@bk.ru

Abstract—Steganography develops tools and methods for hiding the fact of message transmission. The first traces of steganographic methods are lost in ancient times. For example, there is a known method of hiding a written message: the slave's head was shaved, a message was written on the scalp, and after the hair grew back, the slave was sent to the addressee. From detective works, various methods of secret writing between the lines of ordinary text are well known: from milk to complex chemical reagents with subsequent processing. Digital steganography is based on hiding or embedding additional information in digital objects while causing some distortion of these objects. In this case, text, images, audio, video, network packets, and so on can be used as objects or containers. To embed a secret message, steganographic methods rely on redundant container information or properties that the human perception system cannot distinguish. Recently, there has been a lot of research in the field of hiding information in a text container, since many organizations widely use text documents. Based on this, here the MS Word document is considered as a medium of information. MS Word documents have different parameters, and by changing these parameters or properties, you can achieve data embedding. In the same article, we present steganography using invisible Unicode characters of the Space type, but with a different encoding.

Keywords—information hiding; steganography; unicode; text container; MS Word.

I. INTRODUCTION

The development of information and communication technologies has led to the emergence of modern steganography, which deals with information in electronic form, rather than with physical objects and texts. This is mainly because the process of hiding and retrieving a secret message can be automated. This makes it possible to effectively conduct experiments using computer technology and the appropriate algorithms needed to create software applications.

The secret transmission of information and the establishment of hidden relationships has been of interest since ancient times. Text documents are widely used in everyday practice. Steganography can be a vital means by which secret information is embedded in the fairing information that can be observed for transmission, so the information cannot simply be recognized by others. Text steganography has low redundancy and is associated with language rules, which leads to limited text manipulation, so it is a pleasant task to properly hide a message in the text and see such concealment.

Steganography is the science of hiding data inside a coverage object to preserve an invisible secret message without compromising the integrity of the coverage object, so other people cannot recognize the presence of a secret message. A common feature of these methods and algorithms is that the hidden message is embedded in some harmless, non-attracting object, which is transported to the recipient openly [1]. When using cryptography, the presence of an encrypted message itself attracts the attention of an attacker; in the case of steganography, the presence of hidden information remains unnoticeable. The plain text where information will be hidden by the steganographic algorithm is called a container.

Volume, safety, and reliability, which are the three main factors affecting steganography, are in principle factors that contradict each other. Volume is the relative number of bits of secret information that can be hidden in a container. Security is the ability to find out hidden information from the enemy. Reliability refers to the number of modifications a stego environment can withstand before an adversary destroys hidden information [2]. An appropriate balance should be sought between the three aspects following specific requirements.

Many 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 means to cover [3, 4].
From the review given in [3, 4], we can conclude that most text steganography is based on the formats TXT, MS Word, PDF, PPT, and so on. However, here we try to improve the method of invisible characters between words with additional spaces for embedding data in an MS Word document. This article also discusses the existing algorithmic approaches to steganography in MS Word documents to hide additional information in it.

As you know, the popular Microsoft Word software is designed for entering and processing texts in its 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 that allow you to successfully apply it in steganography. And this approach allows you to implement a high-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 variety of methods has not yet led to a qualitative method of text steganography, which is stable and capacious. In contrast to the main concealment methods that use images, text steganography methods indicates that the variety of other types of steganographic methods.

The wbStego4open method also uses a space character, asUniSpace. This method works with the ASCII character value, not bits. As you know, the popular Microsoft Word software is designed for entering and processing texts in its 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 that allow you to successfully apply it in steganography. And this approach allows you to implement a high-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 variety of methods has not yet led to a qualitative method of text steganography, which is stable and capacious. In contrast to the main concealment methods that use images, text steganography methods indicates that the variety of other types of steganographic methods.

III. MATERIAL AND METHOD

In this section, we present some of the well-known approaches to text steganography in MS Word documents. At the same time, the methods of text steganography considered are based on invisible characters or based on Unicode encoding, the implementation of which in various ways allows you to create sequences of bits of a secret message. The study of scientific literature on this topic allows you to create new directions in methods of hiding information. At the same time, we will not focus on the strengths and weaknesses of these methods.

One well-known method is White Steg, which uses the standard Space character to hide a secret message. At the same time, bit encoding is carried out understandably, for example, one space after the word represents bit 0, and two spaces after the word represent bit 1 [9].

The wbStego4open method also uses a space character, together with a null space, which has the code 0x00. At the same time, the space between sentences and between words is used for embedding the payload. To embed a secret message, the space character is replaced with the code value 0x00 for embedding bit 1 or the code value 0x20 for embedding 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 equal to 1, otherwise, the null space will remain unchanged.

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

Certain modifications of these algorithms can be observed in other works. For example, [15] uses features of Arabic writing and presents a steganographic algorithm also based on Unicode encoding. The algorithm proposed here is based on processing only related letters. however, the size and shape of the text remain unchanged.

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

In [16] we propose a method of steganographic algorithm based on the features of the Arabic text, taking into account the Unicode encoding. In this case, the main idea is to process isolated Arabic letters, which use individual letters as hiding data 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 end of words, and not all isolated letters in words.

In [17], a method called UniSpaCh is proposed. This method is an improved version of the White Steg method discussed above. Here, additional characters of the Space type, from Unicode encoding, are inserted between the words suggested. For example, characters such as Punctuation, Thin, En Quad, Em Quad, Hair in sentences between words. The advantage of these spaces over a normal space is that the width of these characters is too small. Therefore, more spaces can be entered, which increases the amount of information that can be hidden in the document container.

As an alternative to the text container in [20], a study is conducted to hide bits in an MS Excel document. This paper also proposes a steganographic method for effectively hiding information using the Unicode character encoding system. In this case, a unique fact is used, namely, seven numbers (9, 8, 7, 3, 2, 1, 0) in the Unicode standard, they have the same form, but different codes in Arabic and Persian. As a result, by alternating these codes, you can hide information in an MS Excel document.

The method called SEFT technique in [21] is useful for our research. This study proposes a new method of text steganography that takes font types into account. This new method depends on the similarity of font types in English. 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 accompanying document, combining different fonts, which are designated as F1, F2, F3, by Combining these fonts, you can encode 27 characters, which is enough for English text. The text steganography method proposed here can work in different accompanying documents of different font types.

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

This study suggests hiding information between words by further 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, under 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 changing characters in the text of an empty container, based on its analog from the Unicode code table. This will cause data to be hidden in each letter in the target word. However, the grammatical form of a word or sentence changes, so we need an algorithm that does not spoil the form of words.

The word-spacing method allows you to embed a message in the text that has a binary format by placing one or two spaces after each word in the text. However, these or similar methods have a small amount of embedding. Based on this, it is suggested to embed ASCII characters instead of binary data. This technology is implemented using the following sequence of codes, which will be the basis for this approach (see Table 2).

| Space          | Unicode |
|----------------|---------|
| THIN SPACE     | 2009    |
| HAIR SPACE     | 200A    |
| ZERO WIDTH SPACE| 200B    |

Thus, this study proposes a new method using characters that have a single character within the Unicode encoding system (i.e. similar characters with different codes in the Unicode table) for embedding a secret message in an MS Word document. In the proposed version, you can hide a secret message in a Word document using various variants consisting of three basic space codes from Table 2.

To compare the one-to-one correspondence of letters from the Latin alphabet, we will use the following scheme (to save space, we will skip the word SPACE in this table, see Table 3).

| 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      |
For the convenience of defining a set of spaces by the character (and then by its code), we will create an array for 3 types Of myspace spaces(3), where the elements of the MySpace(i) array can take one of the values: THIN, HAIR, ZERO WIDTH.

Next, we will use the numeric value of the code to define a set of UNISPACE spaces. For example, let's use the letter 'N' as an example for clarity. According to the table above (see Table 4), this letter has the code icode = 13. Then the set from the MySpace(i) array is defined as follows:

Table IV
NUMERIC ENCODING OF LATIN LETTERS

| Position | The numeric value of the code | Symbol |
|----------|------------------------------|--------|
| 1 2 3    | 0 0 0                        | A      |
|          | 0 0 1                        | B      |
|          | 0 1 0                        | C      |
|          | 0 1 1                        | D      |
|          | 0 2 0                        | E      |
|          | 0 2 1                        | F      |
|          | 0 2 2                        | G      |
|          | 0 2 3                        | H      |
|          | 1 0 0                        | I      |
|          | 1 0 1                        | J      |
|          | 1 0 2                        | K      |
|          | 1 1 0                        | L      |
|          | 1 1 1                        | M      |
|          | 1 1 2                        | N      |
|          | 1 2 0                        | O      |
|          | 1 2 1                        | P      |
|          | 1 2 2                        | Q      |
|          | 2 0 0                        | R      |
|          | 2 0 1                        | S      |
|          | 2 0 2                        | T      |
|          | 2 1 0                        | U      |
|          | 2 1 1                        | V      |
|          | 2 1 2                        | W      |
|          | 2 1 3                        | X      |
|          | 2 2 0                        | Y      |
|          | 2 2 1                        | Z      |
|          | 2 2 2                        |        |

The last combination of the triple ZERO WIDTH can be used as the beginning and end of the hidden text. To digitize this data, we apply a ternary number system to the data in Table 3, namely (we denote THIN-0, HAIR-1, ZERO WIDTH-2) (see Table 4).

Here we took into account the fact that in the UniSpace encoding table (see Table 4) we used the ternary number system. So our Set of UNISPACE spaces for the letter 'N' will be:

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

Let's look at the main algorithm in General terms. The proposed concealment algorithm consists of six stages. In the first step (Step 1), an empty text container is opened, which is pre-prepared and saved in a text file of the type .doc or .docx. In the second stage (Step 2), a hidden text consisting of a sequence of Latin letters only is requested. In the third stage (Step 3), the document container takes the starting point for data insertion and is marked with the code 26. In the fourth stage (Step 4), the container capacity is checked by the length of the embedded message, although this may not be necessary since text files are usually large. At the fifth stage (Step 5), we consistently 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 label with code 26 at the end of the secret message in the Word document and the document file is saved and the process ends.

To implement this idea, the authors developed a software application in the VBA programming language, which is basic in MS Office applications.

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

V. RESULT AND DISCUSSION

The proposed method was implemented using software developed by the authors. At the same time, 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 the programming language is not a matter of principle.

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

**Fig 1. Source text, empty container**

If you hide the word “Nazokat” in this text, for example, we will get the following result after executing the program:
To understand how the program works, in figure 2, after the word “has”, an additional three UniSpace characters are shown alternating. Comparing figures 1 and 2, we can conclude that these two texts are quite difficult to distinguish visually. In principle, this text is very difficult to distinguish from the original. Visually, the text of the stegecontainer does not differ from the original, i.e. an untrained reader will most likely not be able to detect the presence of hidden information in the text being read.

When we reverse read the secret message from this text, we get the word “NAZOKAT”. Please note that the response contains only uppercase letters, although the input was both uppercase and lowercase. This is since the Table 4 letters of the Latin alphabet are encoded only as uppercase.

Thus, the proposed scheme and algorithm for implementing and reading a secret message in a text document MS Word works. In General, this method does not have a limit on the volume of a secret message being implemented. 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 so, 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 extracting can be automated since computers can process data efficiently. Much of the research done in this area is based on digital media such as text, image, audio, video, etc. However, many organizations prefer text documents, so a lot of research has been done based on word processors. In General, secret information can be hidden almost anywhere, and 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 Space characters from a set of Unicode codes. Since the Space symbol has the highest frequency in the text, we can conclude that the amount of embedded information is limited only by the number of this symbol in the text. The proposed steganography algorithm includes both the embedding and extraction process. In this case, each character of embedded data is hidden in the cover file without any noticeable degradation of the cover file itself. As noted, the observed average percentage power of the proposed approach is due to the large space character in the text. And also that this approach works with ASCII character values, not with their binary value.

Although changes are made to the cover file during embedding, the cover and the stege file are the same.

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

Thus, steganography created in ancient times received a new impetus for development due to the advent of computer technology. Digital steganographic methods that use the features of information representation in computer files is a promising area of practical science. These methods can be applied in applications such as copyright protection, electronic document forgery prevention, secret message transmission, 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 the text, form, 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 paper. The complexity of learning how to hide data in different containers requires the purity of experiments and their reproduction. And this requires a conscientious attitude to scientific work. Great progress in the field of steganography has been achieved thanks to such scientists. These scientists are honest to science and have performed the data hiding algorithms described by them, and published the results of their work in such an accessible form that allows other novice researchers 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] Zaynulov N.R., Narzallayev U.Kh., Muhamediev 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/IH- 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.
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.

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.

Murphy, B., Syntactic information hiding in plaintext. Master’s Thesis. Trinity College Dublin. 2001.

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.

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

Kharullah, M. Steganography in bengali unicode text. SUST J. Sci. Technol. 2018.

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.

[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.

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.

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.

Shahreza MS, Shahreza SS. Persian/Arabic Unicode text steganography. In: The fourth international conference on information assurance and security. IEEE; 2008. p. 62-66.

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.

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.

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.

Rahab, K. Steganography-the art of hiding data. Inform/ 2004, Technol. J., 3: 245-269.

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.

Bender, W., D. Gruhl, N. Morimoto and A. Lu. Techniques for data hiding// 1996, IBM Syst. J., 35: 313-336.

[Online] : http://uza.uz/oz/culture/ayel-mu-addas-m-zhiza-yaratuvchi-15-04-2020. [Last accessed on 16.4.2020]