2022 2nd International Conference on Computing and Information Technology (ICCIT)
Jan. 25 - 27, 2022/ FCIT/UT/KSA

A Novel Steganography Approach to Embed Secret Information into a Legitimate URL

Kholood Ayed Almalki
College of Computer Science and Engineering
Jeddah University
Jeddah, Saudi Arabia
kalmalki0127.stu@uj.edu.sa

Roshayu Mohammed
College of Computer Science and Engineering
Jeddah University
Jeddah, Saudi Arabia
romohamad@uj.edu.sa

2022 2nd International Conference on Computing and Information Technology (ICCIT) | 978-1-6654-3605-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICCIT52419.2022.9711647

***Abstract*— Hiding confidential digital information in today's modern communication system is a very challenging task. Even though many algorithms have been developed to ensure the security of information delivered via the communication channel, many flaws have been discovered over time. Steganography is the science of hiding the existence of secret information. The concept of steganography is about embedding the secret data in a cover file in which the cover file could be an image, audio, video, text, or any other medium. Successfully formulating an algorithm that embeds the secret information in a cover file will produce a new file called the stego file. This research aims to develop a new methodology in which the secret information will be embedded into a legitimate URL link. Unlike the usual methods that hide the secret message in the content of another file, this method uses a valid form of the URL link to hide the secret message. Since the URL is a type of text that links users to a web page and is not an informative text to be read, it results in a secure method to pass secret information as no one pays attention to a URL text. This method combines and implements different techniques, abbreviates, searches, and matches bits and converts to allowed URL characters to encode the secret message. The results show that it works well with legitimate URLs, including the website's homepage URL.**

***Keywords*— Steganography, Information Hiding, Text Steganography, URL Steganography.**

I. INTRODUCTION

Steganography means concealing the presence of a secret piece of information message by embedding it into another innocent, harmless type of medium like graphics or audio. The word steganography comprises two Greek words, *steganos*, the secret or the covering, and *graphy*, which means drawing or writing. From ancient times steganography was used as a means to hide secret messages. Histaeus, the Greek ruler, used steganography by shaving the slave's head, tattooing the secret message, waiting for the hair to grow back, and then sending the slave out to deliver the message to the intended receiver. The receiver of the message shaves the slave's head, reads the message, and if he has a message of his own to deliver, he tattoos the slave's head with the secret message and waits for the slave's hair to grow back, and sends it back to the sender. Steganography has also played a significant role during wars, from using invisible ink by British and Americans to using microdots by Germans [1].

To clearly understand the concept of steganography, we must understand cryptography and know the difference between these two concepts of information hiding. Cryptography is a means to hide the meaning of the secret message, not the message's presence like steganography. Moreover, cryptography requires a key to encrypt and decrypt the message [2]. In steganography, multiple medium types are used as covers to hide secret messages, with each type having different techniques to implement. Fig. 1 lists the most commonly used and well-known multimedia methods in steganography.

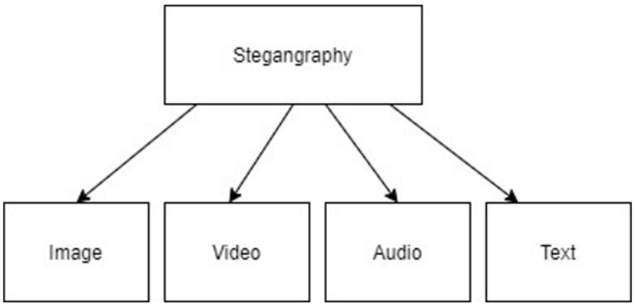


Fig 1. Commonly Known Steganographic Categories

Using images to hide secret information is called Image Steganography. Image file format such as JPEG or bitmap is employed to be altered to hide secret information in it. Common image steganography techniques that alter images to hide secret information are: Spatial domain, Transform domain, Spread spectrum, Statistical methods, and Distortion techniques [3]. Audio Steganography, is done by altering the digital representation of 0s and 1s of the audio file. The secret information is hidden inside the audio signals [4]. This steganography type uses audio formats such as MP3 and WAV to hide secret information. Low Bit Encoding, Phase Coding, and Spread Spectrum are different methods of audio steganography. In Video Steganography, the secret information is hidden in each image of the cover video since the digital video is a sequence of images [4]. The DCT (discrete cosine transform) is usually used to hide the secret information in each image of the cover video. The format files such as MP4 and MPEG are used to hide the secret information. Text Steganography is hiding a secret message into a text by altering the formatting of the text or the text characteristics [5]. A very ancient way to hide a secret message is by hiding it in every *n*th letter of the cover text. Methods that are frequently used in Text Steganography are Format Based Method, Random and Sequence Method, and Linguistic Method. Text files as covers are not preferable to be used since it has a minimal amount of redundant data.

A. Objectives

In order to safely send and receive secret messages over the internet without drawing the attention of third parties, the following objectives shall be accomplished by this research work:

- To develop and embed the secret message inside the URL of the sender's end and link it to a legitimate website.
- To explore the method of how the authorized receiver will extract the secret message on the receiving end.
- To formulate the limit length of the created URL in order not to instill suspicion in other viewers.

B. *Motivation*

This paper tries to achieve two motives. First, since limited researches in steganography uses URLs as covers to hide text messages were found, this research proposes exploring the method that can be used to hide the secret message in the URL without raising any suspicions considering the legitimate length of the URL. Second, this research intend to formulate an efficient steganography algorithm technique. Three criteria should be measured-capacity, robustness, and imperceptibility. Capacity refers to the number of secret information bits the steganographic technique can embed into the cover file without distorting it. The capability of a technique to withstand attacks is referred to as robustness, and imperceptibility measures the similarity between the cover file and the stego file, the encoding process's output. A trade-off between these three criteria will produce an effective and efficient algorithm [6].

II. Literature Review

A. *URL Steganography*

Desoky [7] proposes a novel algorithm that conceals a secret message and its transmission over the internet using legitimate websites' URLs. The algorithm does not hide any secret information in the text body of the webpage. The algorithm converts the ASCII values of the secret message to binary format, then slicing the binary values into parts of six bits each. Then it converts each of the six bits to its corresponding letters, numbers, symbols, and alphanumeric characters allowed in URL Table I. The previous step's result attached to the URL end, and then the URL shortened using any shortening tool. The URL is then sent via public channels such as email and chat to make the transmission legitimate and not raise any suspicion.

TABLE I. CHARACTERS ALLOWED IN URL

A	B	C	D	E	F	G	H	I	J	K	L	M
N	O	P	Q	R	S	T	U	V	W	X	Y	Z
A	b	c	d	E	f	g	h	i	j	k	l	m
N	o	p	q	R	s	t	u	v	w	x	y	z
0	1	2	3	4	5	6	7	8	9	!	*	,
(	)	;	:	@	&	=	+	$	'	/	?	#
[	]	-	_	.	~							

Satir et al. [8] proposes a high capacity steganographic method where secret data is compressed via LZW coding, a method of compression where it compresses a file into a smaller file using a table-based lookup algorithm and hides it in an HTML code of a selected webpage with the best capacity rate from a collection of webpages. The equation that calculates the capacity rate of a webpage uses the following formula:

$$Capacity = \frac{bitsOfCompressedSecretMessage}{bitsOfURL} \quad (1)$$

After selecting the web page, the content is compressed again via LZW coding to enhance the capacity. RSA encryption algorithm will be applied to the content to improve the security and resiliency against possible attacks. The encrypted content will then be embedded to the URL of the corresponding webpage to be used as a steganographic cover, and the URL directs the recipient correctly to the webpage. The sender sends only the URL link of the chosen webpage that contains the secret message, where the receiver of the URL link applies the opposite steps to extract and read the secret message. Unfortunately, there is an unbalance between the URL cover's capacity and imperceptibility because the longer the secret message, the longer the URL will be. This situation would raise suspicion, and the attacker would reveal the secret message.

In [9] they use an approach that detects and extracts all types of hidden URLs in all kinds of images using the least significant bit hiding technique. There are two groups of embedding techniques in image steganography, spatial domain techniques and transform domain techniques. Their study focuses on the spatial domain, which directly embeds the secret message in the image pixels, usually in the least significant bit of the image. To the human eye, these modifications are imperceptible. In the work, the reason behind embedding URLs in an image instead of embedding the actual secret data is that URLs occupy less space in the cover image. This method will help in hiding the URL and prevent corrupting the image. The proposed work in [10], is an enhancement of the previous work. They increase the security of the method by encrypting the URL using the binary operation NOT as an encryption method before embedding it to the cover image, and then extracting and decrypting the URL before revealing the actual URL to the receiver.

B. *Text Steganography*

Hiding secret information of any kind in a text is considered challenging since the amount of redundant data in text files are very small, making texts rarely used as covers compared with other multimedia files [5]. Elmahi et al. [11] employ both cryptography and text-steganography techniques to conceal and hide information at different cover text positions randomly. The algorithm calculates the number of the secret message's characters and then extracts unique letters from the original message. For example, suppose the original message is “Hello World”, the extracted letters are “Helo wrd”. The algorithm generates cover-text randomly from the extracted letters using PRG. The sequence bits of 0s and 1s mix the cover text and the secret text to generate the stego text. Then the stego text will be compressed using the lossless Huffman compression algorithm. If the same message is sent again, the model will produce different cover text and the stego text file's content. The algorithm proposed by Iyer and Lakhtaria [12] based their work on the theory of "nonreplacement," which means the algorithm produces no changes in the cover text. It uses a pairing algorithm that pairs the secret message's alphabets into four parts with two bits each. In the original message, the last word is purposely added, containing the secret message's location. The secret message will be revealed by decoding the location in the last word of the received stego text.

The work in POR and Delina [13] proposes a new approach in text steganography using a hybrid method of inter-word and inter-paragraph spacing. It utilizes the spaces between words and paragraphs in the right justification of text. Combining both algorithms provides a large capacity for embedding bits. Since each 8-bits character requires eight spaces to encode one character- this is plenty of spaces. POR's and Delina's method overcome this drawback by compressing the text message's character to be 3-bits or less instead of 8-bits. The benefit of using white spaces is that white spaces appear in a document more than words. No one would think that there is a secret message hidden in these spaces, and the cover text is dynamically created according to the length of the secret message.

Agarwal [14] presents three novel approaches in text steganography: Missing Letters Puzzle, Hiding Data in Wordlist, and Hiding Data in Paragraph. The first approach, Missing Letters Puzzle, is where each word in a collection of words has one or more letters missing and replaced with a question mark. The approach is solvable if the question marks are replaced by an appropriate letter that makes the word meaningful. The second approach, Hiding Data in Wordlist, overcomes the first approach's issues, using a list of words to conceal a message without using any additional special characters. The cover of the previous two approaches is dynamically generated, and the stego file is a list of words. In the third approach called Hiding Data in Paragraph, the cover file uses any meaningful English text. It can be a text from any source, such as a text from a book or a newspaper, and this approach uses the first and last letters of words in the cover file to hide the message.

The proposed method by [15] enhances the paragraph approach proposed by [14]. The enhancement done is regard to the capacity ratio of the cover text. The capacity ratio goes from 2\% as calculated in [14] to 4\%. The algorithm takes multiple characters of a single word instead of taking the first and last characters to improve this algorithm's capacity ratio. Zero distortion in the cover text means no changes made to the cover file. [16] proposes an algorithm that employs the zero-distortion technique by searching for matching bits between the secret message and the cover text and stores the matched bits' locations in a matrix of locations. The matrix is then encrypted using the Indexed Based Chaotic Sequence. The secret message reconstructed from the matrix contains the locations of the actual information. The algorithm applies text steganography abbreviation method before applying the zero-distortion algorithm. It replaces a word with its abbreviated acronym, slang, or a letter with the same pronunciation stored in a database. These slangs are frequently used by social media users such as Twitter and WhatsApp, (refer Table II). Therefore, the proposed algorithm can store a large amount of data.

TABLE II. EXAMPLES OF WORDS AND ITS COMMONLY USED ABBREVIATIONS FORM

Words	Abbreviation
See	C
To	2
And	&

III. METHODOLOGY

This research adopts design science research principles for computer science research. Design science research involves the creation and evaluation of artifacts in order to solve the intended problems [17] - in the context of this research, it proposes a solution to send a secret message safely via URL link using the steganography method. Research using the 'building artifact' methodology is where a software system model is developed as the research instrument. Several methodologies can be used for computing science research to tackle questions within the discipline. This research adopts the build methodology involving designing the software as an instrument of the research to implement the proposed idea and moves to conducting the system testing and further evaluate the performance of the systems.

In order to address the research objectives specified in the introduction of the paper, a program is developed to exchanging encoded secret messages using legitimate URLs as covers. The software program used to develop this application is divided into modules. Each module performs one task. The encoding algorithm module accepts one URL and the secret message and produces one result: the URL cover. The decoding algorithm module accepts only one value, the URL cover, and produces a plaintext of the secret message—each module connected with one class containing the functions that will accomplish the task. The centralized database contains one table, *Abbreviations*. A centralized database allows the sender and receiver to use the same exact data for the encoding and decoding processes. Moreover, if a new word-abbreviation row is added to the table, the table should be up to date for both sides. Several options to implement the database are available online. However, for simplicity, a centralized localhost database is built to test the proposed method.

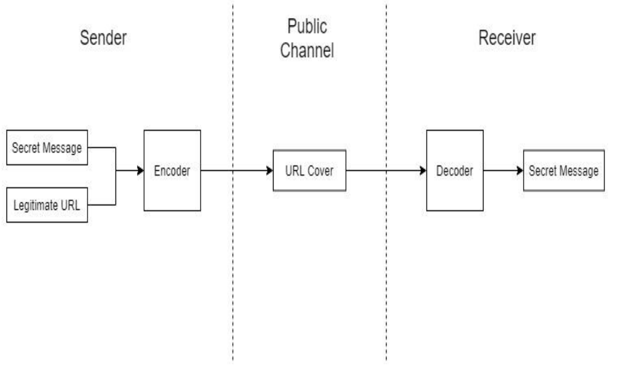


Fig. 1. *URL Steganography Scheme*

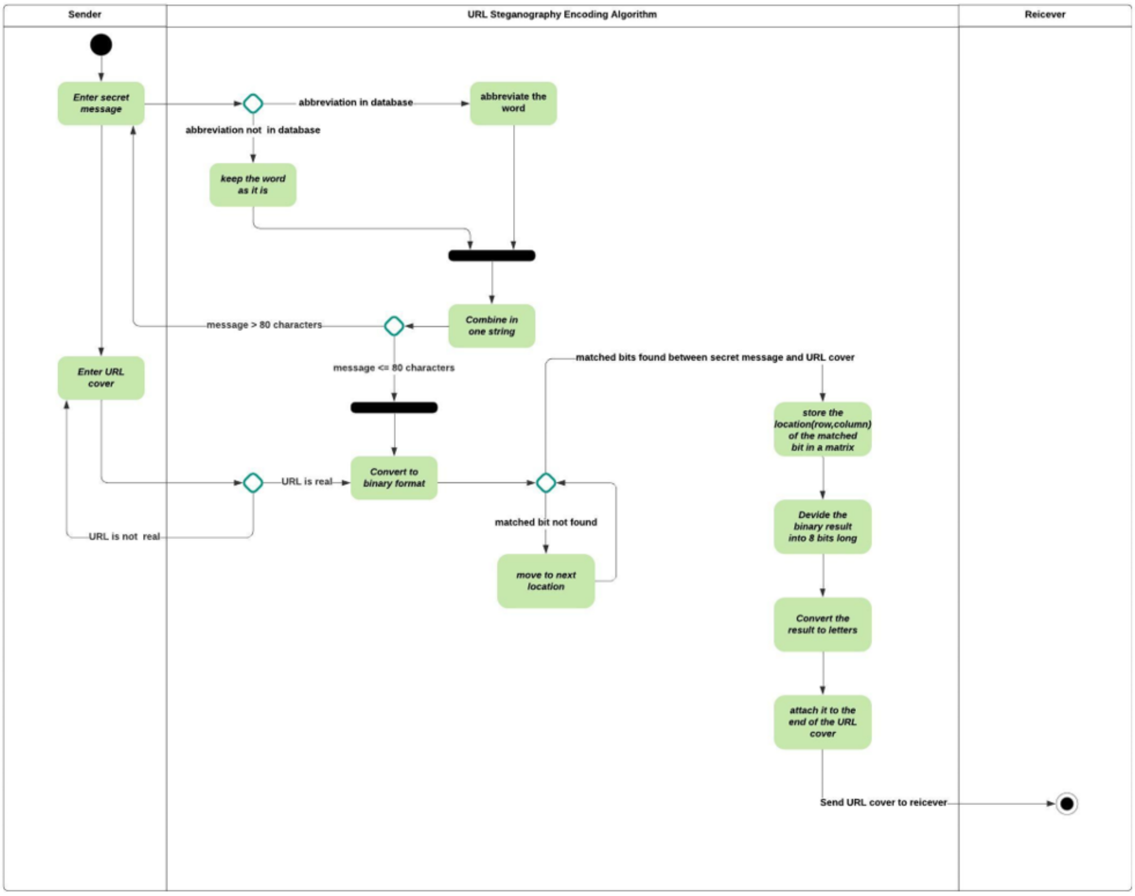

Sender
URL Steganography Encoding Algorithm
Reicever
Enter secret message
abbreviation in database
abbreviate the word
abbreviation not in database
keep the word as it is
Combine in one string
message > 80 characters
message <= 80 characters
Enter URL cover
matched bits found between secret message and URL cover
store the location(row,column) of the matched bit in a matrix
URL is real
Convert to binary format
URL is not real
matched bit not found
move to next location
Devide the binary result into 8 bits long
Convert the result to letters
attach it to the end of the URL cover
Send URL cover to reicever

Fig 3. Activity Diagram of the Encoding Algorithm

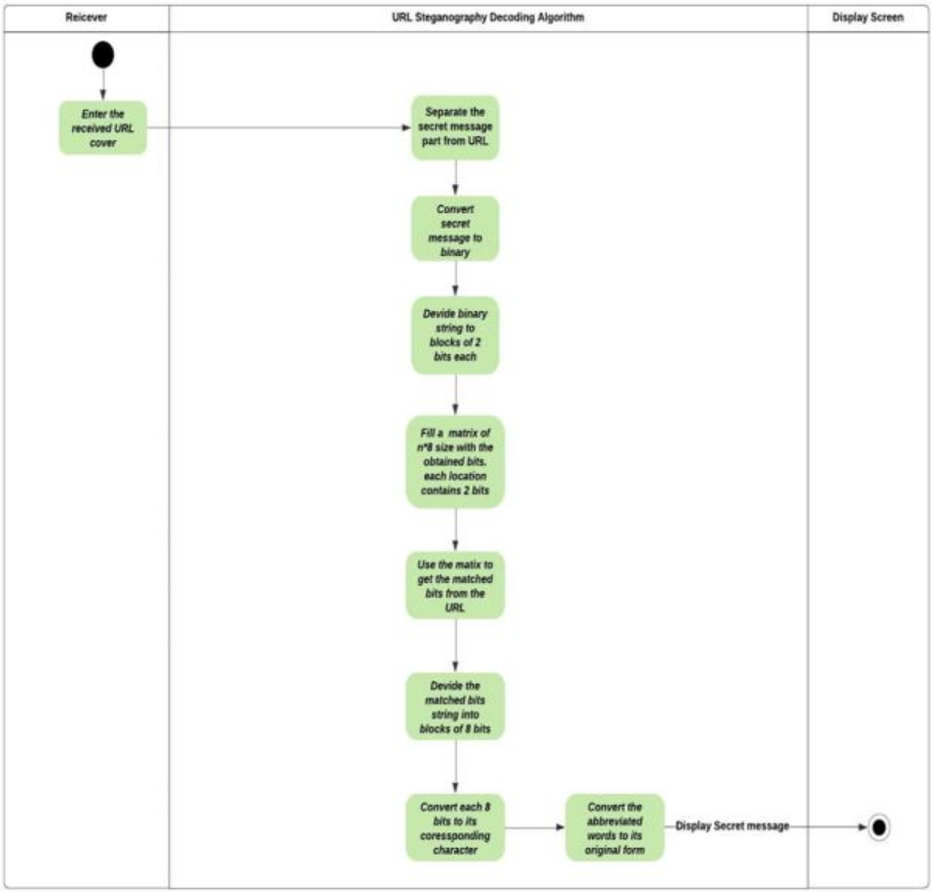

Reicever
URL Steganography Decoding Algorithm
Display Screen
Enter the received URL cover
Separate the secret message part from URL
Convert secret message to binary
Devide binary string to blocks of 2 bits each
Fill a matrix of n*8 size with the obtained bits. each location contains 2 bits
Use the matix to get the matched bits from the URL
Devide the matched bits string into blocks of 8 bits
Convert each 8 bits to its coressponding character
Convert the abbreviated words to its original form
Display Secret message

Fig 4. Activity Diagram of the Decoding Algorithm

Fig. 2 illustrates the URL steganography scheme. The sender chooses the secret message and the URL cover. The encoder algorithm attaches the encoded secret message to the URL, which results in the URL cover. The sender sends out the URL cover to the receiver side through any public channels such as emails. The recipient receives the URL cover on the receiver side, and the decoder algorithm extracts the secret message from it.

A. Encode Secret Message into URL

Fig. 3 is the activity diagram of the encoding algorithm. It goes through the following steps shown in order to achieve the desired results.

B. Decode Secret Message from URL

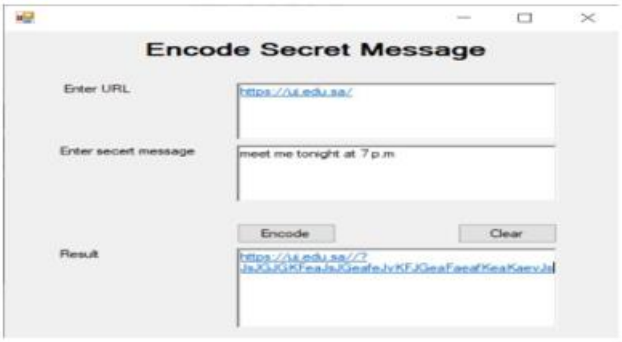
Fig. 4 is the activity diagram of the encoding algorithm. The steps used to reveal the secret message is shown in the diagram.

IV. Results

This method is developed with Visual Studio using C# 2019. The algorithm is divided into two parts. On the sender side, the encoding algorithm is intended to encode the secret message. On the receiver side, the decoding algorithm is intended to reveal the secret message. On the sender side, the result is the camouflaged secret message attached to the end of the URL cover (Fig. 5), and it is ready to send to the authorized receiver through public channels such as social media and emails. On the receiver side, the final result is the secret message's full form in plaintext (Fig. 6). Thus, this study satisfies the objectives mentioned in the introduction of this paper. The experimental results showed that the encoded secret message is twice the length of the abbreviated secret message. The abbreviated secret message length depends mainly on how many words in the original secret message that can be abbreviated. Table III and Table IV compare the proposed method with similar study by Desoky [7] and Satir et al. [8] regarding the produced URL length and whether it succeeded in opening the intended webpage. Both studies failed to open the corresponding URL webpage. In contrast, the proposed method succeeded in opening the webpage using the same secret messages and URL covers as in their works, as shown in Tables III and IV. Moreover, the method proposed by Satir et al. [8] produces a very long URL for such a short secret message of 10 characters concealed in the HTML body, which may raise suspicions.

TABLE III. COMPARISON BETWEEN THE PROPOSED METHOD AND Desoky[7]

	Proposed Method	**Desoky [7]**
URL cover	http://desoky.com/	http://desoky.com/
Secret message	our meeting 8pm	our meeting 8pm
Result	http://desoky.com//?JwKGKeeaFsKFJKeaflKaJs	http://desoky.com/b3VyIg1lZXRpbmcgOHBt
Open the webpage	Succeed	Failed, it gives "This page doesn't seem to exist" error.



Enter URL: https://ut.edu.sa/

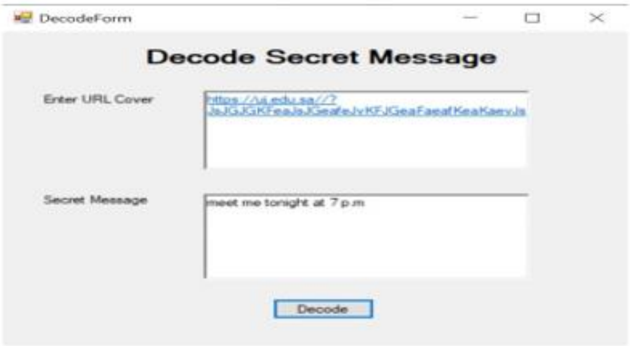
Enter secert message: meet me tonight at 7 p.m

Result: https://ut.edu.sa//?JsJGJGKFeaJsJGeafeJvKFJGeaFaeafKeaKaevJs

Fig 6. Encoding Result

TABLE IV. COMPARISON BETWEEN THE PROPOSED METHOD AND Satir et al.[8]

	Proposed Method	**Satir et al. [8]**
URL cover	http://www.selcuk.edu.tr/	http://www.selcuk.edu.tr/
Secret message	Lorem ipsu	Lorem ipsu
Result	http://www.selcuk.edu.tr?JrJwKeJGJseaJmKaKfKG	http://www.selcuk.edu.tr/index.php?id=&NzYyLzExMDYvMjEvMzYvMzYvMzYvMjEvMzYvMzYvMTc2OS8yMS8zNi8zOC8zNi8yMS8zNi8zOC8xNDUvMjEvMzYvMzgvNjk2LzIxLzM2LzM2Lzc2NC8yS8zNi8zNi82OTYvMjEvMYvMzYvNzYyLzIxLzM2LzM4LzE0NS8yMS8=
Open the webpage	Succeed	Failed, it gives "The resource cannot be found" error.



Enter URL Cover: https://ut.edu.sa//?JsJGJGKFeaJsJGeafeJvKFJGeaFaeafKeaKaevJs

Secret Message: meet me tonight at 7 p.m

Fig 7. Decoding Result

A. Efficinecy Criteria

Capacity, robustness, and imperceptibility are the common evaluation criteria that measure the strength and efficiency of the developed text steganography methods. Even though every steganographic method is relatively prone to these drawbacks, a balance between the three criteria should be considered. The capacity of the proposed method is five times the capacity of the similar studies' capacity rate. The large capacity of the proposed method is due to using two locations to form the matrix of locations. Since the encoding algorithm limits the abbreviated secret messages to accept 80 characters, the actual capacity rate ranges between less than or equal 3.846% and 7.692%. The main issue with texts that will affect the imperceptibility is that texts have less redundant data than other media which means long-secret message entered yields long encoded message, resulting in extended URL cover, which could endanger the imperceptibility. Since the encoded secret message is limited and attached to a URL, not concealed into the URL, and the URL structure does not obey linguistic rules, in most cases, the URL cover does not grab the attention of internet users. The proposed method is robust to the following attacks:

1) *Visual Attack*

Since URLs contain meaningless characters. Most users are more interested in the web site's content whether it looks normal or not rather than the URL. The proposed method of URL cover obeys the URI rules [18] therefore, visual attack in the proposed method is worthless.

2) *Statistical Attack*

The behavior of the URL cover of the proposed method is normal because it does not contain strange characters and is not too long. The main issue of the proposed method is that if the sender uses the same message with a different URL domain, it may produce the same encoded message. However, even if the adversary suspects unusual behavior, he cannot figure out the secret message since the encoding process is not straightforward or easily configured.

3) *Comparison Attack*

The adversary looks for modification and alteration between the original text cover and the text cover after embedding the secret message. However, the proposed method is not subjected to such an attack since a URL cover is not a text document that can be altered.

V. Conclusion

This paper proposed a novel method in text steganography that employs URLs of legitimate websites as a cover to carry encoded secret text messages through public channels such as social media applications and email. First, the secret message is abbreviated using a database of abbreviations to reduce its size and then converted to binary. Next, the binary string of the abbreviated message is compared bit by bit with the binary string of the URL, searching for a match. If there is a match, the location of the matched bit found in the URL binary string is stored in a matrix of location. Next, the binary content of the matrix of locations is divided into blocks of 8-bits. Finally, the blocks are converted to letters and attached to the end of the URL. In conclusion, compared to previous studies, the encoding algorithm results indicate that the algorithm works for every URL used, including the home page URL of the website. Furthermore, it produces a proper URL length since it limits the length of the abbreviated secret message entered by the sender.

References

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

[2] N. F. Johnson and S. Jajodia, "Exploring steganography: Seeing the unseen," Computer, vol. 2, no. 31, pp. 26-34, 1998.

[3] B. G. Banik and S. K. Bandyopadhyay, "Review on Steganography in Digital Media," International Journal of Science and Research (IJSR), vol. 4, no. 2, pp. 265-274, 2015.

[4] J. Kour and D. Verma, "Steganography Techniques –A Review Paper," International Journal of Emerging Research in Management &Technology, vol. 3, no. 5, pp. 132-135, 2014.

[5] J. Abraham and R. Gundla, "Survey on The Different Hiding Types and Techniques in Steganography," Journal of The Gujarat Reasearch Society, vol. 21, no. 14, pp. 174-180, 2019.

[6] M. Zamani, A. Abdul Manaf and R. Daruis, "Azizah Technique for Efficiency Measurement in Steganography," in Proc. 8th International Conference on Information Science and Digital Content Technology (ICIDT2012)., 2012.

[7] A. Desoky, "Uniform Resource Locator Based Steganography Methodology," International Journal of Network Security, vol. 20, no. 6, pp. 1005-1015, 2018.

[8] E. Satir, A. Sargin, T. Kanat and C. Okuducu, "A High-Capacity Html Steganography Method," in Indalam Proceedings of 7th International Conference on Information Security and Cryptology., Istanbul, 2014.

[9] M. M. Aljamea, C. S. Iliopoulos and M. Samiruzzaman, "Detection of URL In Image Steganography," in Proceedings of the International Conference on Internet of things and Cloud Computing., pp. 1-6, 2016.

[10] M. Aljamea, T. Athar, C. S. Iliopoulos and M. Samiruzzaman, "Detection of Hidden Encrypted URL in Image Steganography," in The Ninth International Conferences on Pervasive Patterns and Applications, 2017.

[11] M. Y. Elmahi, T. M. wahbi and M. H. Sayed, "Text Steganography Using Compression and Random Number Generators," International Journal of Computer Applications Technology and Research, vol. 6, no. 6, pp. 259-263, 2017.

[12] S. Iyer and K. Lakhtaria, "New robust and secure alphabet pairing Text Steganography Algorithm," International Journal of Current Trends in Engineering & Research (IJCTER), vol. 2, no. 7, p. 15 – 21, 2016.

[13] L. Y. POR and B. Delina, "Information Hiding: A New Approach in Text Steganography," in 7th WSEAS Int. Conf. on APPLIED COMPUTER & APPLIED COMPUTATIONAL SCIENCE (ACACOS '08), Hangzhou, China, 2008.

[14] M. Agarwal, "Text Steganographic Approaches: A Comparison," International Journal of Network Security & Its Applications (IJNSA), vol. 5, no. 1, pp. 91-106, 2013.

[15] F. X. K. akotoye, Y. E. Yakavor and J. Kwofie, "Character Pair Text Steganography Based on The Enhanced Paragraph Approach," in Proc. IEEE 7th International Conference on Adaptive Science & Technology (ICAST)., 2018.

[16] S. V. K. Yadav and S. Batham, "A Novel Approach of Bulk Data Hiding using Text Steganography," in 3rd International Conference on Recent Trends in Computing 2015 (ICRTC-2015), 2015.

[17] F. Gacenga, A. Cater-Steel, M. Toleman and W.-G. Tan, "A Proposal and Evaluation of a Design Method in Design Science Research," Electronic Journal of Business Research Methods, vol. 10, no. 2, pp. 89-100, 2012.

[18] T. Berners-Lee, R. Fielding, U. Irvine and L. Masinter, "Uniform Resource Identifiers (URI): Generic Syntax," Network Working Group: Fremont, CA, USA, 2005.