Review on feature-based method performance in text steganography

Mohd Hilal Muhammad¹, Hanizan Shaker Hussain², Roshidi Din³, Hafiza Samad⁴, Sunariya Utama⁵
¹Department of Computing & Management Sciences, Universiti Islam Antarabangsa Sultan Abdul Halim Mu'adzam Shah, Malaysia
²SP Intellect Resources, Malaysia
³,⁵School of Computing, College Arts and Sciences, Universiti Utara Malaysia, Malaysia
⁴Kolej Universiti Poly-Tech MARA Kuala Lumpur, Malaysia

Article Info

Article history:
Received Apr 8, 2020
Revised Jun 15, 2020
Accepted Jul 28, 2020

Keywords:
Capacity performance
Hidden message
Robustness performance
Security performance
Word-rule based

ABSTRACT

The implementation of steganography in text domain is one the crucial issue that can hide an essential message to avoid the intruder. It is caused every personal information mostly in medium of text, and the steganography itself is expectedly as the solution to protect the information that is able to hide the hidden message that is unrecognized by human or machine vision. This paper concerns about one of the categories in steganography on medium of text called text steganography that specifically focus on feature-based method. This paper reviews some of previous research effort in last decade to discover the performance of technique in the development the feature-based on text steganography method. Then, this paper also concern to discover some related performance that influences the technique and several issues in the development the feature-based on text steganography method.

This is an open access article under the CC BY-SA license.

Corresponding Author:
Roshidi Din,
School of Computing, College Arts and Sciences,
Universiti Utara Malaysia,
06010, Sintok, Kedah, Malaysia.
Email: roshidi@uum.edu.my

1. INTRODUCTION

The impact of IT application has a vital influence in every activity and productivity of the people at work [1, 2]. However, the existence of intruders in the communication technology enables anyone to easily retrieve and modify information [3]. Hence, one of considerations to deal with this matter is steganography as a part of information. Steganography is the scientific knowledge information that hide the hidden message implementation in several media that are unrecognized by human vision [4]. The steganography implementation could be applied in private communication, security system protection and any confidential data that is commonly used by government, military, industry and etc [5]. From that history, steganography is introduced as information hiding field that hides the confidential information to avoid the message from the third party [6]. Moreover, one characteristic of steganography is securing the important message in every medium to hide in the text [7]. The implementation of steganography knowledge field is divided into two main parts such as digital steganography and natural language steganography. Digital steganography conceals the hidden message through some media which are images, audios, video, network performance and other digitally undetectable codes [8]. Meanwhile, natural language steganography is the implementation of steganography that hides the hidden message in medium of text [4]. This study focuses on natural language steganography that is developed in text domain. It because text medium has a limited space...
to hide information, it is becoming a challenge to this study to implement the technique of steganography [9]. Generally, steganography can be classified into two parts which are digital steganography and natural language steganography. Digital steganography consists of four media such as image, audio, video, and protocol in order to cover the hidden message [8]. On the other hand, natural language steganography hide the hidden message in a medium of text so that the third party is unable to discover the presence of the message. There are two main groups in natural language steganography which are linguistic steganography and text steganography. The linguistic steganography domain is a type of steganography that is dependable with linguistic order of sentence in the text [10]. Meanwhile, text steganography manipulates the component of text such as word, line, space and other componentst of text in order to hide the message [11]. The development of text steganography method consists of two which are word-rule based method and feature-based method. This paper focus on review some technique in feature-based method in order to discover the related the performance and aslo issue that had happened in some of the technique by previous researcher’s effort.

2. RESEARCH METHOD

Text steganography conceals the hidden messages in the cover text based on structure in the text and any other characters of text without influencing the linguistic rule of text. There are two methods of text steganography which are word-rule based and feature-based method. The knowledge about word-rule based method consists of two category of hiding the message which are line-shift coding and word-shift coding. Line-shift coding is implemented to hide a message that can be embedded vertically to conceal the message in the text. This technique measures the length between the centroid texts that calculate the position based on the difference between spacing and the original text [12]. According to Sing et al. [13] about line-shift coding, the shifted development are those of 0 for a line-shifted up and 1 for a line-shifted down. Based on Roy and Manamisti [12], line-shift coding develops a unique shape in some degree of text shifted vertically and also explains the weakness of this technique in retyped text that can destroy secret information in the text. Meanwhile, the word-shift coding can embed the hidden message horizontally to conceal the message in the text.

This technique mostly justifies the document and is not shifted by the first and last word on each line of the text, [13]. Moreover, Liu et al. [14] proposed a technique adjacent character adjusted within sequence words in English text as the word-shift coding technique. This technique encoded the hidden message in matrix mode based on online chat and inferior of encoded shipping adjacent letters in text. The implementation of feature-based method can modify the uniqueness of letter by manipulating in the shape, size, and position in the text. This technique used the cover ext as the medium that embed with the hidden message based on uniqueness of text structure in cover text [15]. The cover text that embed with hidden message is named as stego text that will send to receiver that will extract to discover the hidden message that unknown by intruder [16]. The characteristic of this technique could be used by many researchers based on the characters of language in the world and also can be used in website text [17]. A lot of developments of techniques create several implementations of technique in covering the hidden message. Table 1 illustrates several techniques of feature-based method that is considering with the advantages and drawback that become a discovered issue and achievement in performance of technique. However, there are three main concerns of the performance that are elaborated in the next section.

| Technique                  | Schemes Used                      | Advantages                                                      | Disadvantages                                      |
|----------------------------|-----------------------------------|-----------------------------------------------------------------|----------------------------------------------------|
| Coverless steganography    | Coding matching binary transition | The effectiveness of algorithm in implementing the steganography | The technique does not fully cover the transition order of hidden message |
| Single Bit Rules [16]      |                                   | High robustness because it exists in the back end web page      | Low security that has to combine with cryptography |
| HTML Web page steganography [17] | Based on Khasida-PS of Arabic text | High capacity in the hidden message                             | Low security in extracting process                 |
| Arabic text [18]           |                                   | High capacity secret shares based improvement development technique | Low security against the strength attack            |
| Content-based sharing      | Based on Khasida of Arabic text   | Strong against the attack in the stego text from intruder       | Easily noticed in insertion text and low imperceptible capacity |
| Arabic Text [19]           |                                   | High efficiency for steganography purpose; able to use in Roman text | Easy to detect the changes in text after embedding process |
| Letter shaping in Arabic    | Based on two diacritics (Fatah and Kasrah) of Arabic text |                                           |                                                      |
| Text [20]                  |                                   |                                                 |                                                      |
| Transliteration in Bengali  | Based on characteristic of Bengali text |                                           |                                                      |
| text [21]                  |                                   |                                                 |                                                      |
### Table 1. The feature-based method performance (continue)

| Technique | Schemes used | Advantages | Disadvantages |
|-----------|--------------|------------|---------------|
| Character based of Chinese text [15] | Based on even and odd characteristic of Chinese text | This method is high in embedding capacity, good robustness | Dependable with corpus text that has coverless application |
| RNN-Generated Lyrics in Chinese text [22] | Based on the four sizes of candidate pool and structure lyric in Chinese text | Effective to resist the detection in traditional to avoid steganalysis algorithm | Less quality in capacity, security and robustness performance |
| Font color MS excel [23] | Based on the numbers and colour of Excel cells | High capacity embed that used in MS excel; easy extracting processing | Low robustness of the hidden message shown based on RGB colour. Easy to detect the embed process |
| Coverless English text [16] | Based on the letter in the Text | The security performance is guaranteed | It has plenty of space and low capacity |
| Character pair text [24] | Based on character starting, ending and character depicted | Large capacity in embedding space | Easy to detect any existence of hidden message |
| ATSTeg Via social media [25] | Based on ASCII code characteristics | Large capacity embedding with short cover text | It suffers the low robustness of stego text |
| Binary digit mapping on ASCII letters [26] | Based on ASCII characteristics | High capacity ratio to hide the hidden message in the cover text | Dependable with ASCII characters and easy to detect the changes |
| Arabic text hiding information [27] | Based on Fatah in the Arabic text | High capacity for embedding process in domain | If printed and retyped, the text will destroy; low invisibility |
| Multilayer Partially Homomorphic in Text steganography [28] | Based on numbering stream of text. | It has high hidden capacity than other techniques and high efficient selection letter | Lack robustness in avoiding possible attack in Homomorphic algorithm |
| English text using number oriented [29] | Based numbering and letter in English text | Fast loading time in embedding process steganography | Low robustness of stego text and security the security performance |
| Glyph perturbation [30] | Based on the alphabetical codebook in the text | High robustness in any format conversion the of text | Retyping process will destroy the message in document |
| Content-based Feature extraction [31] | Based on letters of vowel, consonants, lines and writable | Developed with maximum capacity, high embedding ratio with minimum time | Lack robustness and less security protection in encryption algorithm |
| Alphabet Pairing Text [32] | Based on letter and ASCII approach | High robustness and large embedding capacity | The technique is complex that is dependable with ASCII approach |
| Huffman Compression in Email Based [33] | Based on symbol @ in the email address | Large hiding capacity for embedding hidden message | Unavailable to execute in online condition |
| Right remark, Left remark, Zero width joiner, and Zero width joiner [34] | Based on characteristics and position of the letter | Easy to modify with simple requirement | This technique is unable to execute ASCII and Unicode |
| Compression ratio in Email [35] | Based on ratio vector letter on the text | Large capacity for embedding the hidden message | This technique can only develop in e-mail environment |
| Encryption with Cover Text and Reordering (ECR) [36] | Based on characteristic letter and inter-Word in the text | Large capacity embedding; has quick time to embed process | It has complex requirement in embedding and process of technique |
| Chain code using ANOVA [37] | Based on chain-code histogram in Bangla text | It is possible for the letters to embed the hidden message | Low security in covering the hidden message |
| Back end interface web page [38] | Based on characteristic HTML in Web page | It is able to embed large capacity and to transmit in internet | It is time consuming to implement and is only able to use in HTML |
| SKT and CCM [39] | Based on classification letter table dictionary in Chinese text | It has good performance in embedding and extracting | Low security to cover the hidden message |
| Vertical displacement of the point [34] | Based on Khasia variation in Arabic text | It can embed large capacity of hidden message | Easy to attack with retyping that can remove the hidden message |
| Secret steganography code for embedding (SSCE) [40] | Based on letter of a or an in English text | It has high robustness to avoid the intruder to discover the hidden message | It only embeds inconsonant and vowel word that makes low capacity to embed |
| Change alphabet letter pattern (CALP) [11] | Based on characteristic letters in the text | This technique is able avoid steganalysis technique | This technique is dependable with pattern and low robustness |
| Curve subheading (CURVE), vertical straight line (VERT), and quadnuple [41] | Based on characteristic letters in the text | Applicable to the soft-copy texts and cannot decode until it becomes an unaware technique | Low security and easily to detect the changes in the text |
| Numerical code [42] | Based on consonant letters of Hindi text | This technique has high security | This technique takes a long time for embedding the hidden message |
| SSM and HESM [43] | Based on traditional letter of Chinese | It has large capacity for embedding | Low security in covering the hidden message |

Review on feature-based method performance in text steganography (Mohd Hilal Muhammad)
Table 1. The feature-based method performance (continue)

| Technique                        | Schemes used                                      | Advantages                                           | Disadvantages                                          |
|----------------------------------|---------------------------------------------------|------------------------------------------------------|--------------------------------------------------------|
| Specific matra [44]              | Based on syntactic structure and sequence model   | It has large capacity for embedding the hidden message | Retyping of medium text and remove the existences hidden message |
| Reversed Fatah [45]              | Based on characteristic fatah in Arabic text      | It can embed large capacity and high invisibility    | Retyping the text can remove the existence of hidden message |
| Letter point in novel Arabic [46]| Based on characteristic of letter point in the text| It has high security to cover the hidden message      | It has low capacity to embed hidden message            |
| Rectangular region [47]          | Based on occlusive component in Chinese text      | It has transparency in extracting process             | It has limited medium letter that makes it have less capacity to embed |
| Machine translation[48]          | Based on parallel corpus and protocol overhead in the language text | This technique has a constant capacity               | The technique has high possible error                  |
| Mark up letter [49]              | Based on segment nodes in Hypertext               | This technique is useful for hypertext and online environment | This technique has low performance in development and low security |

3. COMPARATIVE PERFORMANCE ON FEATURE-BASED METHODS

According to Febryan, Purboyo and Saputra [50], there are three main performances in steganography that have a relation application performance which influence each other. The three relations are highly possible to become the achievement performance or as the drawback performance in implementing the steganography in covering the hidden message. The three relation performance of steganography is shown in Figure 1.

![Figure 1. The relation of three performances in steganography [50]](image_url)

In Figure 1, there three performances in steganography which are robustness, security, and capacity are considered to conceal the hidden message in some medium. However, the relation of the three performances has a contrary relation that is unable to adjust independently [50]. As such, the robustness and security will be decreased when the capacity performance is increased in applying the implementation of steganography. Based on the three performances, it is considered that the achievement and the possible issue happen in hiding the hidden message in text using the feature-based method of text steganography method. The criteria of the three relation performance the procedure of steganography as as follows:

a. **Robustness**: This is the capability to hide the hidden message in embedding process that is protected from an attacker to protect the stego text [31, 50].

b. **Security**: The level of safety performance that avoids the third party that has no connection with the sender and the receiver in steganography process to detect the existence of hidden message that is embedded in the text [20, 50].

c. **Capacity**: The quantity of data in the hidden message that is able to hide by embedding the hidden message in the text. The capacity data could be classified with size bit, number bit, and length of the text in performing the text steganography [25, 50].

The three relations are anticipated as an indicator to achieve the expected performance in steganography. However, the implementation is more dominant in achieving the high capacity rather than robustness and security in the feature-based of text steganography as shown in Figure 2. Figure 2 illustrates...
the comparison among the performances of robustness, security and capacity in the feature-based of text steganography method based on previous researchers in last decade. This figure classified the three performances based on advantages and disadvantages of the research effort in the development of the feature-based. It clearly seems that there is also four researchers’ effort that achieved high robustness and six researchers’ effort with low performance of robustness in the development technique of feature-based. For security performance, there are only two researchers’ effort that achieve high security and eight researchers’ effort which have some issues about security performance. However, the capacity performance in the development of the technique have 22 researchers’ effort and only three researchers’ effort with some issue in the development of the feature-based of text steganography method in the last decade.

Figure 2. The comparison relation performance in feature-based method in last decade

4. CONCLUSION
This paper is reviewed about feature-based method on text steganography based on several researchers’ effort in develop their techniques. It begins with the classification of the steganography category consisting of digital steganography and natural language steganography. This paper focuses on text steganography as a part of natural language. Then, the development of feature-based method by past researchers’ effort are reviewed along with the advantages and disadvantages between both methods in text steganography. This paper also presents the three relation performances in the development of the feature-based method which are robustness, security and capacity. It has discovered the most achievement performance in previous research effort, which is capacity performance while the highest issue in the feature-based of text steganography method is security performance.

ACKNOWLEDGEMENTS
We would like thank to Dean of School of Computing, Universiti Utara Malaysia (SoC CAS UUM) and Director of Awang Had Salleh Graduate School (AHSGS), Universiti Utara Malaysia for their moral support for the realization of this work.

REFERENCES
[1] R. Kaur, Pooja and Varsha, “A hybrid approach for video steganography using edge detection and identical match techniques,” 2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), Chennai, pp. 867-871, 2016.
[2] M. T. Ahvanooey, et al., “Modern text hiding, text steganalysis, and applications: A comparative analysis,” Entropy, vol. 21 no. 4 pp. 1-29, 2019.
[3] S. D. Torvi, K. B. ShivaKumar and R. Das, “An unique data security using text steganography,” 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, pp. 3834-3838, 2016.
[4] R. Din, et al., “The evaluation performance of letter-based technique on text steganography system,” Bulletin of Electrical Engineering and Informatics, vol. 8, no. 1, pp. 291-297, March 2019.
Review on feature-based method performance in text steganography (Mohd Hilal Muhammad)