Performance analysis on secured data method in natural language steganography

Roshidi Din[1], Rosmadi Bakar[2], Raihan Sabirah Sabri[3], Mohamad Yusof Darus[4], Shamsul Jamel Elias[5]
[1,2,3]School of Computing, College Arts and Sciences, Universiti Utara Malaysia, 06010 UUM Sintok, Kedah, Malaysia
[4]Faculty of Computer and Mathematical Sciences, University Technology of MARA, Shah Alam, Malaysia
[5]Universiti Teknologi MARA (UiTM) Kedah, Malaysia

Article Info

Article history:

Received Sep 30, 2018
Revised Nov 26, 2018
Accepted Dec 11, 2018

Keywords:

Data hiding
Decoding performance
Decrypting performance
Extracting performance

ABSTRACT

The rapid amount of exchange information that causes the expansion of the internet during the last decade has motivated that a research in this field. Recently, steganography approaches have received an unexpected attention. Hence, the aim of this paper is to review different performance metric; covering the decoding, decrypting and extracting performance metric. The process of data decoding interprets the received hidden message into a code word. As such, data encryption is the best way to provide a secure communication. Decrypting take an encrypted text and converting it back into an original text. Data extracting is a process which is the reverse of the data embedding process. The effectiveness evaluation is mainly determined by the performance metric aspect. The intention of researchers is to improve performance metric characteristics. The evaluation success is mainly determined by the performance analysis aspect. The objective of this paper is to present a review on the study of steganography in natural language based on the criteria of the performance analysis. The findings review will clarify the preferred performance metric aspects used. This review is hoped to help future research in evaluating the performance analysis of natural language in general and the proposed secured data revealed on natural language steganography in specific.

Corresponding Author:

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

1. INTRODUCTION

Various techniques including steganography are used to secure information on the internet. The presence of secret messages in steganography has been in use and using the internet has gained its popularity as it is highly preferred. A part of steganography is called natural language which is divided into text steganography and linguistic steganography. Text steganography is referred as hiding a message behind other files, while linguistic steganography contains linguistic device of generated and modified texts. Meanwhile, in most situations, linguistic structure becomes the space to hide the messages. The steganography relies on the type of the cover media used to embed secured data such as image, text, audio, and video.

Changing words in the text, changing the existing text format, generating random sequence or using context-free grammar to generate readable texts are related to text steganography. Natural language becomes complicated due to lack of excessive information available in images, audio or video files. Basically, the studies will focus on decoding, decrypting, and extracting in term of the performance metric. The process of translating the received messages into a specific code word is provided codes known as data decoding. Data encrypting is the best way to provide a secure communication. Decrypting takes an encrypted text and converting it back into the original text. Data extracting is a process which is the reverse of the data embedding process. In general, these methods will be proved through performance analysis to measure the effectiveness and efficiency method. The previous studies have mentioned about the aspects that are measured. From there we will able to see the aspects needed for each of these techniques and define the types of aspects that are emphasized to measure it. Performance analysis essentially is meant to measure a successful technique. To measure is a requirement that determines the performance metric aspect must be the focus. Numerous technique approaches have been proposed by previous researchers in steganography method due to the attention received lately. However, the problem arises because the undesirable intruder tries to steal the information, and because of that, it makes the researcher improve the technical part with the evaluation performance metric. Usually, the security aspect is one of the main concerns by the researcher for the protection of message or data while transmitting it over the networks. This is because to attain data security is the biggest challenging issue to transfer a secret message at this time. In order to get a better secured communication and prevent the hackers, it is a priority for the researcher to consider the performance metric aspect.

2. OVERVIEW OF SECURED DATA METHOD

This section presents the performance metric natural language steganography in evaluation for decoding, decrypting and extracting used in previous studies. There are two types of text steganography and linguistic steganography. There are 11 performance metric aspects in natural language used as shown in Figure 1.

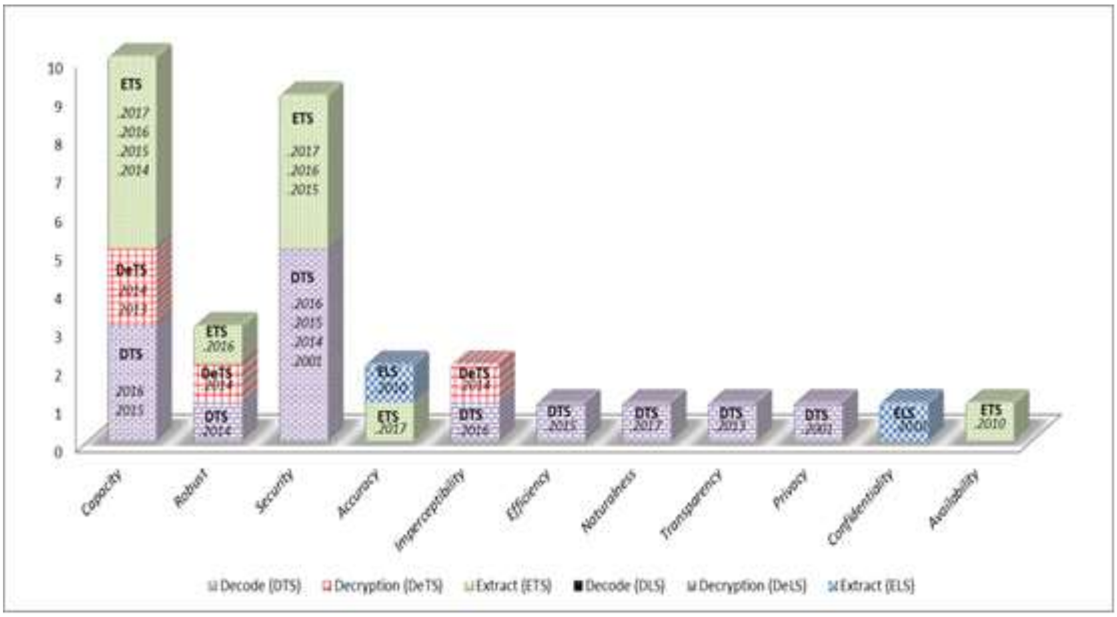

Figure 1. Natural language performance metric

Basically, the bar graph is categorized into decoding, decrypting and extracting in terms of text steganography and linguistic steganography. For example, there are six groups shown in figure 1 such as the decoding Text steganography (DTS), the decrypting text steganography (DeTS), and the extracting text steganography (ETS), the decoding Linguistic steganography (DLS), the decrypting Linguistic steganography (DeLS), and the extracting Linguistic steganography (ELS). The bar graph shows the number of literature collected from year 2001 to year 2017. It is clear that capacity aspect has the highest number in the bar graph with 10 studies. This is followed by security as the second highest with nine kinds of literature. The third highest is robust aspect with three citations. These three aspects dominate the three top most used according to the literature. While the remaining aspects: accuracy, imperceptibility, efficiency, naturalness, transparency, privacy, confidential and availability become the least aspects used. As for DLS and DeLS in data interpret, the number citation is none. However, the most studies are in DTS (14 literature), DeTS (5 literature) and ETS (12 literature).

3. PARAMETER METRICS USED

Table 1 represents the list of parameter metric used. This section represents the parameter metric on natural language type with references based on past research effort. The parameter metric is consisting of decoding, decryption, and extracting performance.

Table 1. Parameter used of natural language steganography

Natural Language	Text steganography			Linguistic steganography		
Years	*2017*		*2010*		*2001*	
Data Security / Parameter	Decoding	Decrypting	Extracting	Decoding	Decrypting	Extracting
Capacity	[1-3]	[4, 5]	[6-10]			
Robust	[10]	[4]	[8]			
Security	[2, 3], [11-13]		[6-8], [11]			
Accuracy			[14]			[15]
Imperceptibility	[2]	[4]				
Efficiency	[11]					
Naturalness	[16]					
Transparency	[17]					
Privacy	[13]					
Confidentiality						[18]
Availability			[19]			

Based on Table 1, natural language is divided into two categories. The first is text steganography and the second is linguistic steganography. There are 11 performance metrics available as a result of relevant search studies. This is considering the performance evaluation parameter of natural language frequently applied by the previous studies. Usually, the researchers used to evaluate their performance metric of approach. Regularly in the certain study, they used numerous aspects to evaluate their techniques, scheme or algorithm for enhancement. Among all those parameter metrics the three highest belong to capacity, security, and robust. All of them become the preference in evaluating the ability of an approach. Thus, they are considered as a priority to measure and find the strengths and weaknesses of any approach done by previous researchers. As can be seen in Table 1, capacity references dominated the literature which means that more researchers use capacity to measure the amount of the secret message that can be embedded without affecting the success of the technique. The second most used belongs to security. Which appears to be a protection against the outsider who might conceal the hidden message. Finally, robust is the third most applicable as it ensures that the algorithm has a better anti-attack capability. Based on table 1, a summary about the parameter use of natural language steganography will be presented in the next section.

3.1. Text steganography

Text steganography in decoding for capacity cited by [2, 3] has been using capacity to evaluate the approached technique. For instance, evaluation technique in terms of capacity and imperceptibility is performed by using a file size of 15.0 kb which contains 3040 whitespace word file poem. The number of characters can be done by looking for ratios after being embedded in a particular text file. For implementation, the evaluation of imperceptibility can examine the stego text content. Instead, this also refers to data security in decoding for an Imperceptibility as cited by [2]. Text steganography in decrypting for capacity cited by [4, 5] mentioned about capacity performance metric. There was a survey that involved hiding a secret message information in a cover text in the text steganography method. It was a technique that made it capable to hide 4-bits in the cover text once at a time through each and every character. From the sender to the recipient during the secret message, the amount of data bytes needs to be sent of n bytes defines as length capacity. For example; Capacity ratio = hidden message (bytes)/(cover text (bytes). For instance, to acquire the percentage capacity, the ratio is multiplied by 100 for one character that has a memory of one byte. If a person wants to hide the n bytes of the secret message, it will require him 2n+1 byte of the cover text =50% approx. Text steganography in extracting for capacity used by [6-10] mentioned about capacity performance metric. Those that are measured in text steganography are tested, between capacity and similarity. Capacity refers to the ability to embed the secured data tested while similarity refers to the cover text and stego text difference tested. Capacity ratio is acquired by dividing the total of hidden (bytes) over the cover text (bytes) size/(cover text (bytes)). The capacity ratio is calculated by using the formula;

$$\textit{Capacity Ratio}=\frac{\text{Size of hidden secret (x)}}{\text{Size of cover secret (n)}} \qquad (1)$$

Text steganography in decoding for robust used by [10] mentioned about robust performance metric. As for that, the most decoding improves detection is by factor 2. Caused by post selections of the protocol is robust against loss. The security that is analysed previously seems to be a straight forward extension. Therefore, protocols are now safe against the separation of beams. It has proposed an efficient phase coding scheme for quantum cryptography using a coherent state and post selection. This encoding increases the efficiency of tracking with the two most frequent factors. This protocol is robust against loss caused by post selection. Text steganography in decrypting for robust cited by [4] mentioned about robust performance metric. The studies stated algorithm that has great capabilities, good imperceptibility and a wide range of applications for the algorithm based on font format modifications, font styles and so on. The studies believe that the proposed approach delivers superior tendencies and thereby supports greater security compared to other methods. This security enables the secret message authenticity delivered by the sender to the receiver, to be checked. Text steganography in extracting for robust cited by [8] mentioned about robust performance metric stated that the main effort compares it in terms of robustness and ability to hide capacity, and to explore different text steganography techniques. Text steganography in decoding for security used by [2, 3], [11-13] mentioned about security performance metric, in which the security refers to the process of hiding information while being protected. It will seem difficult for an attacker to see an existing secret message.

Hence, the security is kept confidential and invisible so that the probability of being attacked is minimal. Then the data is assimilated into any multimedia documentation invisibly. For instance, the insertion of an invisible information into multimedia. Text steganography in extracting for security cited by [6-8], [11] mentioned about security performance metric, revealed the comparisons on the web with a proposed algorithm with four traditional data hiding methods. In comparison with other traditional algorithms, the result of the studies shows that security is better as well. However, the weakness is the page is prove to increase. Generally, the performance of the algorithm is good according to the studies that are about generated cover texts. These text have the ability to extract secret messages again randomly by concealing sensitive information from an unauthorized use by hiding the secret messages. Text Steganography is used as a security rate such as the number of cover-text characters used to simulate the message in it.

Text steganography in extracting for accuracy cited by [14] mentioned about accuracy performance metric. According to the proposed studies the level of security is higher in the format-based of text steganography algorithm which includes a private key cryptography. A different set of sample data has also been evaluated for its accuracy as the evaluation is with another method comparison. The result study shows a higher level of security with a steganography model. Text steganography in decrypting for an imperceptibility cited by [4] mentioned imperceptibility performance metric. Based on the result, it shows that there are advantages of the performance metric that has good imperceptibility and capacity algorithms when the font format and style are modified. The proposed approaches are meant to deliver a superior randomness, and as a result, it is able to support a higher security compared to other methods.

Text steganography in decoding for an efficiency cited by [11] is about efficiency performance metric. The studies are about the proposed approach evaluated in term of efficiency. The result revealed that the proposed method is against the existing method in the encoding process, in which it is more precise and remains in the decoding process. Based on this, the result shows that the precision rate is 9.5 while the decoding rate is 8.1. Meanwhile, evaluations are done in different word groups with the range of 3,5,10, and 15. As for accuracy, the coding is calculated as the ration of the number of word coded, and the decoding is calculated as the number of words ratio described per word volume. Text steganography in decoding for a naturalness cited by [16] stated that RNN has proved effective in generating poetry. The method yields stego-poems with high quality, comparable to the state-of-the-art poetry steganography, but has a much higher embedding rate. To evaluate a steganography system based on text generation, there are two important factors: one is the naturalness of generated text; the other is the embedding rate. Next, we evaluate our proposed method by making a comparison between different poetry steganography algorithms in these two factors. Experimental results show that our method yields stego-poems with high quality, comparable to the state-of-the-art poetry steganography, but has a much higher embedding rate. In the future, we would like to continually improve the quality of generated poems and explore poetry steganography using other genres, e.g. English sonnets or even normal style texts. In order for the proposed method to have a good perceptual transparency, text steganography in decoding for a transparency cited by [17] is based on the similarity of the font type, the capacity that must be high and robust, to the digital copy-paste operation. It is important to note that the capacity is very high of this method. By taking different cover document that have different types of font, text steganography method is tested by hiding some font types into a similar secret message.

Text steganography in decoding for a privacy cited by [13] mentioned that the studies make a comparison in term of concrete security, and some formulations of symmetric encryption schemes to be kept private under a selected plain text attacks are provided. There have been researches that explore how coding and fraud schemes and lead to the security attributes of encryption scripts generated. The studies ensures that the privacy and authenticity are associated with the number in the encoding scheme and become an underlying security in quantitative. Text steganography in extracting for an availability cited by [19] mentioned that the studies are about a technique-based secret message encryption on the length of English text documents in data hiding. The system uses VBA that analyses safety, capabilities, and availability for the evaluation parameter to be implemented.

3.2. Linguistic steganography

Linguistic steganography in extracting for an accuracy cited by [15] mentioned about studies that apply the statistical significant testify to extract the applicable syntactic reordering schemata with an automatic modern Greek paraphrase generation. These paraphrases do not have to be sophisticated when they are being altered. Instead, what is important is the number becomes the paraphrase and will be applied in steganography communication. The processes are robust domain-independent and portable to another language with similar syntax. Between safety and accuracy, the more accurate syntactic scheme is accuracy because security has a low level of usability. Linguistic steganography in extracting for confidentiality cited by [18] revealed that these studies proposed Chinese text information hiding algorithm in Chinese sentences with automatic paraphrasing techniques. Meanwhile Confidentiality would not cause any change to the algorithm format. In addition, there will be no misunderstanding in semantic although the original meaning of the text is changed. After all, the information hiding is based on the text context and the sentences that are paraphrased.

4. RESULT AND ANALYSIS

Provide a statement that what is expected, as stated in the "Introduction" chapter can ultimately result in "Results and Discussion" chapter, so there is compatibility. Moreover, it can also be added the prospect of the development of research results and application prospects of further studies into the next (based on result and discussion). This section presents the trends of decoding, decrypting and extracting performance metrics used in previous studies. The first trend is based on the most frequently used metrics while the second trend is to demonstrate the top three highest performance metrics used.

4.1. Preferred performance metric

As a reference to Figure 2, there are 11 performance metrics mentioned previously. It shows that the highest performance metric dominated in the previous studies are capacity (32%), security (28%) and robust (10%). Which the least remaining performance metrics are imperceptibility (6%), accuracy (6%), efficiency naturalness transparency, privacy, confidentiality, and availability with 3% respectively. The three upmost use belong to capacity, security and robust performance metrics. Thus, all these three are the greatest preferred performance metrics. Security refers to hiding information process safety and be protected. The existing secret message is hardly able to be noticed by an attacker. Meanwhile, capacity refers to the size of capacity that can be embedded without being traced. Otherwise it will be easy for hackers to modify the function of words and auxiliary words by deleting and adding on if the hacker knows the way to embed the secret message.

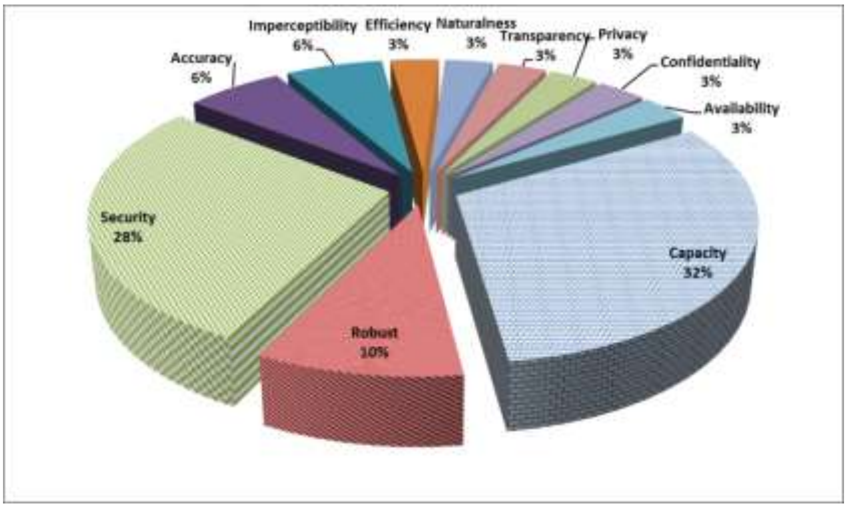

Figure 2. Preferred performance metric type

4.2. Extraction performance metrics

Table 2 is displays the final result of the three highest preferred performance metric types from the previous studies. The percentage obtained is based on the analysis in section 4 in pie chart Figure 2 in which the top three performance metrics are reduced into a majority of focused form. These three highly rated performance metrics apart from the 11 performances type are available in Table 2.

Table 2. Perentage preffered performane metric

Performance Metrics Type	Percentage (%)
Capacity	55
Security	28
Robust	17

Based on Table 2, capacity dominates the percentage use, which refers to the amount or size of the secret message that will affect the effectiveness of the method of text length sentences [18]. Meanwhile, security is related to the ability to succeed without an outsider notice that the exchange of the secret message belongs to the second preferred. Robust is the third influence used because hidden information cannot be destroyed by copying. It is also difficult to receive an attack from the anti-copy attack.

5. CONCLUSION

In conclusion, based on Figure 3 the researcher prefers to choose capacity, security and robust as the most influential aspects for performance metrics for measurement. This is the final result finding study for the aspects that the researcher has chosen.

In conclusion, the classification of the three highest preferred performance metrics, as shown in the above pie chart: capacity, security, and robust dominated 11 performance metric types. Since the capacity is the most popular aspect for performance metric for measuring, it is clear that many researchers give priority on the amount or the size of the secret message that can be embedded. Capacity, which refers to the size of the cover text, is believed to provide a better storage capacity. Security is the second most influential aspect for decryption performance metric for measuring. Many researchers think that the importance of security aspect should be taken seriously. Letting an unauthorized user to detect a secret message can cause breaking the security of the embedded message. That is why the data hiding techniques must be highly secured. Robust refers to the capability for anti-attack from the third party and makes sure that any algorithm or technique can be prevented from attack. In accordance with the main objective of natural language which is to prevent it from being noticed by the third party of its existence as a secret message, robust is placed third. Hence, this review is hoped to help future research in evaluating the performance of a natural language framework in general.

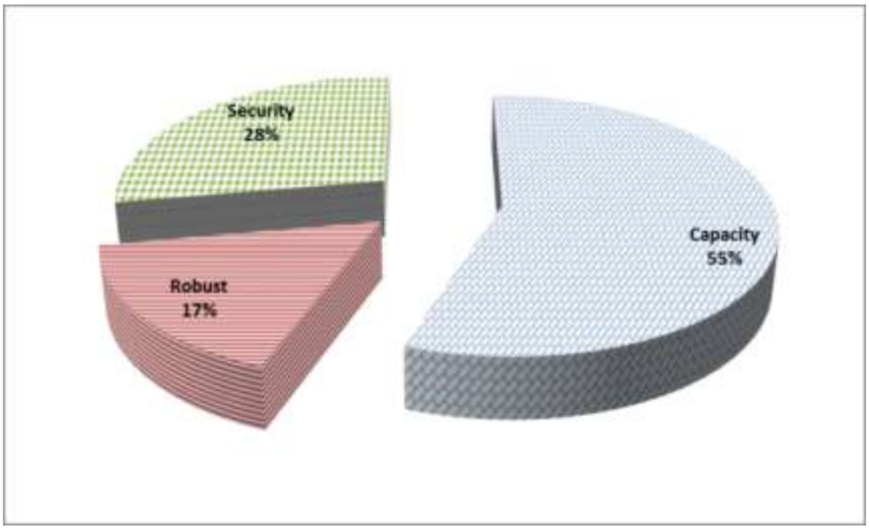

Figure 3. Preferred metric preferred result final

ACKNOWLEDGEMENTS

The author wish to thank the Ministry of Higher Education Malaysia in funding grant under the University Grant Non-PI with S/O 13850, RIMC of Universiti Utara Malaysia, Kedah.

REFERENCES

[1] Gupta S. and Jain, R. An Innovative Method of Text Steganography. *Proc. 2015 3rd Int. Conf. Image Inf. Process. ICIIP 2015*, 2015; pp. 60–64.

[2] Liang, O. W. and Iranmanesh, V. I*nformation Hiding using Whitespace Technique in Microsoft Word*. 2016.

[3] Dongardive, P. and Barde, C. Imp*roved Security Color Grace Steganography with Grace to Text Encoding and LSB*. 2015.

[4] Samanta, S. Dutta, S. and Sanyal, G. A Novel Approach of Text Steganography using Nonlinear Character Positions (NCP). 2014; November 2013, pp. 55–60.

[5] Kataria, S. Kumar, T. Singh, K.and Nehra, M. S. ECR (encryption with cover text and reordering) based Text Steganography*, 2013 IEEE 2nd Int. Conf. Image Inf. Process. IEEE ICIIP*, 2013; pp. 612–616.

[6] Huanhuan, H. Xin, Z. Weiming, Z. and Nenghai, Y. *Adaptive Text Steganography by Exploring Statistical and Linguistical Distortion*. 2017.

[7] Elmahi, M. Y. and Sayed, M. H. *Text Steganography Using Compression and Random Number Generators*. 2017; vol. 6, no. 6, pp. 259–263.

[8] Chaudhary, S.and Dave, M. An Elucidation on Steganography and Cryptography. 2016; pp. 3–8.

[9] Abbasi, A. T.and Ahmad, B. Urdu text steganography : Utilizing isolated letters, 2015;, pp. 37–46,.

[10] Kumar, R. Chand, S. andSingh, S. An Email based High Capacity Text Steganography Scheme Using Combinatorial Compression. *Proc. 5th Int. Conf. Conflu. 2014 Next Gener. Inf. Technol. Summit*, 2014; pp. 336–339.

[11] Vidhya, P. M. and Paul, V. A Method for Text Steganography Using Malayalam Text. *Procedia Comput. Sci.*, 2015; vol. 46, no. Icict 2014, pp. 524–531.

[12] Shah, R. and Chouhan, Y. S. Encoding of Hindi Text Using Steganography Technique. 2014; no. 1, pp. 22–28.

[13] Bellare, P. R. Mihir, *Encode-then-encipher Encryption: How to Exploit Nonces or Redundancy in Plaintexts for Efficient Cryptography*. 2001.

[14] Bhanu, V. Allada, C. and Susarla, M. *Developing an Efficient Solution to Information Hiding through Text Steganography Along with Cryptography*. 2017; vol. 8491, pp. 11–14.

[15] Kermanidis, K. L. Hiding Secret Information By Automatically Paraphrasing Modern Greek Text With Minimal Resources. 2010.

[16] Luo, Y. and Huang, Y. Text Steganography with High Embedding Rate. *Proc. 5th ACM Work. Inf. Hiding Multimed. Secur. - IHMMSec '17,* 2017; pp. 99–104.

[17] Bhaya, W. Rahma, A. M. and Al-nasrawi, D. TEXT STEGANOGRAPHY BASED ON FONT TYPE IN MS-WORD DOCUMENTS," 2013; vol. 9, no. 7, pp. 898–904.

[18] Jin, C. Zhang, D. and Pan, M. Chinese Text Information Hiding Based on Paraphrasing Technology. 2010; pp. 39–42.

[19] Changyan, D. A Data Hiding System Based on Length of English Text. 2017; pp. 161−166.

[20] Mulani, A. O. and Mane, P.B. Watermarking and Cryptography based Image Authentication on Reconfigurable Platform. *Bull. Electr. Eng. Informatics*, 2017;vol. 6, no. 2, pp. 181−187.

[21] Aggarwal, M.S.K. D. and Ahuja, B. A Secure Image Encryption Algorithm Based on Hill Cipher System. *Bul. Tek. Elektro dan Inform. (Bulletin Electr. Eng. Informatics)*, 2012; vol. 1, no. 1, pp. 51–60.

[22] Singh, J. Singh, B. Singh, S. P.and Naim, M. Performance Evaluation of Direct Torque Control with Permanent Magnet Synchronous Motor. 2011; vol. 1, no. January, pp. 165–178.