Substitution-based linguistic steganography based on antonyms

Fawwaz Zamir Mansor[1], Azizan Ismail[2], Roshidi Din[3], Aida Mustapha[4], Noor Azah Samsudin[5]
[1,2,4,5]Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Malaysia
[3]School of Computing, College of Arts and Sciences, Universiti Utara Malaysia, Malaysia

Article Info

Article history:

Received Mar 1, 2019
Revised Jun 5, 2019
Accepted Jun 26, 2019

Keywords:

Linguistic
Steganography
Substitution Approach

ABSTRACT

The study of steganography focuses on strengthening the security in protecting content message by hiding the true intention behind the texts. However, existing linguistic steganography approach especially in synonym-based substitution is prone to attack. In this paper, a new substitution-based approach for linguistic steganography is proposed using antonyms. The antonym-based stego-text generation algorithm is implemented in a tool called the Antonym Substitution-based (ASb). Evaluation of ASb was carried out via verification and validation. The results showed highly favorable performance of this approach.

Corresponding Author:

Azizan Ismail,
Faculty of Computer Science and Information Technology,
Universiti Tun Hussein Onn Malaysia,
86400 Parit Raja, Batu Pahat, Johor, Malaysia.
Email: azizan@uthm.edu.my

1. INTRODUCTION

Among all information medium available, text-based document is considerably an important one until the present time. Indeed, text-based documents are used primarily in both domain applications: business and academic. It is because a lot of important documentation information such as appointment letter, certification, report, confidential document and any other document are available in text. These documents carry valuable information, and therefore considerably prone to intruders' attack when disseminating through any communication channel. Essentially, intruders may temper the information for their own interests. The ultimate goal of any steganography technique is to embed secret messages into a chosen media, such as images [1, 2], text-based documents [3, 4], audio [5-7], and video [5, 7] files without easily noticed by any unintended third party. Essentially, the steganography strategy successfully hides the message in innocuous looking media in a camouflage manner. Therefore, the hidden message will arrive safely to intended recipients with less risk of any illegal attempts.

Figure 1 shows two categories of text steganography: format-based and linguistic-based steganography [8, 9]. The former usually apply either word rule-based or feature-based technique. The word rule-based technique embeds the hidden message based on word pattern by implementing line-shift coding or word-shift coding [10]. While the line-shift coding hides the message with vertically shifting hidden message in text lines, word-shift coding hides the message with horizontally shifting the hidden message in length between words [11]. On the other hand, the feature-based technique can alter unique feature characteristic in text based on code words. The feature based technique hides the message based on pattern letter or length of the word [12]. The existing implementation of linguistic steganography is prone to attack as the existing substitution-based approach emphasize on the use of synonyms. Indeed, the literature has shown that unintended recipients can easily guess the hidden message in the synonym based application of substitution-based approach [13]. The recipients can easily notice the existence of hidden messages during text analysis procedure [14]. In brief, the existing linguistic steganography approach especially in synonym-based substitution is prone to attack [15, 16].

Different from the existing substitution-based approaches which emphasize on synonym usage, this paper presents the use of antonyms to mask the meaning of the original message. The proposed approach will be evaluated using verification test and validation test. The remaining of this paper is as follows. Section 2 presents related works in linguistic steganography. Section 3 presents the methodology for developing the proposed tool for antonym-based linguistic steganography. Section 4 describes the evaluation results. Finally, Section 5 concludes with future plans.

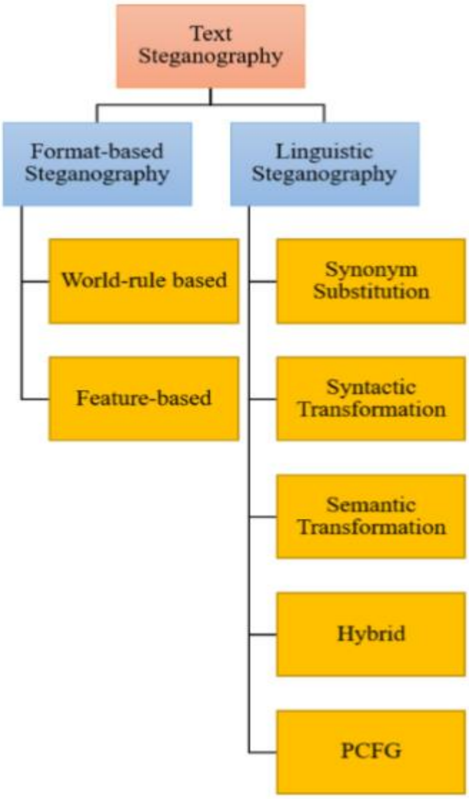


Figure 1. Classification of text steganography

2. RELATED WORK

At present, communication between people highly depends on technology. The dependency of communication technology has put privacy level in communication at risk. One of the potential risks is the existence of intruders in various communication channels. If there is no precaution, the intruders can easily retrieve and misuse the disseminated information sent throughout the communication channel. In fact, such intruders or attackers may disclose and abuse secret information to irresponsible parties [17]. Since the access to Internet is getting easier day after day, the probability of exposing of secret information to intruders and attackers is getting higher too. Therefore, we are taking additional measures to ensure the information right is well guarded or protected. One of the protection measures is to improve the existing linguistic steganography.

2.1. Linguistic Steganography

Linguistic steganography hides secret information by focusing on words usage and linguistic modification in encoding procedure. In linguistic steganography, the use of synonyms in substitution-based approach is one of the popular techniques. Synonym substitution-based suits any language as long as the text concerned has its corresponding synonym word. Table 1 briefly presents some scholarly papers on linguistic steganography within the last decade.

Table 1. Scholarly papers in linguistic steganography

Methods	Advantages	Disadvantages
Chinese text [18]	Simple and effective variants	Limited portion capacity
English text [15]	Printing electronic text only	Low security
Chinese text using WSD [19]	Has watermarking ability and highly robust	Rely of the WSD tool
Malay text [20]	High invisibility	Time-consuming
Synonym paraphrasing in Spanish [16]	Obvious use in Spanish language	Prone to attack and low volume number
English text using lexical [21]	High match for word replacement	High preservation of the grammar text
Synonym substitution [22]	Large volume	Complex algorithm
English text using context-based [23]	Low syntax error	Lack of vocabulary and narrow dictionaries

Apparently, previous studies have explored both, format-based steganography and linguistic steganography. In particular, the substitution-based approach in linguistic steganography really focus on the use of synonyms. However, synonym substitution is considerably prone to attack as the generated text carries the same meaning as plain text. Therefore, in this paper we explore the use of antonyms in the substitution-based approach to overcome the limitation of the synonym-based approach.

2.2. Linguistic semantic of antonyms

The study of meaning in language to express oneself is also known as linguistic semantic [24, 25]. Linguistic semantic solves a problem in language understanding that eventually leads to word selection process. For example, the synonym of the word “presence” is “attend'”, but its antonym is “absence'”. One can indicate both, the synonym and antonym for the given text. Focusing on the antonym approach, an opposite word will replace the given text. Table 2 shows some words with their respective antonyms.

Table 2. Substitution-based Steganography model using antonyms

Word	Antonym
minor	major
success	losses
war	peace
arrest	discharge
fight	truce
presence	absence
accept	decline
heal	ail
agony	pleasure
arrive	bail
cover	offence
chaos	peace

3. MATERIALS AND METHODS

This paper proposes a new method of hiding secret or encrypted message within a cover text medium by using antonyms of selected words in the cover text. This antonym-based method is then implemented as a tool called the Antonym Substitution-based (ASb) Steganographic Tool. based on linguistic steganography. In the proposed antonym-based method, the syntactic and semantic structure of the cover text are maintained so it appears to be normal when seen by the unknowing readers or recipients. Figure 2 illustrates the flow design process of embedding stego message using antonym word. The process started by inserting plain text, then the dictionary will be checked in database to match with its antonym. Then, the antonym text will replace the matched plain text to generating stego message.

Next, Figure 3 shows the proposed substitution-based steganography model using antonyms. The proposed model is implemented into a tool called the Antonym Substitution-based Steganographic Tool (ASb), where the algorithm of the proposed approach is coded into.

Meanwhile, Figure 4 shows the interface of ASb, which allows user to insert secret text within the cover texts. The button ‘Generate to Stego Text’ is used for converting the secret text entered into stego texts. The button ‘Generate to Secret Text’ is used for converting the stego text back to the original text. For validation purposes, ASb will later be tested with two different word types as the input; past tenses and present tenses. 100 set of input using both tenses were taken from newspaper article and fed into the ASb tool. Next, the experiment recorded the stego text generated.

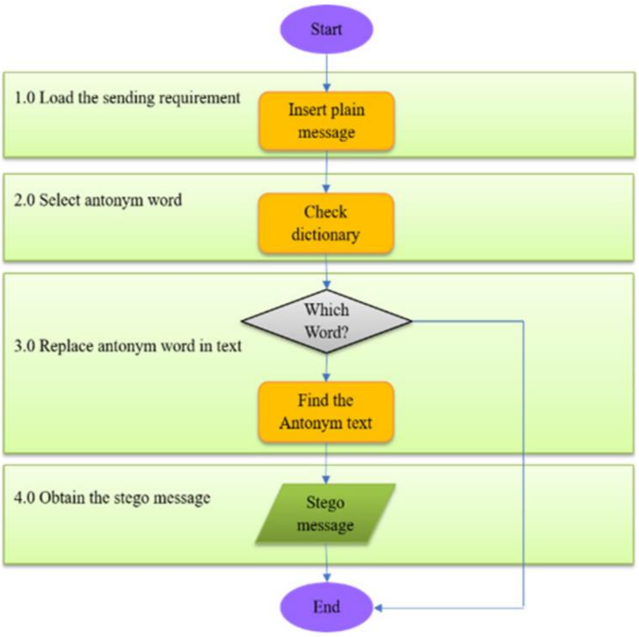


Figure 2. Flowchart of ASb

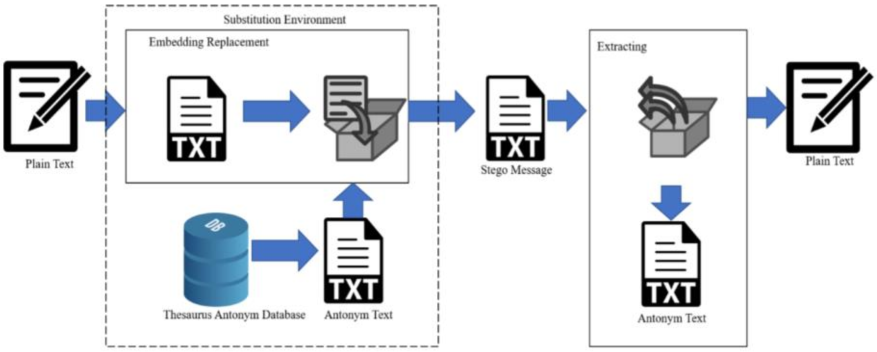
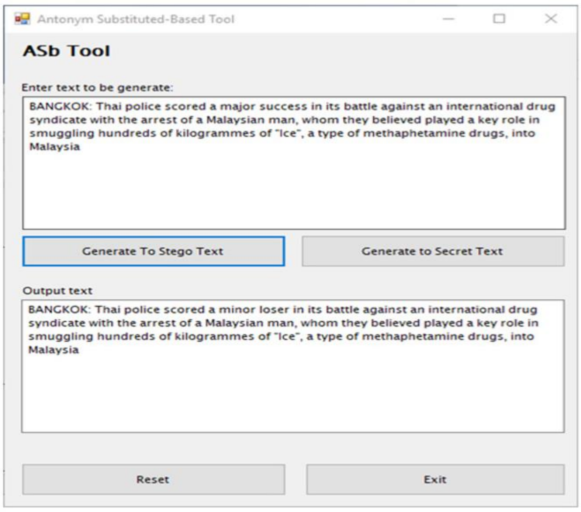


Figure 3. Substitution-based steganography model using antonyms



BANGKOK: Thai police scored a major success in its battle against an international drug syndicate with the arrest of a Malaysian man, whom they believed played a key role in smuggling hundreds of kilogrammes of "Ice", a type of methaphetamine drugs, into Malaysia

BANGKOK: Thai police scored a minor loser in its battle against an international drug syndicate with the arrest of a Malaysian man, whom they believed played a key role in smuggling hundreds of kilogrammes of "Ice", a type of methaphetamine drugs, into Malaysia

Figure 4. Interface of ASb

4. RESULTS AND DISCUSSIONS

The Antonym Substitution-based Steganographic Tool (ASb) developed was evaluated using the verification and validation methodology. The evaluation process involved the performance profiler tool from Microsoft Visual Studio.

4.1. Verification test

Figure 5 shows the wall clock time (seconds) that is running based on the CPU Usage that has been identified. It shows at the beginning of 1 second, the percentage for CPU Usage is 20%. The CPU usage was dropping and gradually increased back at second 6.5 and 9. The pattern of this graph shows that this ASb tool is running smoothly. This analysis result indicates the ASb steganographic tool can load the antonym database without slowing down the CPU usage.

Figure 6 shows the report of the functions doing most of the individual works. It listed the functions and its percentage of exclusive sample being used. Also, it listed the percentage of exclusive samples from the highest to the lowest. The following are the list of exclusive samples; 73.19%, 5.58%, 5.03%, 3.50%, and 2.19%. The functions that are doing most individual works as stated in Figure 5 is the *WindowsFormsApplicationBase.Run* that contains 73.19% of exclusive samples. The lowest percentage contains 2.19% of exclusive sample that is from the *System.Windows.Forms.Control.setFont*.

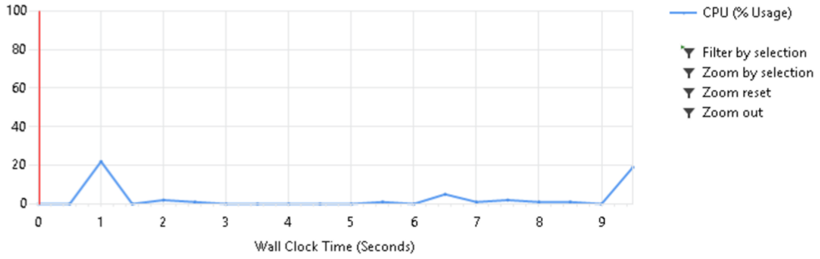
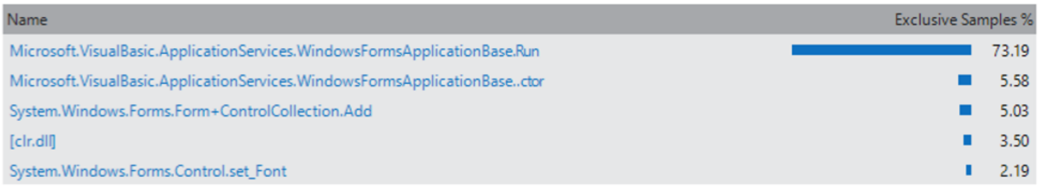


Figure 5. Diagnostic session of CPU usage



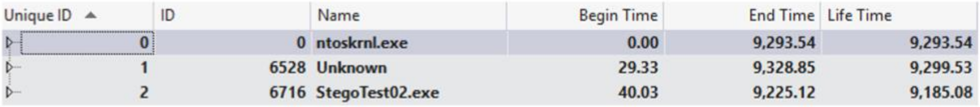
Name	Exclusive Samples %
Microsoft.VisualBasic.ApplicationServices.WindowsFormsApplicationBase.Run	73.19
Microsoft.VisualBasic.ApplicationServices.WindowsFormsApplicationBase..ctor	5.58
System.Windows.Forms.Form+ControlCollection.Add	5.03
[clr.dll]	3.50
System.Windows.Forms.Control.set_Font	2.19

Figure 6. Functions doing most individual works

Figure 7 shows the process time from begin time, end time, and life time. It is stated that *ntoskrnl.exe* at begin time of 0.00, end time 9.293.54, LifeTime 9.293.54. Then, IDof6528 begin time at 29.33, end time 9.328.85, Life Time 9.299.53. Lastly, ID of 6716, named *StegoTest02.exe*, begin time at 40.03, end time at 9.225.12 and life time 9.185.08.

Figure 8 shows the memory profiling report of performance. The wall clock time was been running at 16 seconds. The performance was at 0.5 second, when it reached 60%, it started to decrease at 1(s). It decreased until 3.5 at 20% and increases back at 4 and decreases back at 4.5. The performance is not consistent until end. Next, Figure 9 shows the functions that (a) allocate the most memory, (b) types with the most memory allocated, and (c) types with most instances.

Figure 10 shows that the percentage of bytes for each function, whereby the highest percentage of bytes is from *System.String.Split*(*char*[]) at 42.08. As for the most memory allocated, the highest bytes was generated at 48.34%, followed by 22.49%, 14.38%, 5.35 and lastly 1.68%. *System.String* is the type that has the most memory allocated. Finally, the highest type that has most instances is 54.15%, followed by 10.76%, 10.02%, 9.96%, and lastly 3.43%. *System.String* is the type that has the highest percentage of instances. Figure 10 shows the process memory usage that happens when the ASb tool is running the diagnostic session at 28.848 seconds. At this time, the tool uses 19 MB of memory. The memory is consistent when the ASb was running.



Unique ID	ID	Name	Begin Time	End Time	Life Time
0	0	ntoskrnl.exe	0.00	9,293.54	9,293.54
1	6528	Unknown	29.33	9,328.85	9,299.53
2	6716	StegoTest02.exe	40.03	9,225.12	9,185.08

Figure 7. Processing time

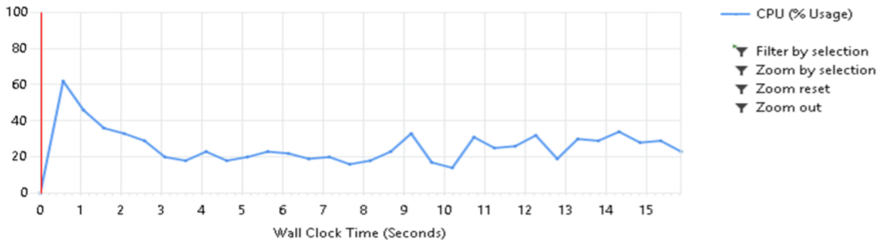
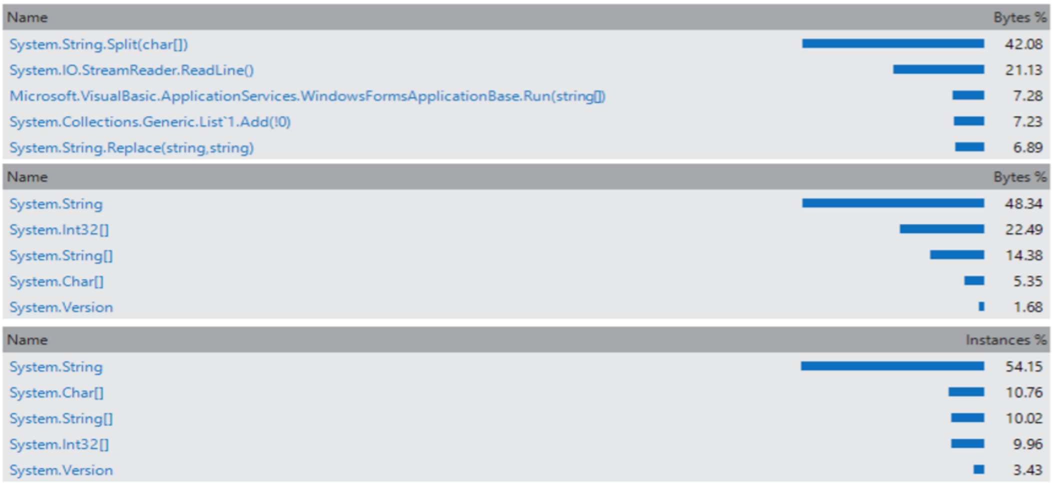


Figure 8. Memory profiling report



Name	Bytes %
System.String.Split(char[])	42.08
System.IO.StreamReader.ReadLine()	21.13
Microsoft.VisualBasic.ApplicationServices.WindowsFormsApplicationBase.Run(string[])	7.28
System.Collections.Generic.List`1.Add(!0)	7.23
System.String.Replace(string,string)	6.89

Name	Bytes %
System.String	48.34
System.Int32[]	22.49
System.String[]	14.38
System.Char[]	5.35
System.Version	1.68

Name	Instances %
System.String	54.15
System.Char[]	10.76
System.String[]	10.02
System.Int32[]	9.96
System.Version	3.43

Figure 9. Memory allocation

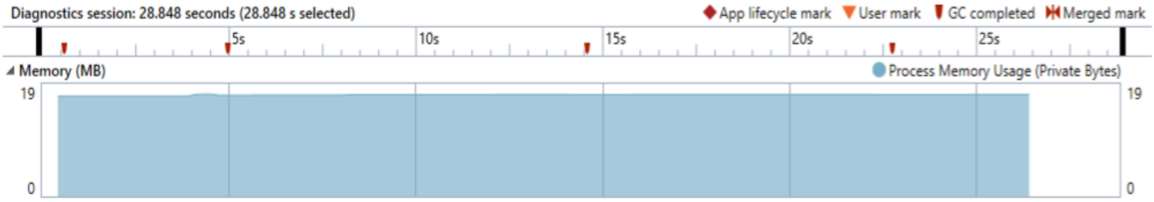


Figure 10. Processing memory usage

4.2. Validation test

The evaluation of performance will validate two types of stego text that were extracted from the dataset selection via past tense and present tense. 100 sets of first paragraph newspaper articles were taken, whereby 100 sets of past tenses and 100 sets of present tenses. Each set of datasets was measured and the outputs were listed. Table 3 shows there are four types of different outcome that are generated from the past tense dataset; true positive, true negative, false positive and false negative. In this case, 100 sets of data were analyzed from the past tense type of sentence and the result was 63/100 were true positive, 6/100 were true negative, 8/100 were false positive. Lastly; 23/100 were false negative.

Table 3. Percentage of outcome value for past tenses

Outcome	Value
True Positive (TP)	63
True Negative (TN)	6
False Positive (FP)	8
False Negative (FN)	23
Total	100

Next is the process of measurement, whereby the evaluation on validation will proceed into precision calculation to get the value that it will produce. In this calculation, 63 plain sets of True Positive from a group of Past Tenses were identified, and divided into a total of True Positive and False Positive (63 + 8 = 71). The result from this calculation is 88.7%. This means it implies a high imperceptibility in steganography terms. This can be proven that the past tense type of sentence is applied and implemented successfully using antonyms. Table 4 shows testing of past tense sentences that were carried to form the measurement of precision, recall, f-measure and accuracy. The precision was rated at 88.7% for this past tense type of sentence, while recall had been rated at 73.3%. For F-measure, it was rated at 80.0% whereas the accuracy was rated at 69%.

Table 4. Validation results for past tenses

Type of Sentence	Precision	Recall	F-Measure	Accuracy
Past Tense	88.7%	73.3%	80.0%	69.0%

The overall result on all the testing of past tense type of sentences showed that this kind of sentence can adapt well in antonym substitution. Precision, recall and accuracy scored more than 70% of each test that has proven that it has achieved steganography measurement. The results are shown in Figure 11.

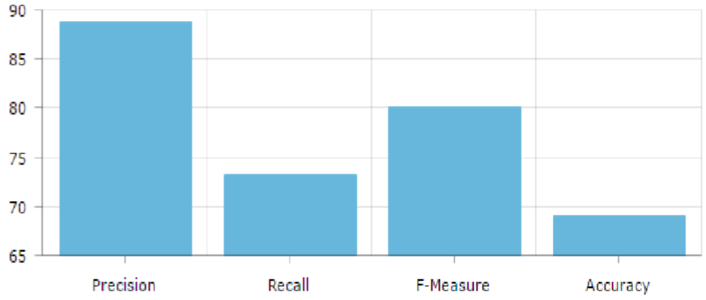


Figure 11. Percentage of validation results for past tenses

Table 5 shows four types of different outcomes that are generated from the present tense dataset. The outcome is true positive, true negative, false positive and false negative. Based on Table 5, 100 sets of data were analysed from present tense type of sentence and the results were 72/100 were true positive, 4/100 were true negative, 18/100 were false positive. Lastly, 6/100 were false negative. Since the dataset was chosen from present tense type of sentence, the result of TP (72/100) was the total of the overall result that shows most of the stego generated from present tense were correctly classified correctly as stego. Meanwhile, the minority value on the table, TN (4/100) shows that the cover medium was classified as cover but not correctly as stego. Then, the rest of the value of FP (18/100), shows that this type of dataset, sometimes generate the cover medium wrongly as stego. Lastly, FN (6/100) shows the cover medium was wrongly classified as cover.

Table 6 shows that the present tense type of sentences had been used to evaluate the few types of measurement like precision, recall, f-measure and accuracy using their equation. The equation had been applied and generated results that can be used for this study to evaluate the value of each measurement. For the present tense, the rate of the precision score is at 80% appears to be quite high, compared to the recall which also scored a high value at 92.0%. Meanwhile, the F-measure value on the present tense scored a value at 92.0% while accuracy scored at 72.0%.

Table 5. Percentage of outcome value for present tenses

Outcome	Value
True Positive (TP)	72
True Negative (TN)	4
False Positive (FP)	18
False Negative (FN)	6
Total	100

Table 6. Validation results for present tenses

Type of Sentence	Precision	Recall	F-Measure	Accuracy
Present Tense	80.0%	92.0%	83.0%	72.0%

The overall result on all the testing of the present tense sentences shows that this kind of sentence can adapt well in antonym substitution. Precision, recall and accuracy scored more than 70% on each test proved that it has achieved steganography measurement. The validation results are shown in Figure 12.

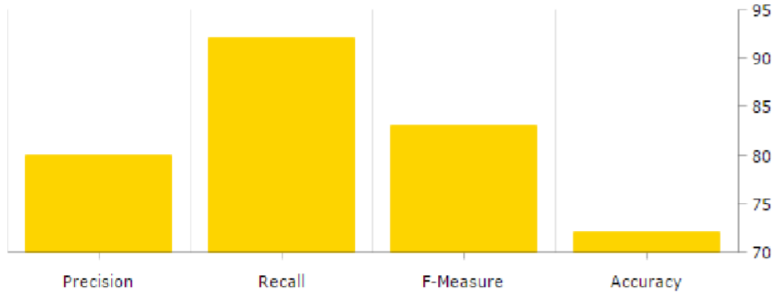


Figure 12. Percentage of validation results for present tenses

Overall, for precision, sentences in both past and present tense scored more than 80%. The rate of precision in input using past tenses is 88.7%, which is higher than the present tense, which is 80%. The difference in this type is very close that is only 8.7%. This implies a high imperceptibility in terms of steganography. This proves that both types of dataset can be used to apply antonym substitution. Note that results from both precision p and recall r generated can be interpreted as a weighted average of the precision and recall where the best value is high. As for F-measure, sentences in past tense sentences resulted in value of 3% lower than the present tense. Both scored high that proves the performance of generating stego text is very good. Finally, past tense scored 69% on accuracy value that is high for accuracy, meanwhile, the present tense scored higher than past tense that is valued at 72%. Both results are high in accuracy which shows that the dataset of the present tense adapted better in the antonym substitution scheme.

In summary, the verification performance was obtained using the performance profiling tool from the algorithm that was implemented in the proposed ASb steganographic tool. It is then followed by the validation performance that consists of output dataset measured using the rates of precision, recall, F-measure and accuracy. The overall results of antonym substitution show positive outcome. The value of each result was high and achieved the measurement and requirement on steganography specification. It is very optimistic to see the proposed scheme of antonym substitution to further investigated in the future.

5. CONCLUSION

This paper presented a new linguistic steganography tool based on antonym substitution called the ASb. The strength and the effectiveness of linguistic steganography is depending on how immune the algorithm is to detection. Antonym-based substitution approach is more robust as compared to simple synonym-based substitution due to the inverse relationship of the word similarity. ASb is hoped to provide a platform for communicating innocuous text as well as to promote utilization of cover texts.

ACKNOWLEDGEMENTS

This work is supported by Universiti Tun Hussein Onn Malaysia under the Tier 1 Grant Scheme U891.

REFERENCES

[1] Pramanik S, Singh R P, Ghosh R. A new encrypted method in image steganography. *Indonesian Journal of Electrical Engineering and Computer Science*. 2019 Jun;14(3):1412-9.

[2] Dogan S. A new data hiding method based on chaos embedded genetic algorithm for color image. *Artificial Intelligence Review*. 2016 Jun 1;46(1):129-43.

[3] Lockwood R, Curran K. Text based steganography. *International Journal of Information Privacy, Security and Integrity*. 2017;3(2):134-53.

[4] Taleby Ahvanooey M, Li Q, Hou J, Rajput AR, Yini C. Modern Text Hiding, Text Steganalysis, and Applications: A Comparative Analysis. *Entropy*. 2019 Apr;21(4):355.

[5] Kakde Y, Gonnade P, Dahiwale P. Audio-video steganography. *In 2015 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)* 2015 Mar 19 (pp. 1-6). IEEE.

[6] Zhang S, Khan I, Ullah Y. Audio Steganography by Additional Channel. *In Recent Developments in Intelligent Computing, Communication and Devices* 2019 (pp. 633-642). Springer, Singapore.

[7] Hossain K, Parekh R. An approach towards image, audio and video steganography. *In 2016 2nd Int. Conference on Research in Computational Intelligence and Communication Networks (ICRCICN)* 2016 Sep 23 (pp. 302-307). IEEE.

[8] Mansor FZ, Mustapha A, Samsudin NA. Researcher's Perspective of Substitution Method on Text Steganography. *InIOP Conference Series: Materials Science and Engineering* 2017 Aug (Vol. 226, No. 1, p. 012092). IOP Publishing.

[9] Taskiran CM, Topkara U, Topkara M, Delp EJ. Attacks on lexical natural language steganography systems. *In Security, Steganography, and Watermarking of Multimedia Contents* VIII 2006 Feb 15 (Vol. 6072, p. 607209).

[10] Li L, Huang L, Zhao X, Yang W, Chen Z. A statistical attack on a kind of word-shift text-steganography. In 2008 *Int. Conference on Intelligent Information Hiding and Multimedia Signal Processing* 2008 Aug 15 (pp. 1503-1507). IEEE.

[11] D. Bharti and A. Kumar, "Enhanced Steganography Algorithm to Improve Security by using Vigenere Encryption and First Component Alteration Technique", *Int. J. Eng. Trends Technology*, 13(5):242-246, 2014.

[12] M. V. Nasab and B. M. Shafiei, "Steganography in Programming", *Aust. J. Basic Appl. Sci.*, 3(12):1496-1499, 2011.

[13] S. Bhattacharyya, A. P. Kshitij, and G. Sanyal, "A Novel Approach to Develop a Secure Image based Steganographic Model using Integer Wavelet Transform", *in Proceedings of the International Conference on Recent Trends in Information, Telecommunication and Computing (ITC)*, 2010, pp. 173-178, 2010.

[14] S. Dulera, D. Jinwala, and A. Dasgupta, "Experimenting with the Novel Approaches in Text Steganography", *Int. J. Netw. Secur. Its Appl.*, 3(6): 213-225, 2011.

[15] M. H. Shirali-Shahreza and M. Shirali-Shahreza, "A New Synonym Text Steganography", *in Proceedings of the 4th International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, pp. 1524-1526, 2008.

[16] A. Munoz, J. Carracedo, I. A. Alvarez, "Measuring the security of linguistic steganography in Spanish based on synonymous paraphrasing with WSD", *in Proceedings of the 10th IEEE International Conference on Computer and Information Technology*, pp. 965-970, 2010.

[17] M. M. Amin, M. Salleh, S. Ibrahim, M. R. Katmin and M. Z. I. Shamsuddin, "Information Hiding using Steganography", *in Proceedings of the 4th National Conference on Telecommunication Technology*, pp. 21–25, 2003.

[18] L. Yuling, S. Xingming, G. Can and W. Hong, "An Efficient Linguistic Steganography for Chinese Text", *in Proceedings of the 2007 IEEE International Conference on Multimedia and Expo*, pp. 2094-2097, 2007.

[19] H. Lu, L. JianBin, L. TianZhi and F. DingYi, "An anti-attack watermarking based on synonym substitution for Chinese text", *in Proceedings of the 5th International Conference on Information Assurance and Security*, pp. 356-359, 2009.

[20] H. Z. Muhammad, A. Shakil and S. M. S. A. A. Rahman, "Synonym-based Malay linguistic text steganography", *in Proceedings of Innovative Technologies in Intelligent Systems and Industrial Applications*, pp. 423-427, 2009.

[21] C.-Y. Chang and S. Clark, "Practical Linguistic Steganography using Contextual Synonym Substitution and Vertex Colour Coding", *Emnlp*, 40(2):403-448, 2010.

[22] C. Qi, S. Xingming and X. Lingyun, "A secure text steganography based on synonym substitution", *in Proceedings of the IEEE Conference Anthology*, pp. 1–3, 2013.

[23] F. Wang, L. Huang, Z. Chen, W. Yang and H. Miao, "A Novel Text Steganography by Context-Based Equivalent Substitution", 2013.

[24] Nida EA. A componential analysis of meaning: An introduction to semantic structures. Walter de Gruyter GmbH & Co KG; 2015 Jun 3.

[25] E. L. Coderre, M. Chernenok, B. Gordon and K. Ledoux, "Linguistic and Non-Linguistic Semantic Processing in Individuals with Autism Spectrum Disorders: An ERP Study", *J. Autism Dev. Disord.*, 47(3):795-812, 2017.