STEVENOGRAPHY CAPACITY ANALYSIS OF TEXT IN TEXT WITH
SENTENCE STRUCTURE IN INDONESIAN

R. Fanry Siahun,1, Amran Sitohang2, Ibran Febrians1, Widia Putri3
1STMIK Pelita Nusantara
Jl. Iskandarmuda No.1 Medan
E-mail: *rfanry@gmail.com

ABSTRACT-Data security issues become very important when the computer has been used as a communication tool on a global network (internet). One method that is quite popular for securing data from irresponsible parties is text-based steganography, where confidential data or information is hidden or inserted into other text media so as not to arouse suspicion from other parties. Based on the input data calculated on the number of words, it has a significant influence on the results of the stegotext. This can be seen from the comparison of the length of the input message which is calculated in the number of words with the length of the output message (stegoteks) which is influenced by the length of the style (pattern) of the sentence structure used. For input data with a word length of 1, the average capacity of the stegotext is 8 to 9 words or 12.4%; for input word length 2, the average capacity of the stegotext is 22 words or 8.85% and the input word length is 4 then the capacity of the stegotext is between 30 to 33 words or equivalent to 12.91%.

Keywords: Text Steganography; Sentence Pattern; Payload Balancing Capacity

1. PRELIMINARY

In data communication on a computer network, one important thing that must be understood and paid attention to is information security, one of the important aspects of information security is the aspect of integrity (authenticity), this aspect guarantees the authenticity of information on computer network users so that the information cannot be changed by unauthorized parties, having the right to change it[1]. Along with advances in technology, all the information needed can be obtained easily, including top secret information. Because with the help of technology, all confidential information that is locked or stored properly can even be opened and accessed by irresponsible parties if the methods used in information security are simple or predictable.[2][and][3].

Many acts of theft of information-information users such as the username and password of an account or important data because there is no protection against confidentiality and integrity aspects.[4]. In a computer network, this will certainly be fatal to the network users. Therefore, it is necessary to conduct an analysis that aims to measure the level of confidentiality through the capacity of a message or user information in a computer network to develop the network to be even better in terms of protecting the confidentiality of user information.[5],[6][and][7].

With current technological developments, secret messages hidden in media or encrypted messages can be opened by steganalysis and cryptanalysts because the secret messages are only inserted or only encoded using certain methods. Based on the results or output of the text-based steganography process for a message, text steganography in the Indonesian language pattern will be applied in analyzing the capacity (message insertion results) in the Indonesian sentence structure.[8][9].

Some of the results of research on the use of steganographic media that have been carried out until now include the proposed Bayesian method which performs better than the existing steganalysis method for detecting multiple steganography in low-bit compressed messages AbS-LPC.[3][and][10].

2. RESEARCH METHODS

The research method used in this research is descriptive qualitative where qualitative data is descriptive data in the form of numeric symbols. Qualitative data is carried out to understand empirical phenomena that are focused on finding as much picture as possible without detailing the relationships between variables[13] which is carried out in several stages, namely:

1. Data collection technique

In carrying out this research, the data collection techniques used are described in the following stages:
a. Research sites

Determining the location of the research in this study is very important to obtain the data needed in the study. The research location was conducted at STMIK Pelita Nusantara.
b. Data collection

The internal method used in data collection is descriptive qualitative, the data used is sourced from the Big Indonesian Language Dictionary in 2008, which has 7 elements of word types.

<table>
<thead>
<tr>
<th>Kode</th>
<th>Keterangan</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adj</td>
<td>Adjektive</td>
</tr>
<tr>
<td>Adv</td>
<td>Adverbia</td>
</tr>
<tr>
<td>Noun</td>
<td>Nounia</td>
</tr>
<tr>
<td>Num</td>
<td>Numeralia</td>
</tr>
<tr>
<td>Par</td>
<td>Partikel</td>
</tr>
<tr>
<td>Pron</td>
<td>Pronominal</td>
</tr>
<tr>
<td>Ver</td>
<td>Verba</td>
</tr>
</tbody>
</table>

Table 1. Word Class

Sumber: KBBI, 2008

The preprocessing stage is the filtering or data selection stage that aims to get the right data to be used in the next process. In this preprocessing stage, two input components are formed, namely a dictionary and a pattern. These components along with the secret message will be input. The secret message entered is in the form of ASCII characters that the user typed into the program. The cover media where the message is inserted in this study is generated from the dictionary and pattern components along with the insertion process.

In this research pattern, the sentence used is a combination of the grammatical functions of sentences in Indonesian, namely subject (S), predicate (P), object (O), complement (Pel), and adverb (K). The five grammatical functions form eight simple single sentence patterns[14],[15].

![Diagram](image-url)

Figure 1. D Dictionary Development Process

Data Analysis Techniques

Pattern made by listing all possible combinations (cross products) of grammatical elements from 8 basic sentence patterns. Each combination of grammatical elements of each pattern is placed in a pattern file. To overcome the issue of the length of the message being limited by the pattern length, the function is used

\[ SIZER(C) + C + R \]

To ensure that the secret message can be completely hidden, the SIZER function processes the binary string C from the message with the ASCII code entered in the system, into a new binary string (Stegoteks). The resulting stegotext SIZER consists of a string of a certain length that represents the length of C (Cbit), followed by a binary message string (C), followed by a random string R. The function DESIZER(SIZER(C)) = C is the inverse of the SIZER function to get the return message string C from Stegoteks. The SIZER and DESIZER functions that are applied to the message insertion and extraction process to get the C message to become:

\[ DESIZER\left( EMBED_2\left( SIZER(C) \right) \right) = C \]

Testing and Implementation

Conclusion

Figure 3. Research stages
The pattern generation process begins with defining a grammatical function. Each function is filled with elements in the form of grammatical categories. Category filler in this study is Part of Speech (POS) in the form of 7 types of words from word class which are also used to group words in the D dictionary.

Filling element function subject and predicate categorized as adjectives, nouns, numerals, pronouns, and verbs. Object filler elements are nouns and pronouns. Complementary grammatical functions are filled with nouns, numerals, pronouns, and verbs, while adverb grammatical functions can be filled with adverbs, nouns, and pronouns.

Table 1. Combination of grammatical elements of SP. sentence patterns

<table>
<thead>
<tr>
<th>Subject-Predicate</th>
<th>Part of Speech</th>
</tr>
</thead>
<tbody>
<tr>
<td>adj</td>
<td>adj-adj addr</td>
</tr>
<tr>
<td>nom</td>
<td>nom-nom adj-nom</td>
</tr>
<tr>
<td>num</td>
<td>num-num adj-num</td>
</tr>
<tr>
<td>par</td>
<td>par-par adj-par</td>
</tr>
<tr>
<td>pro</td>
<td>pro-pro adj-pro</td>
</tr>
<tr>
<td>ver</td>
<td>ver-ver adj-ver</td>
</tr>
</tbody>
</table>

Each word class is arranged so that it has the same bit length during the message insertion process with the goal being that during the inverse or decryption process the stegotext can work well. The bit length for each word class is 8 bits for each selected style.

Table 2. Bit length in Style

<table>
<thead>
<tr>
<th>Code</th>
<th>Word Class</th>
<th>Bit Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>adj</td>
<td>adjectiva</td>
<td>8</td>
</tr>
<tr>
<td>adv</td>
<td>adverbial</td>
<td>8</td>
</tr>
<tr>
<td>nom</td>
<td>nomina</td>
<td>8</td>
</tr>
<tr>
<td>num</td>
<td>numeralia</td>
<td>8</td>
</tr>
<tr>
<td>par</td>
<td>partikel</td>
<td>8</td>
</tr>
<tr>
<td>pro</td>
<td>pronomina</td>
<td>8</td>
</tr>
<tr>
<td>ver</td>
<td>verba</td>
<td>8</td>
</tr>
</tbody>
</table>

3. RESULTS AND DISCUSSION

The default bit length representing the message length (n) is obtained from \( \frac{1929}{\log_2} \) with values rounded up.

\[
n = \frac{1929}{\log_2} = 11 \text{ bit}
\]

Representation of message length C in 11 bits (Cbit): 00000000100

The pattern used and the total bits that can hide the message in table 4 above:

<table>
<thead>
<tr>
<th>No.</th>
<th>ASCII message</th>
<th>Message length pesan (word)</th>
<th>(character)</th>
<th>(bits)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Dosen</td>
<td>1</td>
<td>15</td>
<td>40</td>
</tr>
<tr>
<td>2</td>
<td>Penelitian</td>
<td>2</td>
<td>19</td>
<td>152</td>
</tr>
<tr>
<td>3</td>
<td>STMIK Pelita Nusantara Medan</td>
<td>4</td>
<td>28</td>
<td>224</td>
</tr>
</tbody>
</table>

The following is an example of the original message that will be inserted into the system and analyzed for the capacity of the container media.

SIZER process of inserting the word “Dosen”:

If \( p + n \), so \( x = \left\lfloor \frac{p+n}{s} \right\rfloor \), so that \( r = (s * x) - (p + n) \)

1. \( x = \left\lfloor \frac{40 + 11}{16} \right\rfloor = \left\lfloor \frac{51}{16} \right\rfloor = 3 \)

The Random (R) string that is formed is: 0000001011010113

Stegobiner: Cbit | C | R
0000000101000-010001000110111011100110010101110011001110-00000001101101
(16 bit)

Stegoteks:

\[
\text{bade ampat. absurd ampat. apas asta. akta asta}
\]

2. \( x = \left\lfloor \frac{40 + 11}{24} \right\rfloor = \left\lfloor \frac{51}{24} \right\rfloor = 2 \)

It means the pattern (adj-num-par) will be used 3 times to hide the
message “Lecturer” so that \( r = (24 \times 3) - (40 + 11) = 72 - 51 = 21\)
The Random (R) string that is formed is:

\[
0000000110101 (21 \text{ bits})
\]

Stegoteks:

\[
\text{Cbit} \mid \text{C} \mid \text{R}
\]

1. x = \( \left\lceil \frac{40 + 11}{32} \right\rceil = \frac{51}{32} \approx 2 \) it means the pattern (pro-adj-nom-num) will be used 2 times to hide the message “Lecturer” so that \( r = (32 \times 2) - (40 + 11) = 64 - 51 = 13 \)
The Random (R) string that is formed is:

\[
0000000110101 (13 \text{ bits})
\]

Stegoteks:

\[
\text{Cbit} \mid \text{C} \mid \text{R}
\]

2. \( x = \left\lceil \frac{152 + 11}{24} \right\rceil = \frac{163}{24} \approx 7 \) it means the pattern (adj-num-par) will be used 7 times to hide the message “Lecturer” so that \( r = (24 \times 7) - (152 + 11) = 168 - 163 = 7 \)
The Random (R) string that is formed is:

\[
0000000110101 (7 \text{ bits})
\]

Stegoteks:

\[
\text{Cbit} \mid \text{C} \mid \text{R}
\]

3. \( x = \left\lceil \frac{152 + 11}{32} \right\rceil = \frac{163}{32} \approx 6 \) it means the pattern (pro-adj-nom-num) will be used 6 times to hide the message “Lecturer” so that \( r = (32 \times 6) - (152 + 11) = 182 - 163 = 29 \)
The Random (R) string that is formed is:

\[
0000000110101 (29 \text{ bits})
\]

Stegoteks:

\[
\text{Cbit} \mid \text{C} \mid \text{R}
\]

1. meaning that the pattern (adj-num) will be used 15 times to hide the message “STMIK Pelita Nusantara Medan” so that \( r = (16 \times 15) - (224 + 11) = 240 - 235 = 5 \)
The Random (R) string that is formed is:

\[
0000000110101 (5 \text{ bits})
\]

Stegoteks:

\[
\text{Cbit} \mid \text{C} \mid \text{R}
\]

2. \( x = \left\lceil \frac{224 + 11}{24} \right\rceil = \frac{235}{24} \approx 10 \) meaning that the pattern (adj-num-par) will be used 10 times to hide the message “STMIK Pelita Nusantara Medan” so that \( r = (24 \times 10) - (224 + 11) = 240 - 235 = 5 \)
The Random (R) string that is formed is:

\[
0000000110101 (5 \text{ bits})
\]

Stegoteks:
3. The pattern (pro-adj-nom-num) will be used 8 times to hide the message "STMIK Pelita Nusantara Medan" so that \( r = (32 \times 8) - (224 + 11) = 256 - 235 = 21 \) bits.

The Random (R) string that is formed is: 0000000110101 (13 bits)

Stegobiner: Cbit | C | R 0001100000-01010011010100010011010001001001010011001000010101100100110011001010011100110000101101010011011000101110101001001010010110010000001001101010010100101100000000010011010111001100111000101110010011000010111001011000010010000001001110011101010111001101100001011011100111010001100111100010100100100101001011001000010110111001000001011100101100001001000000100110101100101011001000110000101101110

Stegoteks:
badang argumentatif aforisme ampat. badang babil ablaut aneka. apakah antep aerograf asta. badang apik ambisius asta. badang awam afridisiak aku. badang adib abses aneka. badang adem arsip aneka. badang awet antung-antung aneka.

Based on the results of the data verification above, it is shown in the following table.

<table>
<thead>
<tr>
<th>Sentence Pattern</th>
<th>Style</th>
<th>Jawahir koma (prosen)</th>
<th>Stegoteks</th>
<th>Jawahir koma (prosen)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pro-adj-nom-num</td>
<td>1</td>
<td>(224 + 11) / 32</td>
<td>7</td>
<td>0000000110101</td>
</tr>
<tr>
<td>Pro-adj-nom-num</td>
<td>2</td>
<td>(235 - 6) / 32</td>
<td>9</td>
<td>11000110100010010100100101001001010100101001001010010010100100110010000101101010011011000101110101001001010010110010000001001101010010100101100000000010011010111001100111000101110010011000010111001011000010010000001001110011101010111001101100001011011100111010001100111100010100100100101001011001000010110111001000001011100101100001001000000100110101100101011001000110000101101110</td>
</tr>
</tbody>
</table>

The results of the steganographic capacity process based on sentence patterns in Indonesian indicate that based on the input data calculated on the number of words, has a significant influence on the results of stegotext. This can be seen from the comparison of the length of the input message which is calculated in the number of words with the length of the output message (stegoteks) which is influenced by the length of the style (pattern) of the sentence structure used. For input data with a word length of 1, the average capacity of the stegotext is 8 to 9 words or 12.4%, for input word length 2, the average capacity of the stegotext is 22 words or 8.85% and the input word length is 4 then the capacity of the stegotext is between 30 to 33 words or equivalent to 12.91%.

**4. CONCLUSION**

Based on the results of the analysis of the capacity of text-based steganography with Indonesian sentence patterns, steganography using the Indonesian Pattern Sentence method (IPSM) is feasible to be used as one of the text-based steganographic models. The capacity of text-based steganography depends on the sentence pattern used. If the sentence pattern used is getting longer, the
capacity of text-based steganography (which is calculated in the number of words) will be smaller.

5. CLOSING
We would like to thank the STMIK Pelita Nusantara institution which has been facilitated through LPPM in carrying out this research.

BIBLIOGRAPHY


Contact person Author: R. Fanry Siahaan
Hp : 081265815557