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

R. Fanry Siahaan*[1], Amran Sitohang[2], Ibnu Febrian[3], Widia Putri[4]
[1,2,3,4] STMIK 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. have 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 rate compressed messages AbS-LPC.[3]and[10]. research in text-based steganography in the field of information hiding because there is no modification to the text carrier that will increase the steganography hiding because the parity feature shows a uniform distribution in each character[2],[11]and[12].

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:

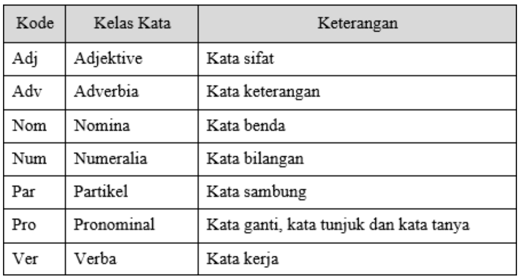
a. Research sites

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 1. Word Class



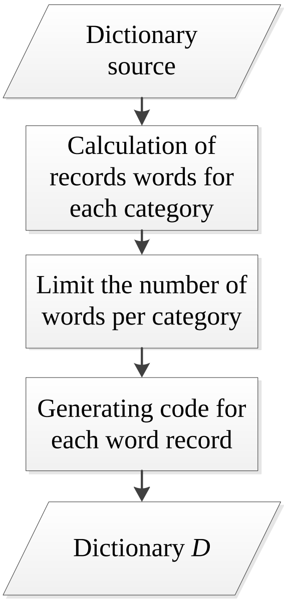
Kode	Kelas Kata	Keterangan
Adj	Adjektive	Kata sifat
Adv	Adverbia	Kata keterangan
Nom	Nomina	Kata benda
Num	Numeralia	Kata bilangan
Par	Partikel	Kata sambung
Pro	Pronominal	Kata ganti, kata tunjuk dan kata tanya
Ver	Verba	Kata kerja

Sumber: KBBI, 2008

c. Preprocess

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. The number of words in each type of word (f) is limited to a number to the power of two by calculation

g =$\lfloor 2^{\log_2 f} \rfloor$…………………………….. (1)

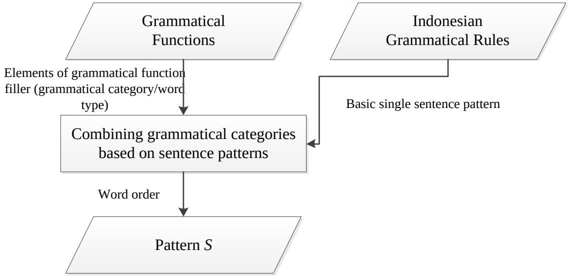


Dictionary source
Calculation of records words for each category
Limit the number of words per category
Generating code for each word record
Dictionary *D*

Figure 1. D Dictionary Development Process

In this researchpatternThe 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] namely SP, SPK, SPO, SPO-Pel, SPO-Pel-K, SPOK, SP-Pel, and SP-Pel-K.

The data that has been collected is then analyzed for data to filter (cleaning and filtering) appropriate data to be used as raw data to be processed.



Grammatical Functions
Indonesian Grammatical Rules
Elements of grammatical function filler (grammatical category/word type)
Basic single sentence pattern
Combining grammatical categories based on sentence patterns
Word order
Pattern *S*

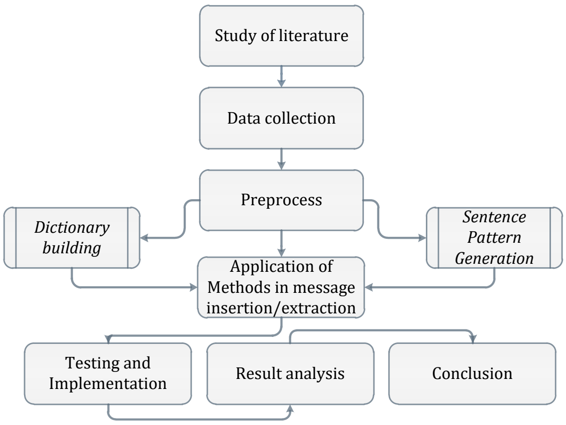
Figure 2. The process of building the S . pattern

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

SIZERR(C)+ C + R……………………(2)

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 c (Stegoteks). The resulting stegottext 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:

$$DESIZER\left(EKSTRACT_D\left(EMBED_{D,S}(SIZER_S(C))\right)\right) = C. \text{.....(3)}$$



Study of literature
Data collection
Preprocess
Dictionary building
Sentence Pattern Generation
Application of Methods in message insertion/extraction
Testing and Implementation
Result analysis
Conclusion

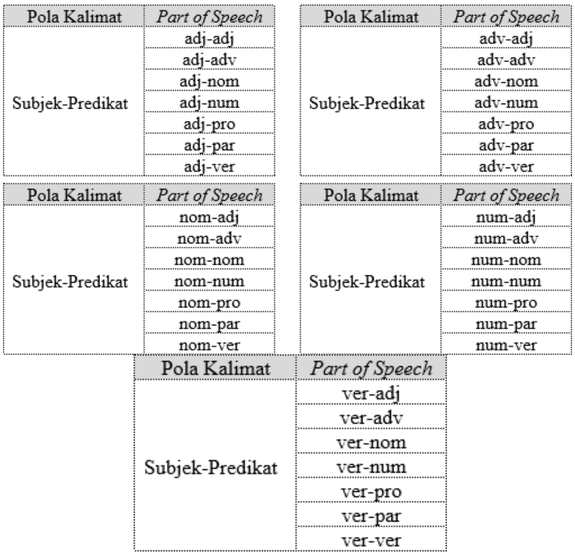
Figure 3. Research stages

2. Data Analysis Techniques

The pattern generation process begins with defining a grammatical function. Each function is filled with elements in the form of grammatical categories.CategoryThe grammatical function 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



Pola Kalimat	*Part of Speech*
Subjek-Predikat	adj-adj
	adj-adv
	adj-nom
	adj-num
	adj-pro
	adj-par
	adj-ver

Pola Kalimat	*Part of Speech*
Subjek-Predikat	adv-adj
	adv-adv
	adv-nom
	adv-num
	adv-pro
	adv-par
	adv-ver

Pola Kalimat	*Part of Speech*
Subjek-Predikat	nom-adj
	nom-adv
	nom-nom
	nom-num
	nom-pro
	nom-par
	nom-ver

Pola Kalimat	*Part of Speech*
Subjek-Predikat	num-adj
	num-adv
	num-nom
	num-num
	num-pro
	num-par
	num-ver

Pola Kalimat	*Part of Speech*
Subjek-Predikat	ver-adj
	ver-adv
	ver-nom
	ver-num
	ver-pro
	ver-par
	ver-ver

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

Code	Word Class	Bit Length
adj	adjectiva	8
adv	adverbial	8
nom	nomina	8
num	numeralia	8
par	partikel	8
pro	pronomina	8
ver	verba	8

3. Research Sample

The data used is sourced from the Big Indonesian Dictionary in 2008, as shown in the following table.

Table 3 Number of words from KBBI used

Word Class	Total Word Dictionary
Nomina	1.265
Pronomina	4
Adjectiva	232
Adverbia	14
Verba	368
Numeralia	5
Partikel	41
Total:	**1.929**

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.

Table 4 The data used in the study

No.	ASCII message	Message length pesan		
		(word)	(character)	(bits)
1	Dosen	1	15	40
2	Penelitian Internal	2	19	152
3	STMIK Pelita Nusantara Medan	4	28	224

3. RESULTS AND DISCUSSION

The default bit length representing the message length (n) is obtained from $\frac{Total\ Kamus\ kata}{\log 2}$with values rounded up.

n=$\left\lceil \frac{1929}{\log 2} \right\rceil = 11$bit

Representation of message length C in 11 bits (Cbit): 00000001000

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

a. adj(8) + num(8) = 16 bit
b. adj(8) + num(8) + par(8) =24 bit
c. pro(8) + adj + nom(8) + num(8) =32 bit

SIZER process of inserting the word "Dosen":

If p + n s, so$x = \left\lceil \frac{p+n}{s} \right\rceil$, so that r = (s * x) – (p + n)

1. $x = \left\lceil \frac{40+11}{16} \right\rceil = \left\lceil \frac{51}{16} \approx 4 \right\rceil$it means the pattern (adj-num) will be used 4 times to hide the message "Lecturer" so that r = (16 * 4) – (40 + 11) = 64-51 =13
   The Random (R) string that is formed is: 0000000110101 (13 bits)

Stegobiner: Cbit | C | R
00000101000-
0100010001101111011100110110010101101110-
0000000110101 (16 bit)
Stegoteks:
bade ampat. absurd ampat. apas asta. akta asta

2. $x = \left\lceil \frac{40+11}{24} \right\rceil = \left\lceil \frac{51}{24} \approx 3 \right\rceil$ it means the pattern (adj-num-par) will be used 3 times to hide the

message “Lecturer” so that r = (24 * 3) – (40 + 11) = 72-51 =21

The Random (R) string that is formed is: 0000000110101 (21 bits)

Stegobiner: Cbit | C | R
00000101000-0100010001101111011100110110010101101110-000000000000000111110 (21 bit)

Stegoteks:
aus asta aho. afirmatif asta adakan. afektif aku adakan

3. $x = \left\lceil \frac{40+11}{32} \right\rceil = \left\lceil \frac{51}{32} \approx 2 \right\rceil$ it means the pattern (pro-adj-nom-num) will be used 2 times to hide the message “Lecturer” so that r = (32 * 2) – (40 + 11) = 64-51 =13

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

Stegobiner: Cbit | C | R
00000101000-0100010001101111011100110110010101101110-0000110011100 (13 bit)

Stegoteks:
mengapa absurd aci-acian asta. mengapa ajun angguk ampat

SIZER process of inserting the word “Penelitian Internal”:

1. $x = \left\lceil \frac{152+11}{16} \right\rceil = \left\lceil \frac{163}{16} \approx 11 \right\rceil$ meaning the pattern (adj-num) will be used 11 times to hide the message “Intenal Research” so r = (16 * 11) – (152 + 11) = 176-163 =13

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

Stegobiner: Cbit | C | R
00010011000-0101000001100101011011100110010101101100011010010111010001101001011000010110111000100000010010010110111001110100011001010111001001101110011000-0101101100000 (13 bit)

Stegoteks:
argumentatif ampat. ala-bihalal asta. ambring-ambringan ampat. alpa ampat. bade aneka. anasional ampat. antesis aku. abasah asta. aneka ragam aku. bacar ampat. azmat asta

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

   The Random (R) string that is formed is: 0000000110101 (7 bits)

Stegobiner: Cbit | C | R
00010011000-010100000110010101101110011001010110110001101001011101000110100101100001011011100010000001001001011011100111010001100101011100100110111001100001011011-0010011 (7 bit)

Stegoteks:
antun aku apa. asin asta awat. abrar aku bahwasanya. apas asta ah. agung ampat atau. abu-abu aneka ajak. babar ampat ai

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

   The Random (R) string that is formed is: 0000000110101 (29 bits)

Stegobiner: Cbit | C | R
00010011000-0101000001100101011011100110010101101100011010010111010001101001011000010110111000100000010010010110111001110100011001010111001001101110011000010110110
0-00000000000000000001000101001 (29 bit)

Stegoteks:
badang aktual antraks ampat. mengapa bacak androlog aku. badang akrofobia audivus asta. apakah awah advertensi aneka. apakah akustik anafilaksis ampat. badang asor anateksis ampat

SIZER process for inserting the word “STMIK Pelita Nusantara Medan”:

1. meaning that the pattern (adj-num) will be used 15 times to hide the message "STMIK Pelita Nusantara Medan" so that r = (16 * 15) – (224 + 11) = 240-235 =5

   The Random (R) string that is formed is: 0000000110101 (5 bits)

Stegobiner: Cbit | C | R
00011100000-01010011010101000100110101001001010010110010000001010000011001010110110001101001011101000110000100100000010011100111010101110011011000010110111001110100011000010111001001100001001000000100110101100101011001000110000101101110-01111 (5 bit)

Stegoteks:
akrobatik ampat. antun ampat. antagonis ampat. antisosial asta. asih ampat. akut asta. angit asta. amorf asta. afektif asta. aktif asta. asin aku. automatis asta. asyik asta. asepsis aneka. acak-acakan ampat

2. $x = \left\lceil \frac{224+11}{24} \right\rceil = \left\lceil \frac{235}{24} \approx 10 \right\rceil$ meaning that the pattern (adj-num-par) will be used 10 times to hide the message "STMIk Pelita Nusantara Medan" so that r = (24 * 10) – (224 + 11) = 240-235 =5

   The Random (R) string that is formed is: 0000000110101 (5 bits)

Stegobiner: Cbit | C | R
00011100000-010100110101010001001101010010010100101100100000010100000110010101101100011010010111010001100001001000000100111001110101011100110110000101101110011101000110000101110010

01100001001000000100110101100101011001000110000101101110-00110 (5 bit)
Stegoteks:
awal aku akan. aneka jenis asta arkian. aboral aneka alhasil. ayu asta antar. anggak ampat amin. anggal aku atau. adaptif ampat abong-abong. abstrak aneka awat. amorf ampat bahwa. akademik ampat apalagi

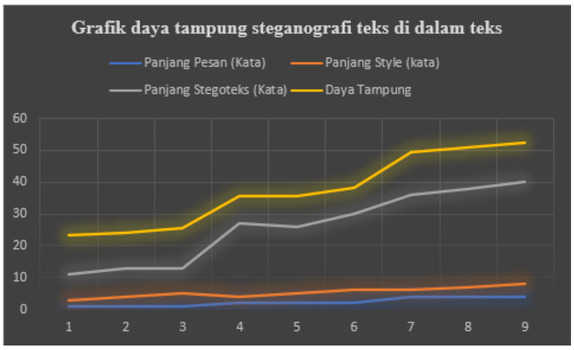
3. $x = \left\lceil \frac{224+11}{32} \right\rceil = \left\lceil \frac{235}{32} \approx 8 \right\rceil$ meaning that the pattern (pro-adj-nom-num) will be used 8 times to hide the message "STMIK Pelita Nusantara Medan" so that r = (32 * 8) – (224 + 11) = 256-235 =21
   The Random (R) string that is formed is: 0000000110101 (13 bits)

Stegobiner: Cbit | C | R
00011100000-
01010011010101000100110101001001010010110010000001010000011001010110110001101001011101000110000100100000010011100111010101110011011100001011011100111010001100001011100100110000100100000010011010110010101100100011000010110111-0000000000010101011001 (21 bit)
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 5. analysis of the capacity of stegotext messages

Pesan	Style	Jumlah kata (pesan)	Stegoteks	Jumlah kata (stegoteks)
Dosen	Adj-num	1	*bade ampat. absurd ampat. apas asta. akta asta*	8
	Adj-num-par	1	*aus asta aho. afirmatif asta adakan. afektif aku adakan*	9
	Pro-adj-nom-num	1	*mengapa absurd aci-acian asta. mengapa ajun angguk ampat*	8
Penelitian Internal	Adj-num	2	*argumentatif ampat. ala-bihalal asta. ambring-ambringan ampat. alpa ampat. bade aneka. anasional ampat. antesis aku. abasah asta. aneka ragam aku. bacar ampat. azmat asta*	23
	Adj-num-par	2	*antun aku apa. asin asta awat. abrar aku bahwasanya. apas asta ah. agung ampat atau. abu-abu aneka ajak. babar ampat ai*	21
	Pro-adj-nom-num	2	*badang aktual antraks ampat. mengapa bacak androlog aku. badang akrofobia audivus asta. apakah awah advertensi aneka. apakah akustik anafilaksis ampat. badang asor anateksis ampat*	24
STMIK Pelita Nusantara Medan	Adj-num	4	*akrobatik ampat. antun ampat. antagonis ampat. antisosial asta. asih ampat. akut asta. angit asta. amorf asta. afektif asta. aktif asta. asin aku. automatis asta. asyik asta. asepsis aneka. acak-acakan ampat*	30
	Adj-num-par	4	*awal aku akan. aneka jenis asta arkian. aboral aneka alhasil. ayu asta antar. anggak ampat amin. anggal aku atau. adaptif ampat abong-abong. abstrak aneka awat. amorf ampat bahwa. akademik ampat apalagi*	31
	Pro-adj-nom-num	4	*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*	32



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

[1] J. Y. Babys, Kusrini, and Sudarmawan, "Analisis Aspek Keamanan Informasi Jaringan Komputer (Studi Kasus : STIMIK Kupang)," *Semin. Nas. Inform. 2013*, 2013.

[2] K. Wang and Q. Gao, "A Coverless Plain Text Steganography Based on Character Features," *IEEE Access*, 2019, doi: 10.1109/ACCESS.2019.2929123.

[3] J. Yang, P. Liu, and S. Li, "A common method for detecting multiple steganographies in low-bit-rate compressed speech based on bayesian inference," *IEEE Access*, 2019, doi: 10.1109/ACCESS.2019.2939629.

[4] N. Wu *et al.*, "STBS-Stega: Coverless text steganography based on state transition-binary sequence," *Int. J. Distrib. Sens. Networks*, 2020, doi: 10.1177/1550147720914257.

[5] X. Duan, D. Guo, N. Liu, B. Li, M. Gou, and C. Qin, "A New High Capacity Image Steganography Method Combined with Image Elliptic Curve Cryptography and Deep Neural Network," *IEEE Access*, 2020, doi: 10.1109/ACCESS.2020.2971528.

[6] Y. Huang, C. Liu, S. Tang, and S. Bai, "Steganography integration into a low-bit rate speech codec," *IEEE Trans. Inf. Forensics Secur.*, 2012, doi: 10.1109/TIFS.2012.2218599.

[7] N. Hamid, A. Yahya, R. B. Ahmad, and O. M. Al-Qershi, "Image steganography techniques: an overview," *Int. J. Comput. Sci. Secur.*, 2012.

[8] M. Taleby Ahvanooey, Q. Li, J. Hou, H. Dana Mazraeh, and J. Zhang, "AITSteg: An innovative text steganography technique for hidden transmission of text message via social media," *IEEE Access*, 2018, doi: 10.1109/ACCESS.2018.2866063.

[9] B. Gupta Banik and S. K. Bandyopadhyay, "Novel Text Steganography Using Natural Language Processing and Part-of-Speech Tagging," *IETE J. Res.*, 2020, doi: 10.1080/03772063.2018.1491807.

[10] R. Gawade, P. Shetye, V. Bhosale, and P. N. Sawantdesai, "Data Hiding Using Steganography For Network Security," *Int. J. Adv. Res. Comput. Commun. Eng.*, 2014.

[11] S. Edward Jero and P. Ramu, "Curvelets-based ECG steganography for data security," *Electron. Lett.*, 2016, doi: 10.1049/el.2015.3218.

[12] A. A. Fikhri and H. Hendrawaty, "IMPLEMENTASI STEGANOGRAFI TEXT TO IMAGE MENGGUNAKAN METODE ONE BIT LEAST SIGNIFICANT BIT BERBASIS ANDROID," *J. Infomedia*, 2018, doi: 10.30811/jim.v3i1.623.

[13] H. Annur, "Klasifikasi Masyarakat Miskin Menggunakan Metode Naive Bayes," *Ilk. J. Ilm.*, 2018, doi: 10.33096/ilkom.v10i2.303.160-165.

[14] I. G. N. K. Putrayasa, "Jenis - Jenis dan Pola Kalimat Bahasa Indonesia," *Https://Repositori.Unud.Ac.Id/Protected/Storage/Upload/Repositori/C5Af5469574856E21718C34882583925.Pdf*, 2016.

[15] A. Hasan, "Tata Bahasa Baku Bahasa Indonesia," *Dep. Pendidik. dan Kebud. Republik Indones.*, 2007.

Contac person Author: R. Fanry Siahaan
Hp : 081265815557