International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 12, No.1 (Dec-2022)

https://dx.doi.org/10.12785/ijcds/1201125

Steganalysis of Markov Chain-Based Statistical Text Steganography

Nujud Alghamdi [1], Lamia Berriche [2], and Maha Alrabiah [1]

[1]Computer Science Department Al-Imam Mohammad Ibn Saud Islamic, Riyadh, Kingdom of Saudi Arabia
[2]Computer Science Department Prince Sultan University, Riyadh, Kingdom of Saudi Arabia

Received 22 Apr. 2021, Revised 27 Jul. 2022, Accepted 24 Aug. 2022, Published 31 December 2022

Abstract: Text steganography is the art of hiding a secret message in a text. Conversely, text steganalysis is the art of detecting a hidden message in a text. In this work, we studied the detectability performance of a Markov chain (MC) based statistical text steganography technique. We started by analyzing Arabic texts of different types: economy, sports, international news. Then, the MC-based encoder was used to hide Arabic messages of various lengths. Subsequently, we extracted specific features from the stego-texts and the natural texts and applied them to a support vector machine (SVM) classifier. We noticed that detectability depends on the cover message type, the length of the concealed message, the embedding rate, and the extracted features. We noticed that lower the embedding rate and the smaller the text-size, less accurate is the classification. Moreover, the accuracy of an SVM classifier was less than 67% for 1 KB stego-texts generated with Arabic economy or sports cover texts with an embedding rate of 4 bits per word (bpw). Besides, more than 62% of stego-texts were classified as natural texts for 1 KB text-sizes when we considered the word distribution feature.

Keywords: Steganography; steganalysis; Arabic text steganography; statistical steganography; Markov chains; text generation.

1. INTRODUCTION

Information security systems aiming to protect sensitive data are divided into cryptography and information hiding. Cryptography converts a plain-text into a cipher, preventing unauthorized access to its content. Conversely, information hiding conceals a secret message in a cover medium which makes it invisible. Information hiding can be classified into watermarking and steganography. Watermarking entails providing proof of authenticity, while the aim of steganography is the unnoticeable transmission of secret information in a cover medium [1], [2], [3]. The cover medium can be an image, an audio file, a video, or text. In steganography, the message to be hidden is called the secret message, and the medium used to hide it is called the cover medium. The stego-object is the cover medium holding the secret message, and the stego-key is the key used to encode the secret message by the sender, to be decoded at the receiver's side. Text is the most widely used information carrier in daily life; using text for information hiding has attracted many researchers [4], [5], [6], [7], [8], [9]. Text steganography is divided into three techniques: format-based, linguistic-based, and statistical-based. In a format-based approach, text format such as spacing, line shifting is exploited to embed the secret message [8]. Linguistic-based approaches rely on modifying some linguistic properties of the text such as syntax and semantic properties [10], whereas statistical-text steganography is based on the generation of a stego-text mimicking the statistical properties of the cover text [7], [11].

On the contrary, steganalysis aims to detect a stego-text from a cover text [3] by extracting some text features such as word distribution [12], entropy [13], the correlation between words [14], and other features [15], [16]. Then these features are analyzed to determine whether the text contains a secret message.

In [11], we proposed a Markov chain (MC) based steganographic system for the Arabic language. We investigated the system's performance with regard to its embedding capacity. In this work, we focus on the

E-mail address: nujoud_ali@hotmail.com, lberriche@psu.edu.sa, msalrabiah@imamu.edu.sa

undetectability property based on our work in [11] and study the different features of our dataset.

This paper is organized as follows: Section II gives a short overview of the MC encoder proposed in [11]. Section III provides an overview of text steganalysis techniques. Section IV presents the proposed method. Section V gives the results of the steganalysis techniques. And finally, section VI concludes this paper.

2. MARKOV-CHAIN BASED ENCODER

In this part, we will briefly describe our encoder that was previously described in [11]. The main objective of the steganography system encoder is to generate a stego-text out of an input text that is hidden using an MC of an Arabic corpus; the statistical analysis of an Arabic text generates its MC, which is used to encode the Huffman encoded secret message. Hence, a stego-text mimicking the statistical properties of the Arabic corpus text is generated.

The secret message is first encoded with a Huffman code. Next, it is divided into n-bit sized blocks and each block is converted into a decimal number within the range R = $[0, 2^n - 1]$. The basic function for both encoding and decoding processes is sub-ranging, which divides recursively the range R. The sub-ranging is done such that each outbound is assigned a subrange $r_{s_i s_j}$ based on its transition probability $p_{s_i s_j}$ using below equation [7].

$$\frac{\text{length } (r_{s_i s_j})}{\text{length } (R)} \approx p_{s_i s_j} \qquad (1)$$

For each decimal number, representing a block of bits of the secret message, the edges of the MC holding an interval containing it are traversed till an interval of length one is reached. The encoder converts each decimal number to the sequence of traversed successive states.

The subrange $r_{s_i s_j}$ should be of a minimal length one and $\sum r_{s_i s_j}$ = R.

For example, in Figure 1, decimal number 7 is encoded to the sequence of states {S0 S3 S8}.

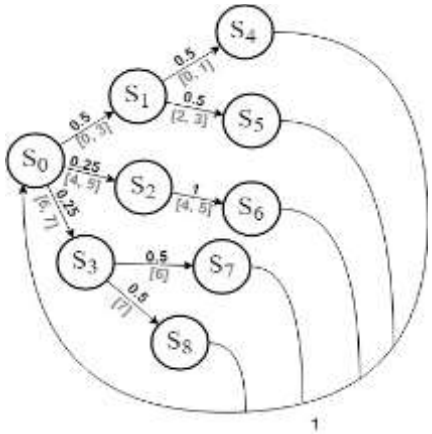

Figure 1: Example of MC State Diagram with Ranges.

3. STEGANALYSIS

Steganalysis is the science of detecting whether the information is hidden in a cover or not [3]. Linguistic steganalysis methods are generally based on the linguidtic features of natural language texts [12], [17]. They are based on the use of natural text features to classify text chunks as natural or not. The closer the statistical features of the stego-text to the natural text, the more undetectable is the stego-text.

Linguistic and statistical steganography methods need efficient detection algorithms due to the diversity of syntax and the ambiguity of semantics. In [12], the authors presented a novel statistical-linguistic steganography detection algorithm. They used the word distribution feature in the text segment, conducting their experiments on three linguistic steganography methods: NICETEXT, TEXTO, and a MC-based method using an SVM classifier. They used three corpora: a Charles Dickens corpus, a Bad corpus containing texts generated from the three methods, and a third corpus, the S-corpus, extracted from novels written by some novelists whose last name begins with the letter "S". The training data set is composed of Charles Dickens corpus and the Bad corpus stego-texts. They used the S-corpus and the Bad corpus to generate the testing data set. They reached a 99.57% accuracy for 40 KB segment sizes and 87.39% accuracy for segment sizes of 5 KB. They improved their detection accuracy in [13] where they used two different classification attributes: a variable of information entropy-like and its variance. This method focuses on detecting text segments with a small size. By experimenting with the three different linguistic steganography methods: NICETEXT, TEXTO, and MC-based, the accuracy exceeded 90% for segments with a size not exceeding 5 KB.

In [14], the statistical characteristics of correlations between the general service words such as n-window mutual information are used to classify the given text segments into stego-text segments and normal text segments, by an SVM classifier. The classifier accuracy reached 94.01%, 98.48%, and 97.96% for MC-based English stego-texts, NICETEXT, and TEXTO for 20 KB size texts. They noticed that 11% of MC-based stego-texts were classified as natural. On the other hand, all stego-texts generated by NICESTEXT and TEXTO were detected correctly.

In [18], the authors used the perplexity feature for steganalysis. They built a Good corpus from The New York Times and a Bad corpus from texts generated by the NICETEXT system. The training data set contained Bad corpus texts and the testing data set contained both Good and Bad texts. Their accuracy exceeded 99% when text size is 400 bytes.

In [15], the authors presented a method based on immune clone mechanism and meta-features. The mechanism of the immune clone selects the optimal and

effective features among meta-features to establish an effective detector. They tested their detector on NICETEXT, TEXTO, and MC-based English stego-texts. The accuracy of detecting stego-texts exceeded 88% for 1 KB text size, 95% for 3 KB texts, and 98% for 5 KB texts.

In recent years, deep neural network technology has gained an interest in Natural Language Processing. Neural network stego-text generation appeared in [9] and [19]. Traditional steganalysis techniques are not effective with deep learning-based steganography. On the other hand, deep learning-based steganalysis showed promising results; in [19], the authors proposed a text steganalysis method based on word correlation features that are fed to a softmax classifier. They trained their model by using the text steganalysis dataset called the T-Steg. They attained a high detection accuracy which reached 99.9% for a 5 bpw.

4. METHODOLOGY

First, we started by preprocessing the collected texts from the Watan-2004 corpus. Then, we generated the MC of different texts. The MC-based encoder was then used to hide Arabic messages. Moreover, we extracted specific features from the stego-texts and the natural texts and applied them to an SVM classifier.

A. Dataset Collection and Preprocessing

For our dataset, we used the Watan-2004 corpus [20], which is a collection of online newspaper texts collected from thousands of articles written in modern standard Arabic. We selected three topics: economy, sports, and international news. Each topic corpus has almost one million words, as shown in Table I. We preprocessed each corpus by removing repeated punctuations and diacritics if any.

Table I. Corpora details [20].

Corpus name	Size in bytes	Size in words
Economy	10325 KB	969271
Sports	10119 KB	979887
International news	7445 KB	709126

Then, we extracted bigrams and their frequencies from the cover texts to generate the Markov models.

Each corpus was divided into two sets: natural texts (NTs) set and another set which was used for generating stego-texts (STs). Also, the NTs and STs datasets were divided into four groups of texts of different sizes: 1 KB, 3 KB, 5 KB, and 9 KB.

B. Feature Extraction

To distinguish between NTs and STs, we have extracted 8 different features; complexity, variety, entropy, Standard Deviation (SD) of the entropy, entropy-like [13], the variance of the entropy-like, word distribution [12], and variance of the word distribution [12].

The complexity measures the average length of sentences in a text and could be calculated as [21]:

$$\text{Complexity} = W \times \log(M) \qquad (2)$$

where W is the average word length in characters and M is the average length of the sentence in words.

The variety measures the variation of expressions in a text [21]. It is a type to token ratio. In linguistics, tokens represent all words in a text while types represent distinct words. Variety could be formulated as [21]:

$$\text{Variety} = t / \log(k) \qquad (3)$$

where t is the number of types and k is the number of tokens. Variety shows the size of vocabulary used in a text. Complexity and variety measures are related to the nature of the text [21]. Variety is low in scientific texts where the content is more important than the writing style and becomes high in the texts that have different expressions and use synonyms [21] (e.g. Arabic poetry or literature).

In the definition of information theory, first introduced by Shannon [22], entropy in texts represents the amount of information in a text. The higher the entropy of a given text the more informative it is. Information entropy is expressed as:

$$H(x) = -\sum_{i=0}^{n} p(x_i) \log p(x_i) \qquad (4)$$

The entropy-like called also Detection Entropy (DE) was first introduced by [13] to improve the detectability of linguistic steganography. The entropy-like is given by:

$$DE = -\sum_{i=0}^{n-1} S_i \log S_i \qquad (5)$$

where S_i represents the score of a word i. The score is computed as:

$$S_i = \frac{1}{C} \sum_{k}^{n_i} k \qquad (6)$$

Where n_i represents the number of occurrences of the word i and C represent the total number of occurrences of all words. The main difference between Shannon entropy and entropy-like is that the score function amplifies the k^{th} occurrence of the word. A small variation in the occurrence affects the entropy-like more than Shannon entropy. The result being that even a little change in the distribution of word frequencies can cause a great difference in the score.

The Average Word Distribution (AWD) is given by [12]:

$$AWD = \sum_{i=0}^{m} WD(w_i)\ \frac{n_i}{n} \qquad (7)$$

where $WD(w_i) = \frac{1}{n_i}\sum_{k=0}^{n_i}(w_{i_{lk}} - \frac{1}{n_i}\sum_{k}^{n_i} w_{i_{lk}})^2$, which is the variance of the word w_i location, with $w_{i_{lk}}$ representing the k[th] location of the word w_i.

5. RESULTS AND DISCUSSION

Table II and Table III give the average values of the used features for different text sizes - 1 KB, 3 KB, 5 KB, and 9 KB. We notice that the values of features are closer to each other for small text sizes. The feature that best discriminates between NTs and STs is the variety. We notice also that the entropy of natural texts is higher than the entropy of stego-texts, but the entropy-like feature of STs is higher than that of NTs.

Table II. Average Features Values for Different Texts' Sizes.

Text size (KB)	Complexity		Variety		Entropy	
	NTs	*STs*	*NTs*	*STs*	*NTs*	*STs*
1	7.54	7.29	39.78	36.31	6.17	5.95
3	7.84	7.75	84.35	71.26	7.4	7
5	7.96	7.57	123.38	95.62	7.99	7.36
9	7.85	7.51	194.84	131.46	8.67	7.73

Table III: Average Features Values for Different Texts' Sizes

Text size (KB)	SD (Entropy)		Entropy-like		Variance of (entropy-like)	
	NTs	*STs*	*NTs*	*STs*	*NTs*	*STs*
1	6.12	5.91	6.79	6.89	38.40	35.84
3	7.43	7.06	9.24	9.60	59.46	54.32
5	8.05	7.47	10.73	10.94	72.35	64.41
9	8.77	7.9	12.70	12.70	89.93	79.51

A. Detectability Analysis

In this section, we studied the detectability of the generated stego-texts. Our steganalysis results are measured via accuracy and False Negative Rate (FNR) [23].

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \qquad (8)$$

$$FNR = \frac{FN}{FN+TP} \qquad (9)$$

where True Positives (TP) are the STs predicted as STs, True Negatives (TN) are the NTs predicted as NTs, False Positives (FP) are the NTs predicted as STs and False Negatives (FN) are STs predicted as NTs. Therefore, the FNR calculates the FN from the actual STs.

First, we studied the impact of increasing the ST size on the detectability property. Then, we studied the effect of increasing the embedding capacity on detectability. After that, we compared our encoder detectability property to state of art steganography encoders [12], [13].

1) Effect of Stego-Text Size on Detectability

This experiment was elaborated based on four features - the complexity, the variety, the entropy, and the standard deviation of the entropy. We computed both the accuracy and the FNR of the SVM classifier for different text sizes. From Table IV we notice that when the size of the dataset is increased, the STs are more detectable. **Error! Reference source not found.** show the effect of different NTs and STs sizes on the detectability based on Table II and Table III. We can notice the increase of the detectability of STs and NTs as the text size increases. Nevertheless, we can also observe that FNR is high in 1K. A high FNR means that many STs were considered NTs.

Table IV: Comparison of Detectability Results for Different Text Sizes.

Text size (bytes)	Accuracy	FNR
1 K	73.33%	37.10%
3 K	90%	8.16%
5 K	94.44%	6.12%
9 K	97.78	0%

2) Effect of Text Type on Detectability

In Table V: Comparison of Detectability for Different Text Types

Embedding Rate	Accuracy	FNR
4 bpw	72.22%	38.78%
8 bpw	73.33%	36.73%
12 bpw	80%	24.39%
Embedding Rate	**Accuracy**	**FNR**

, we compared the detectability of our MC encoder for different text types - economy, sports, and news. We noticed that the lowest accuracy, which is 66.67%, is obtained for 1KB text size for the economy and sports texts, where the FNR was 53.84%. However, for greater text sizes the FNR per text type is greater than the FNR with a dataset of mixed text types.

Table V: Comparison of Detectability for Different Text Types

Embedding Rate	Accuracy	FNR
4 bpw	72.22%	38.78%
8 bpw	73.33%	36.73%
12 bpw	80%	24.39%
Embedding Rate	**Accuracy**	**FNR**

.

3) Effect of the Embedding Rate on Detectability

In this part, we studied the effect of increasing the embedding rate on the stego-text detectability for 1KB texts.

The experiment was elaborated on 300 texts (70% training and 30% testing). We found that when the embedding capacity rate measured in bit per word (bpw) was increased, the STs is more detectable. Embedding fewer bits in a word makes the stego-text more similar to the natural text, as shown in Table VI.

Table VI: Effect of the Embedding Rate on Detectability

Embedding Rate	Accuracy	FNR
4 bpw	72.22%	38.78%
8 bpw	73.33%	36.73%
12 bpw	80%	24.39%

B. Comparison with other Markov Chain based Encoders

In this section we compare our MC encoder detectability performance to other MC encoders investigated in [12] and [13]. In [12], the authors studied the detectability performance of an MC-based encoder proposed in [24] using the word distribution feature [12]. In [13], the authors studied the detectability performance of the same MC-based encoders using an entropy-like and its variance features. We used the same features as inputs for our SVM classifier. First, we compared the detectability performance of our encoder for the two models. Then we compared our MC encoder to the ones studied in [12] and [13].

We notice from Table VII that the detection using the entropy-like feature is more effective than using word distribution; Also, for a small-sized text, the majority of stego-texts, 62.5% to be precise, are not correctly detected.

Table VII. Results of Applying Different Features

Text type	Text size	Accuracy (%)	FNR (%)
Economy	1 K	66.67	53.84
	3 K	80	7.69
	5 K	83.33	0
Sport	1 K	66.67	53.84
	3 K	90	0
	5 K	100	0
International	1 K	83.33	38.46
	3 K	96.67	0
	5 K	100	0

In **Table VIII**, we compared the detectability of our MC encoded stego-text to the encoder used in [13] based on the entropy-like and its variance. We noticed that the classifier gives higher accuracy on our stego system. However, the FNR would be more appropriate for this comparison. Also, as noticed previously, embedding fewer bits per word makes the stego-text less detectable. Our embedding rate for the experiment was 8 bpw but [13] did not provide the used embedding rate. On the other hand, the comparison with [12] in Table IX shows that our encoding is less detectable when the word distribution feature is used.

Table VIII. Comparison of our Classification Results with [13].

	Accuracy	FNR
Size	***Entropy-like***	***Word Distribution***
1 K	74.76%	62.85%
3 K	88.09%	71.90%
5 K	91.42%	76.67%
	Accuracy	**FNR**

Table IX. Comparison of our Classification Results with [12].

	Accuracy	
Size	***[12]***	***Ours***
5 K	87.39%	76.67%

6. CONCLUSION AND FUTURE WORK

In this paper, we investigated the detectability of stego-texts generated by the MC-based encoder proposed [11]. We noticed that the detectability depends on the text type, the stego-text size, and the embedding rate. Hiding a short message in a sports or economy text is less detectable than concealing it in an international news cover. Also, the larger the stego-text or the embedding rate, the higher is the detection accuracy. Besides, detectability depends on the extracted features; our proposed steganography technique is less detectable when the word distribution feature is used. In the future, the performance of the proposed technique in [11] will be considered for different languages, as well as the perceptibility property.

ACKNOWLEDGMENT

The authors would like to thank the scientific government institution, King Abdulaziz City for Science and Technology (KACST), for financial support [grant number: 1-17-02-008-0002].

REFERENCES

[1] M. P. Uddin, M. Saha, S. J. Ferdousi, "Developing an efficient solution to information hiding through text steganography along with cryptography," in *the 9th International Forum on Strategic Technology (IFOST)*, Cox's Bazar, 2014.

[2] R. Mishra and P. Bhanodiya, "A review on steganography and cryptography," in *International conference onAdvances in Computer Engineering and Applications (ICACEA)*, Ghaziabad, 2015.

[3] M. T. Ahvanooey et al, "Modern Text Hiding, Text Steganalysis, and Applications: A Comparative Analysis," *Entropy,* pp. 350-381, 2019.

[4] M. Chapman, "Hiding the hidden: A software system for concealing ciphertext as innocuous text," Wisconsin-Milwaukee Univ., USA, 1998.

[5] A. Desoky, "Listega: list-based steganography methodology," *International Journal of Information Security,* vol. 8, no. 4, p. 247–261, 2009.

[6] E. Satir and H. Isik, "A Huffman compression based text steganography method," *Multimedia Tools and Applications,* vol. 70, no. 3, p. 2085–2110, 2014.

[7] H. Moraldo, *An Approach for Text Steganography Based on Markov Chains,* Argentine: 4th Workshop on Information Security (WSegI), 2012, pp. 21-35.

[8] W. Bhaya, A. Rahma and D. AL-Nasrawi, "Text Steganography Based on Font Type in MS-Word Documents," *Journal of Computer Science,* vol. 9, no. 7, pp. 898-904, 2013.

[9] Y. Luo and Y. Huang, "Text Steganography with High Embedding Rate: Using Recurrent Neural Networks to Generate Chinese Classic Poetry," in *IH&MMSec '17: Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security*, Philadelphia, 2017.

[10] B. G. Banik and S. K. Bandyopadhyay, "Novel Text Steganography Using Natural Language Processing and Part-of-Speech Tagging," *IETE Journal of Research,* vol. 66, pp. 1-12, 2018.

[11] N. Alghamdi and L. Berriche, "Capacity Investigation of Markov Chain-Based Statistical Text Steganography: Arabic Language Case," in *Asia Pacific Information Technology Conference (APIT)*, Jeju Island, pp. 37-43, 2019.

[12] C. Zhi-li et al., "A statistical algorithm for linguistic steganography detection based on distribution of words," in *The 3rd International Conference on Availability, Security, and Reliability*, Barcelona, 2008.

[13] Z.L. Chen, et al., "Effective linguistic steganography detection," in *8th IEEE International Conference on Computer and Information Technology Workshops*, Sydney, 2008.

[14] Z. Chen et al., "Linguistic Steganography Detection Using Statistical Characteristics of Correlations between Words," *Proc of Information Hiding,* pp. 224-235, 2008.

[15] H. Yang and X. Cao, "Linguistic Steganalysis Based on Meta Features and Immune Mechanism," *Chinese Journal of Electronics,* vol. 19, no. 4, pp. 661- 666, 2010.

[16] L. Zhu, "A Linguistic Steganalysis Approach Base on Source Features of Text and Immune Mechanism," *Computer and Information Science,* vol. 10, p. 60, 2017.

[17] S. Samanta, S. Dutta and G. Sanyal, "A Real Time Text Steganalysis by using Statistical Method," in *2nd IEEE International Conference on Engineering and Technology (ICETECH)*, India, 2016.

[18] P. Meng et al., "Linguistic Steganography Detection Based on Perplexity," in *International Conference on MultiMedia and Information Technology*, Three Gorges, 2008.

[19] Z. L. Yang et al., "RNN-Stega: Linguistic Steganography Based on Recurrent Neural Networks," *IEEE Transactions on Information Forensics and Security,* vol. 14, no. 5, pp. 1280 - 1295, 2018.

[20] M. Abbas, "Arabic Corpora," google, [Online]. Available: https://sites.google.com/site/mouradabbas9/corpora.

[21] Y. Benajiba and P. Rosso, "Towards a measure for arabic corpora quality," 2007.

[22] C. Shannon, "A Mathematical Theory of Communication," *The Bell System Technical Journal,* vol. 27, pp. 379–423, 623–656,, 1948.

[23] M. Kubat, An Introduction to Machine Learning, Cham: Springer, 2015.

[24] W. Shu-feng and H. Liu-sheng, Research on Information Hiding, University of Science and Technology of China, 2003.

Nujud Alghamdi received the master's degree in computer science from Imam Mohammad Ibn Saud Islamic University, Saudi Arabia. She is currently a Lecturer at Saudi Electronic University, Saudi Arabia. Her research revolves

mostly around the areas of information security and Networking.

Lamia Berriche received the engineering degree from the French National School of Civil Aviation (ENAC) in 2001. In the same year, she received the master's degree in computer science, Networking and Telecommunication from the National Polytechnical Institute Toulouse, France. In 2006, she received her PhD in signal and image processing from Telecom Paris Tech in France. From 2007 to 2019, she was an assistant professor at the Computer Science Department, Imam Mohammad Ibn Saud Islamic University, Saudi Arabia. Currently, she is an assistant professor in the Computer Science Department at Prince Sultan University in Saudi Arabia.

Maha Alrabiah received the Ph.D. degree in computer science from King Saud University, Saudi Arabia, in 2014. She is currently an Assistant Professor with the Computer Science Department, Imam Muhammad Ibn Saud Islamic University, Saudi Arabia. Her research interests include computational linguistics and artificial intelligence.