A Significant Survey on Text Steganalysis Techniques

Sabyasachi Samanta
Haldia Institute of Technology, Haldia, WB, INDIA
E-mail id:sabyasachi.smnt@gmail.com

Solanki Pattyanayak
Haldia Institute of Management, Haldia, WB,INDIA
E-mail id: solankipattanayak16@gmail.com

ABSTRACT - Data concealing methods privacy, unapproved access to the substance and inaccessibility. Steganography is the craftsmanship that conceals any data through a suitable spread bearer like, picture, content, sound and video. Here, we have examined various sorts of content - text steganalysis strategies. There are unmistakable strategies for content steganalysis. Various applications have various prerequisites of the text steganalysis procedure utilized.

Keywords — Information Hiding, Text Steganalysis, Information Security, Steganography

I. INTRODUCTION

Data is a significant resource of humanity, whose security is a fundamental concern. Hazard increments which taking an attempt to continuous frameworks that incorporate the financial framework, railroads, flights and so on. Odds of assault increment when we transmit information by means of the web. A few sorts of assaults are conceivable, for example, listening in, the man in the centre assault, phishing assault, refusal of administration, and so on. So to verify our information, we are left with three principle arrangements which are by utilizing a private devoted channel, cryptography, and steganography. A private committed is tedious and the client is limited to a physical point. Cryptography shapes the message in some other structure. The team of cryptography and steganography can likewise be utilized which are known as transformative cryptography [12].

The primary motivation behind steganography, which signifies 'sending secluded from everything is to conceal information in a spread media with the goal that others won't have the option to see it (Figure 1). While cryptography is tied in with ensuring the substance of messages, steganography is tied in with disguising their very presence. The utilizations of data concealing frameworks primarily extend over a wide zone from military, knowledge offices, online decisions, web banking, medicinal imaging, etc. These assortments of utilizations make steganography a hotly debated issue for study. The spread medium is normally picked to remember the sort and the size of the mystery message and a wide range of bearer document configurations can be utilized. In the present circumstance, computerized pictures are the most well-known bearer/spread records that can be utilized to transmit mystery data [17].

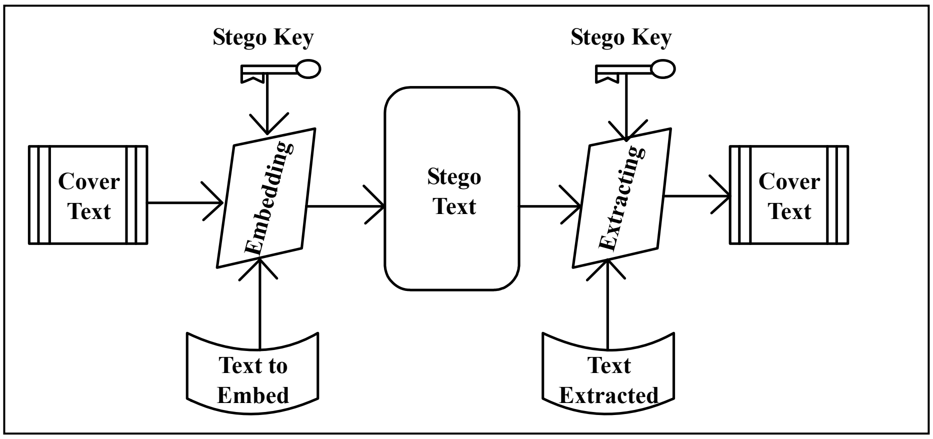

Figure 1: Text Steganography Scheme

Text Steganography is the way of hiding information inside the text. It's like the changing of existing text format, changing words within text, generating random character sequences etc. Various techniques are like format based method, random and statistical generation and linguistic method (as in Figure 2)[18].

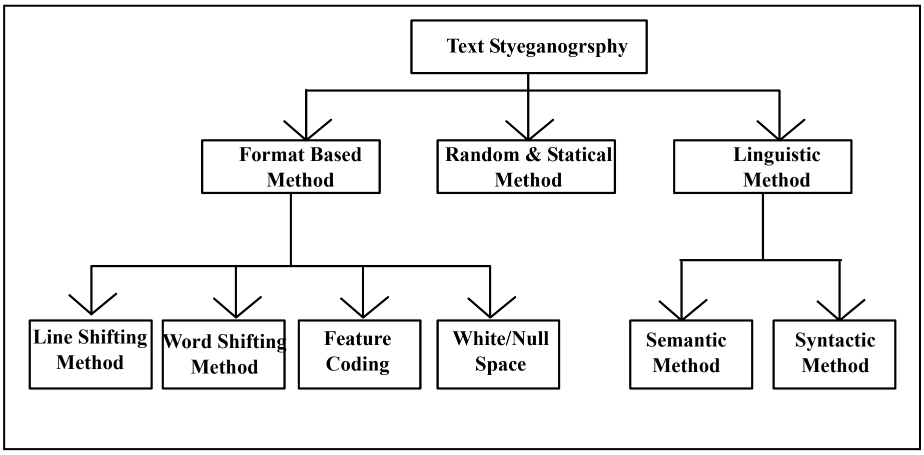

Figure 2: Nature of Text Steganography

The counter-technique of steganography is known as steganalysis. It begins by identifying the object that exist in the suspect stego file. An attacker may also embed irregular information over the existing concealed information. But the basic goal of text steganalysis is to investigate prohibited hidden information from the text stego files[19].

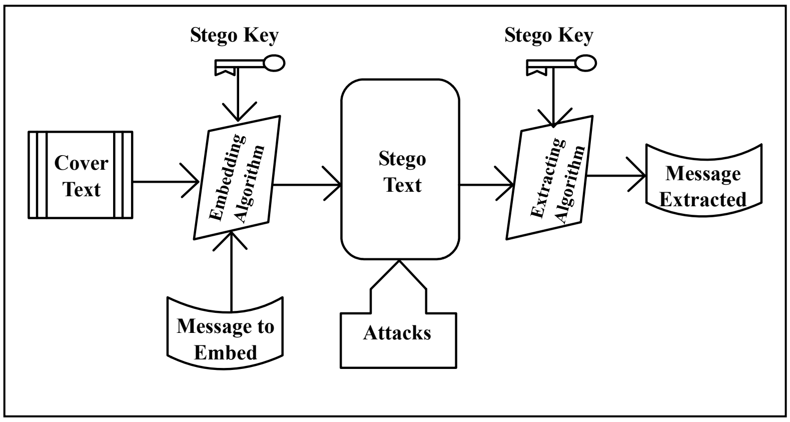

Figure 3: Text Steganalysis Scheme

Text steganalysis exploits the embedding information in carrier text by changing the statistical or exterior properties. Text steganalysis may be classified into three categories in more or less. They are similar to format-based, invisible character-based and linguistics, respectively. Linguistic steganalysis endeavour to notice secret messages in natural language texts. Due to multiplicity of syntax in natural language, it is complicated to observe the alterations in stego-text [20].

Section 2 represents works related to text steganalysis. Section 3 represents a comparison of different existing techniques. Section 4 draws a conclusion.

II. LITERATURE REVIEW

Zhongliang Yang et al. [1] proposed a productive book steganalysis strategy. They investigated the relationships between's words in these created steganographic writings. At that point, they mapped each word to a semantic space and utilized a concealed layer to extricate the relationships between's these words. In light of the separated connection highlights, they utilized the softmax classifier to order the info content. Test results show that the proposed model can accomplish a high location exactness, which shows a cutting edge execution.

Zhongliang Yang et al. [2] proposed the content steganalysis technique (TS-CNN) in view of semantic examination, which utilizes a convolutional neural system (CNN) to extricate elevated level semantic highlights of writings, and finds the inconspicuous conveyance contrasts in the semantic space when implanting the mystery data. To prepare and test the proposed model, they gathered and discharged an enormous book steganalysis (CT-Steg) dataset, which contains an all outnumber of 216,000 writings with different lengths and different implanting rates. The test result shows that the proposed model can accomplish about 100% accuracy and review, outflanks all the past techniques according to their course of action. In addition, the proposed model can figure the limit of the shrouded data inside.

Xin Zuo et al. [3] proposed a content semantic steganalysis dependent on word embedding. An intertwined include called word implanting and measurable list of capabilities (WESF) which comprises of WEF and factual component dependent on word recurrence is intended to improve recognition execution. Here all equivalent words are supplanted by spaces. For each clear, there is a synset. They have to pick the one that best fits the setting of this synset. So if the equivalent word shows up in the content and our decision isn't the equivalent, this equivalent word is crisscrossed with its specific situation. The quantity of criss-crosses can be utilized to recognize spread from stego.

Licai Zhu [4] proposed a phonetic steganalysis approach dependent on the source highlights of content and insusceptible instruments. This strategy has two attributions. The first is the fundamental measurable highlights of the content which is utilized for dazzle steganalysis. The subsequent one is the resistant method, picked to manufacture a two-level identification system to identify two classifications of stego message individually. One of which is Success-Stego-content and another is False-Stego-content. Fitting discoveries are produced and ideal highlights are agreed upon. Trials demonstrate the methodology has higher precision than current steganalysis calculations. Particularly when the portion size of the content is more noteworthy than 3kB, the correctness's of identifying for common content and stego content are both over 95%.

Roshidi Din et al. [5] proposed a content steganalysis framework utilizing the hereditary based technique. They measure the discovery execution dependent on the hereditary calculation technique and factual strategy so as to characterize the dissected content as stego content. Three viewpoints like time taken, a normal of cost work and the normal of mean and standard deviation have been utilized to quantify the exhibition strategies among measurable and proposed GA based.

Roshidi Din et al. [6] also proposed a content steganalysis utilizing the advancement calculation approach. They present another option of steganalysis strategy so as to distinguish shrouded messages in content steganalysis called Evolution Detection Steganalysis System (EDSS) in light of the development calculation approach under Java Genetic Algorithms Package (JGAP). The consequence of the EDSS can be partitioned into two gatherings dependent on wellness esteems which are acceptable wellness and terrible wellness. EDSS returns great wellness esteems when it can recognize every concealed message inside the sentences. Else, it returns terrible wellness esteems.

AB. RAUF et al. [7] introduced a novel text steganalysis technique based on color coded text visualization. The encoding scheme for text visualization is designed by analyzing text features with respect to colors to detect whitespace pattern. The mean of this methodology here is to differentiate between natural and stegano text.

Roshidi Din et al. [8] proposed a formalization of the hereditary calculation technique so as to identify shrouded messages on a dissected book. They utilized five metric parameters, for example, running time, wellness esteem, the normal mean likelihood, change likelihood, and standard deviation likelihood to quantify the location execution between measurable strategies and hereditary calculation techniques. Tests leading by the two techniques demonstrated that the hereditary calculation strategy performs obviously superior to a measurable technique, particularly in identifying short broke down writings.

Roshidi Din et al. [9] exhibited another elective strategy for content steganalysis dependent on an advancement calculation, executed utilizing the Java Evolution Algorithms Package (JEAP). The fundamental target was to recognize the presence of concealed messages dependent on the wellness estimations of a text description. The detection explains by two groups of fitness values which are good fitness and bad fitness value.

YongJian Bao et al. [10] proposed a LSTM-CNN model to tackle the text steganalysis problem. Firstly they mapped words into semantic space for better exploitation of the semantic feature in texts. Then they utilize a combination of Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) recurrent neural networks to capture both local and long-distance contextual information in stego-text. Also they applied attention mechanism to recognize and attend to important clues within suspicious sentences. Finally they used a softmax layer to categorize the input text as cover or stego. Experiment shows that their model may achieve the state-of-art result in text steganalysis framework.

Peng Meng et al. [11] designed a pre-processor to refine the given texts and to expand the frequency differences between normal texts and stego-text. They used a SVM classifier to classify given texts to normal texts and stegotexts.

Zhongliang Yang et al. [12] analyzed the correlations between words in generated steganographic texts. Then they mapped each word to a semantic space and used a hidden layer to extract the correlations between these words. Finally, based on the extracted correlation features, they used the softmax classifier to classify the input text.

Sabyasachi Samanta et al. [13] proposed a text steganalysis by using Bayesian Estimation and Correlation Coefficient based steganalysis methodologies. To measure the similarity between two carriers the order of moment has been taken up to 10th for Bayesian estimation. The values of PSNR, MSE, RMSE, Pearson Correlation Coefficient, Spearman's Rank Correlation Coefficient and Kendall Tau Rank Correlation Coefficient have considered calculating the statistical similarity between cover and stego text using Correlation Coefficient Steganalysis for Text (CCST).

Zhenshan Yu et al. [14] proposed a steganalysis of synonym-substitution based natural language watermarking. They evaluated the fitness of words for their context. Then the suitability sequence of words direct the final judgment made by a SVM classifier. IDF is also used to influence words' suitability in order to balance common words and uncommon words. Experimental results show the accuracy achieves up to 90.0%.

Zhili Chen et al. [15] proposed statistical characteristics of correlations between the general service words gathered in a dictionary. It's to classify the given text segments into stego-text segments and normal text segments. In the experiment three different linguistic steganography approaches are used. These are Markov-Chain-Based, NICETEXT and TEXTO. The accurateness to find out stego-text segments and normal text segments is found to be 97.19%.

Lingjun Li et al. [16] proposed a statistical analysis using word-shift text-steganography by neighbour difference. They have categorized the character of PDF document into two cases: "minor-change" class and "notable-change" class. When embedding rate is greater than 5%, more than 98% stego-text document can be distinguished out. The detection error is kept under 10% when embedding rate is more than 10%. They proposed the concept of "neighbor difference" which is susceptible to word-shift like steganographic scheme.

III. COMPARISON OF VARIOUS STEGANOGRAPHY TECHNIQUES

Authors	Year of Approach	Techniques used	Result/Analysis
Zhongliang Yang et. al. [1]	APRIL 2019	➢ Mapped each word to a semantic space and used a hidden layer to extract the correlations between these words. ➢ Based on the extracted correlation features, they used the softmax classifier to classify the input text.	➢ High detection accuracy
Zhongliang Yang et al. [12]	2019	➢ Analyzed the correlations between words. ➢ Mapped each word to a semantic space and used a hidden layer to extract the correlations between these words. ➢ Used the softmax classifier to classify the input text.	➢ High detection accuracy.
YongJian Bao et.al. [10]	December, 2019	➢ Used LSTM-CNN model. ➢ Mapped words into semantic space for better exploitation of the semantic feature. ➢ Applied attention mechanism to recognize and attend to important clues within suspicious sentences. ➢ Use a softmax layer to classify the input text as cover or stego.	➢ Achieved the state-of-art result.
Zhongliang Yang et. al. [2]	18 Oct 2018	➢ Based on semantic analysis. ➢ Uses Convolutional Neural Network (CNN).	➢ Proposed model can achieve nearly 100% precision and recall, outperforms all the previous methods. ➢ Can guess the capacity of the hidden information

			inside. ➢ Estimated accuracy rate is above 70%.
Xin Zuo et.al. [3]	2018	➢ Statistical feature based on word frequency is designed to improve detection performance. ➢ Getting an 11-D feature set whose detection performance is better than any other feature sets.	➢ Detect errors of WESF is smaller than both WEF and Xiang et al.'s features.
Licai Zhu [4]	2017	➢ Linguistic steganalysis approach. ➢ Use basis statistical features & immune technique.	➢ If the section size of text is more than 3kB, both SR and NR are more than 95%.
Roshidi Din et.al. [5]	MAY 2016	➢ Used genetic algorithm based method ➢ Used statistical method to classify the analyzed text as stego text. ➢ Time taken, average of cost function and average of mean and standard deviation have been used to measure the performance methods between statistical and proposed GA based.	➢ Shows the distribution of GA based method is more accurate than the statistical method.
Sabyasachi Samanta et al. [13]	2016	➢ Used Bayesian Estimation and Correlation Coefficient based Steganalysis techniques. ➢ Order of moment has been taken up to 10th for Bayesian estimation. ➢ The values of PSNR, MSE, RMSE, Pearson Correlation Coefficient, Spearman's Rank Correlation Coefficient and Kendall Tau Rank Correlation Coefficient have considered for CCST.	➢ High detection accuracy.
Roshidi Din et.al. [6]	2015	➢ Used Genetic Algorithms Package (JGAP) and Evolution Detection Steganalysis System (EDSS). ➢ Divided into two groups as good fitness and bad fitness.	➢ EDSS may be an alternative method to detect hidden messages within a text based natural language environment using EA approach.
Roshidi Din et.al. [8]	August 2015	➢ Formalization of genetic algorithm method in order to detect hidden message on an analyzed text. ➢ Five metric parameters were used to measure the detection performance between statistical methods and genetic algorithm methods.	➢ Genetic algorithm method performs much better than statistical method, especially in detecting short analyzed texts.

AB. RAUF et.al. [7]	2014	➢ Color coding technique is implemented to detect a format-based steganography on whitespace method. ➢ Employed two text feature: the distribution of space character and the average length of word.	➢ The detection performance accuracy successfully reaches 96.67% with remarkably high precision and recall.
Roshidi Din et.al. [9]	2013	➢ Used Java Evolution Algorithms Package (JEAP). ➢ Influenced by two groups of fitness values as good fitness value and bad fitness value.	➢ Proposed a sequential searching mode based on a natural language environment.
Peng Meng et al. [11]	2010	➢ Designed a pre-processor to refine between normal texts and stegotexts. ➢ Used a SVM classifier to classify normal text and stego-text.	➢ High detection accuracy.
Zhenshan Yu et al. [14]	2009	➢ Computed the concluding judgment made by a SVM. ➢ Utilized the influence words' suitability in order to balance common words and rare ones by IDF (inverse document frequency).	➢ Experimental results show 90.0% accuracy, 86.8% precision and 82.5% recall rate.
Zhili Chen et al. [15]	2008	➢ Three different linguistic steganography approaches: Markov-Chain-Based, NICETEXT and TEXTO. ➢ Strength of the correlation is measured by N-window mutual information (N-WMI).	➢ Total accuracy of discovering stego-text segments and normal text segments is found to be 97.19%.
Lingjun Li et al. [16]	2008	➢ Uses statistical attack. ➢ Categorized the PDF documents into two cases as "minor-change" and "notable-change" class.	➢ If embedding rate is greater than 5%, distinguish rate more than 98%.

IV. CONCLUSION

A number of research works is done in the area of text steganography. Also a number of attempts to detect the hidden information naming as text steganalysis is also developed. Here some of them have been included. Different techniques like softmax Classifier, CNN, GA, SVM, Markov-Chain-Based, NICETEXT and TEXTO etc. have been used by different authors to detect the presence of information in carrier. High detection accuracy also resulted by different authors. At rest the length of embedding message is not perfectly measurable by the exiting methodologies. So there is an extensive way to develop the new methodologies in the field of text steganalysis.

REFERENCES

[1] Zhongliang Yang , Yongfeng Huang and Yu-Jin Zhang, "A Fast and Efficient Text Steganalysis Method", IEEE SIGNAL PROCESSING LETTERS, VOL. 26, NO. 4, APRIL 2019, pp. 627-631

[2] Zhongliang Yang, Nan Wei, Junyi Sheng, Yongfeng Huang, Yu-Jin Zhang, "TS-CNN: Text Steganalysis from Semantic Space Based on Convolutional Neural Network", arXiv:1810.08136v1 [cs.CR], 18 Oct 2018

[3] Xin Zuo, Huanhuan Hu, Weiming Zhang(B), and Nenghai Yu, "Text Semantic Steganalysis Based on Word Embedding", ICCCS 2018, LNCS 11066, pp. 485–495, 2018

[4] Licai Zhu, "A Linguistic Steganalysis Approach Base on Source Features of Text and Immune Mechanism", Computer and Information Science; Vol. 10, No. 4; 2017, ISSN 1913-8989, pp. 60-66

[5] Roshidi Din, Faudziah Ahmad, H. S. Hussain, Shima Sabri, Nik Zulkarnaen Khidzir and Muzaaliff Musa, "A Performance of Text Steganalytic System using Genetic-Based Method", ARPN Journal of Engineering and Applied Sciences, VOL. 11, NO. 10, MAY 2016 ISSN 1819-6608, pp. 6216- 6221

[6] Roshidi Din, T. Zalizam T. Muda, Puriwat Lertkrai, Mohd Nizam Omar, Angela Amphawan And Fakhrul Anuar Aziz, "Text Steganalysis Using Evolution Algorithm Approach", Advances in Computer Science, ISBN: 978-1-61804-126-5, pp. 444-449,2015

[7] AB. RAUF, Rose Hafsa and JAMAL, Nurhafizah, “Feasibility of Text Visualization in Text Steganalysis”, 13th International conference on Intelligent Software Methodologies, Tools and Techniques, SOMET 2014, Langkawi, Malaysia Volume: 265:103-115

[8] Roshidi Din, Shafiz Affendi Mohd Yusof, Angela Amphawan, Hanizan Shaker Hussain, Hanafizah Yaacob, Nazuha Jamaludin and Azman Samsudin, “Performance Analysis on Text Steganalysis Method Using A Computational Intelligence Approach”, Proceeding of International Conference on Electrical Engineering, Computer Science and Informatics (EECSI 2015), Palembang, Indonesia, 19 - 20 August 2015, pp. 67-73

[9] Roshidi Din, Azman Samsudin, T. Zalizam T. Muda, P. Lertkrai, Angela Amphawan, and Mohd. Nizam Omar, “Fitness Value Based Evolution Algorithm Approach for Text Steganalysis Model”, International Journal of Mathematical Models And Methods In Applied Sciences, Issue 5, Volume 7, 2013, pp. 551-558

[10] YongJian Bao, Hao Yang, ZhongLiang Yang, Sheng Liu and YongFeng Huang, “Text Steganalysis with Attentional LSTM-CNN”, arXiv:1912.12871v1 [cs.MM] 30 Dec 2019

[11] Peng Meng, Liusheng Hang, Zhili Chen, Yuchong Hu, andWei Yang, “STBS: A Statistical Algorithm for Steganalysis of Translation-Based Steganography”, IH 2010, LNCS 6387, pp. 208–220, 2010

[12] Zhongliang Yang, Yongfeng Huang and Yu-Jin Zhang,” A Fast and Efficient Text Steganalysis Method”, IEEE Signal Processing Letters Volume: 26, Issue: 4, pp 627-631 DOI: 10.1109/LSP.2019.2902095, 2019

[13] Sabyasachi Samanta, Saurabh Dutta and Goutam Sanyal, “A Real Time Text Steganalysis by using Statistical Method”, 2nd IEEE International Conference on Engineering and Technology (ICETECH), 17th & 18th March 2016, Coimbatore, TN, India

[14] Zhenshan Yu, Liusheng Huang, Zhili Chen, Lingjun Li, Xinxin Zhao, and Youwen Zhu, “Steganalysis of Synonym-Substitution Based Natural Language Watermarking” International Journal of Multimedia and Ubiquitous Engineering Vol. 4, No. 2, April, 2009, pp.21-34

[15] Zhili Chen, Liusheng Huang, Zhenshan Yu, Wei Yang Lingjun Li, Xueling Zheng and Xinxin Zhao, “Linguistic Steganography Detection Using Statistical Characteristics of Correlations between Words”, 10th International Workshop, IH 2008, Santa Barbara, CA, USA, May 19-21, 2008, Revised Selected PapersOctober 2008 Pages 224–235

[16] Lingjun Li, Liusheng Huang, Xinxin Zhao, Wei Yang and Zhili Chen, “A Statistical Attack on a Kind of Word-Shift Text-Steganography”, International Conference on Intelligent Information Hiding and Multimedia Signal Processing, 978-0-7695-3278-3/08, 2008 IEEE, DOI 10.1109/IIH-MSP.2008.42

[17] Bin Li, Junhui He, Jiwu Huang and Yun Qing Shi, “A Survey on Image Steganography and Steganalysis”, Journal of Information Hiding and Multimedia Signal Processing, Volume 2, Number 2, April 2011, pp.142-172

[18] Souvik Bhattacharyya, Indradip Banerjee and Gautam Sanyal, “ A Survey of Steganography and Steganalysis Technique in Image, Text, Audio and Video as Cover Carrier”, Journal of Global Research in Computer Science, IISN 2229-371X, Volume 2, No. 4, April 2011 , pp. 1-16

[19] Milad Taleby Ahvanooey, Qianmu Li, Jun Hou, Ahmed Raza Rajput and Chen Yini, “Modern Text Hiding, Text Steganalysis, and Applications: A Comparative Analysis”, *Entropy* 2019, *21*, 355; doi:10.3390/e21040355, pp. 1-29

[20] Roshidi Din and Azman Samsudin, “Digital Steganalysis: Computational Intelligence Approach”, INTERNATIONAL JOURNAL OF COMPUTERS, Issue 1, Volume 3, 2009, pp. 161-170