TStego-THU: Large-Scale Text Steganalysis Dataset

Zhongliang Yang, Jin He, Siyu Zhang, Jinshuai Yang, and Yongfeng Huang(✉)

Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
yfhuang@tsinghua.edu.cn

Abstract. In recent years, with the development of natural language processing (NLP) technology, linguistic steganography has developed rapidly. However, to the best of our knowledge, currently there is no public dataset for text steganalysis, which makes it difficult for linguistic steganalysis methods to get a fair comparison. Therefore, in this paper, we construct and release a large-scale linguistic steganalysis dataset called TStego-THU, which we hope to provide a fair enough platform for comparison of linguistic steganalysis algorithms and further promote the development of linguistic steganalysis. TStego-THU includes two kinds of text steganography modes, namely, text modification-based and text generation-based modes, each of which provides two latest or classical text steganography algorithms. All texts in TStego-THU come from three common transmitted text medias in cyberspace: News, Twitter and commentary text. Finally, TStego-THU contains 240,000 sentences (120,000 cover-stego text pairs), each steganographic sentence is generated by randomly choosing one of these four steganographic algorithms and embedding random bitstream into randomly extracted normal texts. At the same time, we also evaluate some latest text steganalysis algorithms as benchmarks on TStego-THU, the detail results can be found in the experiment part. We hope that TStego-THU will further promote the development of universal text steganalysis technology. The description of TStego-THU and instructions will be released here: https://github.com/YangzlTHU/Linguistic-Steganography-and-Steganalysis.

Keywords: TStego-THU · Text steganalysis · Dataset

1 Introduction

According to Shannon's summary in his monograph on information theory [1], security systems in cyberspace can be divided into three categories: encryption system, privacy system and concealment system. Among them, encryption system and privacy system mainly encrypt important information or control the access rights to ensure the security of information content. Concealment system mainly protects the information from the perspective of communication behavior security, which hides the existence of important information to ensure the security of it. However, the convert communication system's powerful information concealing ability may also be used by criminals to transmit some dangerous information, thus endangering cyberspace security and public safety [2].

X. Sun et al. (Eds.): ICAIS 2021, CCIS 1424, pp. 335–344, 2021.
https://doi.org/10.1007/978-3-030-78621-2_27

Therefore, making in-depth researches in steganalysis and detecting the potential harmful information in the network information flow is of great significance for maintaining cyberspace security.

In theory, any information carrier can serve as a medium for covert communication. In cyberspace, the most common information carriers for covert communication are image [3], audio [4], and text [5–7]. For different periods of historical development (whether in the paper-media age or the electronic-media age), text is one of the most important information carriers, playing a very important role in human lives. Therefore, in recent years, there has appeared more and more information hiding methods based on texts [5–7]. These text steganographic methods, embedding some secret information into common texts, pretending as normal texts and disseminating widely in public networks, pose a great threat and challenge to the security of cyberspace [2]. Currently, the main text steganography algorithms can be divided into two categories: steganography based on text modification [8–10] and steganography based on text generation [5–7]. Steganography based on carrier modification is mainly used to embed covert information by modifying the original texts, such as synonym substitution and syntactic structure substitution. This kind of methods can only make very little affect on the semantic expression of the texts to achieve a very high concealment, which makes it very difficult to be detected. However, usually, such methods are difficult to achieve a higher embedding rate.

Another kind of methods is information hiding based on text automatic generation [5–7]. The most important characteristic of this kind of methods is that they do not need to be given the specify texts beforehand, but can automatically generate steganographic texts according to the convert information that need to be transmitted. They usually implement secret information hiding by encoding words during the automatic text generation process. Especially with the rapid development of natural language processing technologies in recent years, text automatic generation technology based on neural networks has made rapid development, and has made great progress in the fields of human-computer dialogue [7], automatic translation [11] and image captioning [12]. With the help of these technologies, researchers can generate high quality natural steganographic texts according to the secret information and ensure reliable concealment [5, 6]. At the same time, since almost every word in the generated texts can be embedded with secret information, this kind of methods can usually achieve a high information hiding capacity [5, 7]. However, it is worth noting that although information hiding technology plays an important role in protecting the privacy of users in cyberspace. But this technology may also be used by criminals and even terrorists to build covert communication systems and transmit dangerous information, endangering public safety and cyberspace security [2]. Therefore, we also need to conduct deeply research on text steganalysis to detect whether the texts transmitted in the cyberspace contains dangerous information.

Current steganalysis methods basically adopt the same framework, that is, by constructing specific statistical features or analytical methods, to find the differences in statistical distribution in a certain dimension between the covertext and the stegotext to conduct steganalysis [13–19]. Traditional text steganalysis methods usually firstly construct a series of text statistical features manually, then analyze the changes of these features before and after steganography, and finally design the corresponding discriminators [17–19]. However, most of the statistical features utilized by these methods are simple, such as word frequency distribution [17], transition probability distribution between words [18], and so on [19]. However, with the development of artificial neural network technology and natural language processing technology, more and more neural network-based text steganalysis models have emerged in recent years [13–16]. Such models usually need a large number of labeled texts to train the corresponding neural network model, and then judge whether the input text is normal text or steganographic text according to the extracted features.

However, as far as we know, currently there is no public dataset for text steganalysis. In order to provide an open test platform for researchers in the field of text steganography and steganalysis, and to promote the development of text steganography and steganalysis technology, in this paper, we construct and release a large-scale universal text steganalysis dataset, called TStego-THU. This dataset contains 240,000 sentences that may contain hidden information. The original sentences in TStego-THU comes from some widely used network platforms. And in order to make the dataset more universal, we use different steganography methods, including two modification-based and two generation-based steganography algorithms, to embed random bit streams into sentences.

In the remainder of this paper, Sect. 2 introduces the details of the construction process of TStego-THU, including data collection and preprocessing, information embedding algorithms. The following part, Sect. 3, describes the steganalysis benchmarks we used and their performance on TStego-THU dataset. Finally, conclusions are drawn in Sect. 4.

2 The Construction of TStego-THU

In this section, we will introduce the details of the construction procedure of TStego-THU, including source text collection, text preprocessing and information hiding methods. Finally, we give the overall distribution characteristics of TStego-THU.

2.1 Source Text Collection

In order to ensure that the samples in the constructed dataset are close enough to the real scene, in the process of constructing TStego-THU, we mainly used three common transmitted text medias in cyberspace, namely Twitter, movie reviews and News. For Twitter, we chose the sentiment140 dataset published by Alec Go et al. [20]. It contains 1,600,000 tweets extracted using the Twitter API. For the movie review dataset, we chose the widely used IMDB dataset published by Maas et al. [21], which contains 25,000 movie reviews from IMDB. The texts of the two datasets above are of the social media type. In addition, we also chose a news dataset [22] containing relatively more standard texts. It contains 143,000 articles from 15 American publications, including the New York Times, Breitbart, CNN, etc. The topics of the dataset are mainly politically related and the published time is mainly between 2016 and July 2017.

2.2 Text Preprocessing

Before steganography, we need to pre-process the collected original texts. Firstly, we unified all the texts in TStego-THU into lowercase. Secondly, we divided all the collected text into single sentences, that is, each sample in TStego-THU contains only one sentence. Then, we removed special symbols from the texts, such as emoticons, web links, special formatted text, etc. Finally, we built a dictionary of all the text in TStego-THU and filtered out words whose word frequency is below a certain threshold (usually between 5–10), mainly to avoid some misspellings and low-frequency words. After pre-processing, the text details of each source are shown in Table 1.

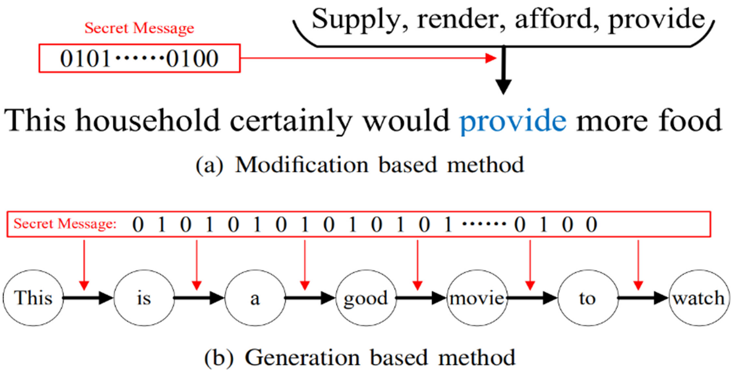
Table 1. The details of each text source.

Dataset	Twitter	IMDB	News
Average length	9.68	19.94	22.24
Sentence number	2,639,290	1,283,813	1,962,040
Words number	25,551,044	25,601,794	43,626,829
Unique number	46,341	48,342	42,745

2.3 Information Hiding

Since currently text steganography algorithms mainly include two types: text modification based and text generation based. Steganographic samples in TStego-THU mainly used these two kinds of algorithms for random bits stream embedding. The general framework of these two algorithms is shown in Fig. 1. For each type of steganographic methods, we chose two steganographic algorithms for embedding. For modification based methods, one of them we choose T-Lex, which is a very classic and widely used text steganographic model proposed by K. Winstein et al. [23], and the other one we chose the latest algorithm proposed by H. Hu et al. [24]. Both of these two algorithms are based on synonym substitution (SS), which embeds secret information by changing cover words for their synonyms. In the process of building TStego-THU, we first used WordNet [25] to build a synonym list of different data sources. Then we randomly selected 10,000 sentences from each data source, for each of which, we randomly used one of the two text steganography algorithms to embed the random bit stream into the text to form 10,000 cover-stego sentence pairs. It is worth noting that in some cases, even if no words are replaced in the whole sentence, it may also carry hidden information (e.g. “0”).

For generation based steganographic methods, we mainly used the two steganographic models proposed by Yang et al. [5, 26]. These two models first learn the statistical language model of a large number of normal texts and then encode the conditional probabilities of each word in the generation process to achieve the hiding of secret information. The difference lies in the generation model used by both: the model proposed in [26] uses the Markov chain model and the model in [5] uses the recurrent neural network (RNN) model. In the process of constructing TStego-THU, we first trained these two models with all the texts from different sources to learn the overall statistical language model. Then, in the process of text generation, the conditional probability space is dynamically encoded by the Huffman tree. And finally the steganographic texts are generated according to the random bits stream. Both methods can adjust the overall embedding rate by adjusting the number of bits embedded in each word. To ensure high concealment, we embeded only one bit per word (bpw = 1) when constructing TStego-THU. It is also worth noting that, considering the text generation models can also generate texts without hidden information, and in fact many texts on the network are automatically generated by computers, such as some man-machine dialogues, news summaries and so on. Therefore, in order to be more consistent with the real scene, when constructing the cover set of TStego-THU, we also incorporated some texts that are automatically generated by the models without covert information.


Secret Message
0101······0100
Supply, render, afford, provide
This household certainly would provide more food
Secret Message: 0 1 0 1 0 1 0 1 0 1 0 1 0 1 ······ 0 1 0 0
This
is
a
good
movie
to
watch

(a) Modification based method

(b) Generation based method

Fig. 1. The general frameworks of two types text steganographic algorithms: text modification based and text generation based.

2.4 Embedding Rate

After completing the information embedding, we calculated the actual embedding rate of different embedding algorithms. The embedding rate is calculated as the ratio of the number of bits actually embedded in all texts and the number of bits occupied by the texts stored in the computer. Through the actual test, the embedding rate of each steganography algorithms on different source of texts is shown in Table 2. As can be seen from Table 2, the modification-based steganography algorithm usually has an embedding rate of less than 1%. The generated steganographic model, even if each word is only embedded in 1 bit, greatly exceeds the modification-based steganography algorithms. Finally, the average embedding rate of the TStego-THU is 1.543%.

Table 2. The embedding rate of each steganography algorithms on different sources of texts.

Stegnography	Twitter	IMDB	News	Average
T-lex [23]	0.773%	0.689%	0.539%	0.667%
STC [24]	0.584%	0.476%	0.364%	0.475%
Markov [26]	2.424%	2.182%	2.087%	2.231%
RNN-Stega [5]	3.060%	2.814%	2.763%	2.879%
Average	1.710%	1.540%	1.438%	1.563%

2.5 Overall Details of TStego-THU

After all these operations, we obtained 10,000 cover-stego text pairs by using different steganography algorithms for different source of texts. We mixed them together to form a large-scale steganalysis texts dataset containing 240,000 sentences (120,000 cover-stego text pairs) in TStego-THU. The overall distribution characteristics of TStego-THU are shown in Table 3.

Table 3. The overall distribution characteristics of TStego-THU.

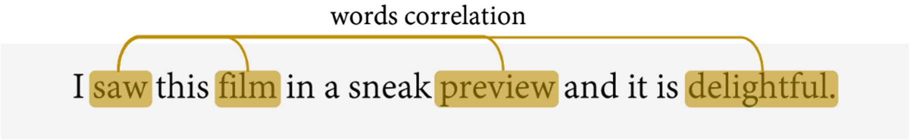
Stegnography	Twitter	IMDB	News	Total
T-lex (Cover:Stego)	10,000: 10,000	10,000: 10,000	10,000: 10,000	10,000: 10,000
STC (Cover:Stego)	10,000: 10,000	10,000: 10,000	10,000: 10,000	10,000: 10,000
Markov (Cover:Stego)	10,000: 10,000	10,000: 10,000	10,000: 10,000	10,000: 10,000
RNN-Stega (Cover:Stego)	10,000: 10,000	10,000: 10,000	10,000: 10,000	10,000: 10,000
Total (Cover:Stego)	40,000: 40,000	40,000: 40,000	40,000: 40,000	40,000: 40,000

3 Experiment

In order to evaluate the difficulty of TStego-THU and also to provide benchmark results for subsequent researchers, in this section, we will evaluate some latest text steganalysis algorithms on TStego-THU.

3.1 Benchmark Methods and Evaluation Metrics

We choose three latest text steganalysis models as the benchmark models of TStego-THU, which are proposed in [13, 14, 16]. These three text steganalysis models are all based on neural networks. Among them, the authors in [13] described the semantic correlations between words in texts, which has been explained in Fig. 3. They analyzed that once the covert information was embedded in the text, it may destroy the semantic correlations between words. So they mainly use word vector and fasttext model to extract the semantic correlations of words and conduct text steganalysis. The authors in [14] further refined the correlations of words into successive word correlation, cross word correlation and cross sentence correlation, and used multi-channel convolution kernels to extract these words correlations for text steganalysis.



The authors in [16] analysised that the conditional probability distribution of words in the sentence may be effected once the covert information was embedded. Therefore, they use bidirectional Long Short-Term Memory (Bi-LSTM) model [27] model to extract the conditional probability distribution features in texts, and then conduct text steganalysis according to the distribution differences. The experimental results show that they can even estimate the amount of covert information that may be contained in steganographic sentence based on the distribution of text features [16].

3.2 Detection Results of Benchmark Methods

We randomly selected 80% sentences from TStego-THU as the training samples to train each steganalysis model, and the remaining 20% used as the test samples to evaluate model's steganalysis ability. The final test results are shown in Table 4.

Table 4. The overall performance of each benchmark methods.

Methods	Accuracy	Precision	Recall	F1
Fasttext [13]	0.593	0.593	0.593	0.593
CNN [14]	0.584	0.584	0.584	0.584
LSTM [16]	0.591	0.591	0.591	0.591
Bi-LSTM [16]	0.578	0.578	0.578	0.578

From the results in Table 4, the first thing worth noting is that the detection performance of these three models are very close, and their detection accuracy are all less than 60%. This shows that TStego-THU dataset is very difficult and very challenging for steganalysis. We think there are several reasons. The first and most important reason is that we used a very low embedding rate when building the dataset, as shown in Table 2. Secondly, most of the previous text steganalysis algorithms were aimed at one single steganographic algorithm. In order to encourage researchers to explore more general text steganalysis algorithms, TStego-THU contains four different text steganography algorithms, which is more challenging. Thirdly, for the same purpose, in order to encourage the study of text steganalysis algorithms in more general scenarios, texts in TStego-THU comes from three different sources and is ultimately mixed together. Since the statistical distribution patterns of each different type of text have certain differences, mixing them together further increases the versatility of the dataset, but at the same time further increases the difficulty.

Based on the above analysis, we can see that TStego-THU has the characteristics of low embedding rate, multiple steganography algorithms, and multiple text types. We encourage follow-up researchers to further develop universal text steganalysis models for multi scenarios that can effectively detect low embedding rates. This will further promote the development of text steganalysis in a more general and practical direction.

Further, we want to know if different types of texts will affect the text steganographic detection performance. Therefore, according to the detection results in Table 4, we further counted the detecting accuracy of texts from different sources in the test samples. The results are shown in Table 5. From Table 5, it seems that Twitter text is relatively easier to detect, while the accuracy of other two types of text detection is relatively close. This also reflects that the statistical characteristics of various source texts are indeed different.

Table 5. The detection results of texts from different sources in the test samples.

Steganalysis	Dataset	Accuracy	Precision	Recall	F1
Fasttext [13]	IMDB	0.560	0.560	0.560	0.560
	News	0.586	0.586	0.586	0.586
	Twitter	0.633	0.633	0.633	0.633
CNN [14]	IMDB	0.575	0.575	0.575	0.575
	News	0.595	0.595	0.595	0.595
	Twitter	0.581	0.581	0.581	0.581
LSTM [16]	IMDB	0.565	0.565	0.565	0.565
	News	0.597	0.597	0.597	0.597
	Twitter	0.610	0.610	0.610	0.610
Bi-LSTM [16]	IMDB	0.551	0.551	0.551	0.551
	News	0.567	0.567	0.567	0.567
	Twitter	0.616	0.616	0.616	0.616

4 Conclusion

In this paper, we construct and release a large-scale univeral text steganalysis dataset called TStego-THU. In TStego-THU, we used text from several different sources, used multiple steganographic algorithms, and kept the embedding rate low. We hope TStego-THU can provide a fair enough platform for comparison of text steganalysis algorithms and further promote the development of universal text steganalysis.

Acknowledgement. Authors are deeply indebted to anonymous reviewers for their constructive suggestions and helpful comments.

Funding Statement. This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFB0804103 and the National Natural Science Foundation of China (No. U61972168, No. U1705261 and No. U1636113).

Conflicts of Interest. The authors declare that they have no conflicts of interest to report regarding the present study.

References

1. Shannon, C.E.: Communication theory of secrecy systems. Bell Syst. Tech. J. **28**(4), 656–715 (1949)
2. Theohary, C.A.: Terrorist Use of the Internet: Information Operations in Cyberspace. DIANE Publishing (2011)
3. Yang, Z., Wang, K., Ma, S., Huang, Y., Kang, X., Zhao, X.: Istego100k: Large-scale image steganalysis dataset. In: Wang, H., Zhao, X., Shi, Y., Kim, H.J., Piva, Alessandro (eds.) Digital Forensics and Watermarking. LNCS, vol. 12022, pp. 352–364. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-43575-2_29
4. Yang, Z., Peng, X., Huang, Y.: A sudoku matrix-based method of pitch period steganography in low-rate speech coding. In: Lin, X., Ghorbani, A., Ren, K., Zhu, S., Zhang, A. (eds.) Security and Privacy in Communication Networks, vol. 238, pp. 752–762. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-78813-5_40
5. Yang, Z.L., Guo, X.Q., Chen, Z.M., Huang, Y.F., Zhang, Y.J.: RNN-stega: linguistic steganography based on recurrent neural networks. IEEE Trans. Inf. Forensics Secur. **14**(5), 1280–1295 (2018)
6. Yang, Z.L., Zhang, S.Y., Hu, Y.T., Hu, Z.W., Huang, Y.F.: VAE-Stega: linguistic steganography based on variational auto-encoder. IEEE Trans. Inf. Forensics Secur. **16**, 880–895 (2020)
7. Yang, Z., Zhang, P., Jiang, M., Huang, Y., Zhang, Y.J.: Rits: real-time interactive text steganography based on automatic dialogue model. In: Sun, X., Pan, Z., Bertino, E. (eds.) ICCCS 2018. LNCS, vol. 11065, pp. 253–264. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00012-7_24
8. Xiang, L., Luo, G., Xie, Y., Chen, W.: Research on the coding strategies for synonym substitution-based steganography. J. Comput. Inf. Syst. **10**(22), 9877–9888 (2014)
9. Xiang, L., Wu, W., Li, X., Yang, C.: A linguistic steganography based on word indexing compression and candidate selection. Multimed. Tools Appl. **77**(21), 28969–28989 (2018). https://doi.org/10.1007/s11042-018-6072-8
10. Xiang, L., Wang, X., Yang, C., Liu, P.: A novel linguistic steganography based on synonym run-length encoding. IEICE Trans. Inf. Syst. **100**(2), 313–322 (2017)
11. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)
12. Yang, Z., Zhang, Y.J., Rehman, S., Huang, Y.: Image captioning with object detection and localization. In: Zhao, Y., Kong, X., Taubman, D. (eds.) ICIG 2017. LNCS, vol. 10667, pp. 109–118. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-71589-6_10
13. Yang, Z., Huang, Y., Zhang, Y.J.: A fast and efficient text steganalysis method. IEEE Signal Process. Lett. **26**(4), 627–631 (2019)

14. Yang, Z., Huang, Y., Zhang, Y.-J.: TS-CSW: text steganalysis and hidden capacity estimation based on convolutional sliding windows. Multimed. Tools Appl. **79**(25–26), 18293–18316 (2020). https://doi.org/10.1007/s11042-020-08716-w
15. Wen, J., Zhou, X., Zhong, P., Xue, Y.: Convolutional neural network based text steganalysis. IEEE Signal Process. Lett. **26**(3), 460–464 (2019)
16. Yang, Z., Wang, K., Li, J., Huang, Y., Zhang, Y.J.: TS-RNN: text steganalysis based on recurrent neural networks. IEEE Signal Process. Lett. **26**(12), 1743–1747 (2019)
17. Yang, H., Cao, X.: Linguistic steganalysis based on meta features and immune mechanism. Chin. J. Electron. **19**(4), 661–666 (2010)
18. Meng, P., Hang, L., Yang, W., Chen, Z., Zheng, H.: Linguistic steganography detection algorithm using statistical language model. In: 2009 International Conference on Information Technology and Computer Science, vol. 2, pp. 540–543. IEEE, July 2009
19. Samanta, S., Dutta, S., Sanyal, G.: A real time text steganalysis by using statistical method. In: 2016 IEEE International Conference on Engineering and Technology (ICETECH), pp. 264–268. IEEE, March 2016
20. Go, A., Bhayani, R., Huang, L.: Twitter sentiment classification using distant supervision. CS224N project report, Stanford **1**(12), 2009 (2009)
21. Maas, A., Daly, R.E., Pham, P.T., Huang, D., Ng, A.Y., Potts, C.: Learning word vectors for sentiment analysis. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pp. 142–150, June 2011
22. Thompson, A.: Kaggle. https://www.kaggle.com/snapcrack/all-the-news/data
23. Winstein, K.: Lexical steganography through adaptive modulation of the word choice hash (1998)
24. Huanhuan, H., Xin, Z., Weiming, Z., Nenghai, Y.: Adaptive text steganography by exploring statistical and linguistical distortion. In: 2017 IEEE Second International Conference on Data Science in Cyberspace (DSC), pp. 145–150. IEEE, June 2017
25. Miller, G.A.: WordNet: a lexical database for English. Commun. ACM **38**(11), 39–41 (1995)
26. Yang, Z., Jin, S., Huang, Y., Zhang, Y., Li, H.: Automatically generate steganographic text based on Markov model and Huffman coding. arXiv preprint arXiv:1811.04720 (2018)
27. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)