

Provably Secure Generative Linguistic Steganography

Siyu Zhang, Zhongliang Yang, Jinshuai Yang, Yongfeng Huang

Department of Electronic Engineering, Tsinghua University, Beijing 100084, China

{zhangsiy19, yjs20}@mails.tsinghua.edu.cn

yangz115@tsinghua.org.cn

yfhuang@mail.tsinghua.edu.cn

Abstract

Generative linguistic steganography mainly utilized language models and applied steganographic sampling (stegosampling) to generate high-security steganographic text (stegotext). However, previous methods generally lead to statistical differences between the conditional probability distributions of stegotext and natural text, which brings about security risks. In this paper, to further ensure security, we present a novel provably secure generative linguistic steganographic method **ADG**, which recursively embeds secret information by **Adaptive Dynamic Grouping** of tokens according to their probability given by an off-the-shelf language model. We not only prove the security of ADG mathematically, but also conduct extensive experiments on three public corpora to further verify its imperceptibility. The experimental results reveal that the proposed method is able to generate stegotext with nearly perfect security.

1 Introduction

Steganography is the technology of hiding secret information within an innocent natural carrier (such as image (Hussain et al., 2018), audio (Mishra et al., 2018), video (Liu et al., 2019), text (Krishnan et al., 2017), etc) in order to avoid eavesdropping. Steganography differs from cryptography in that cryptography only conceals the content of secret information, whereas steganography even conceals its very existence, which makes it more secure and reliable in some scenarios (Anderson and Petitcolas, 1998).

Natural language is suitable as a carrier of steganography by virtue of its high robustness in transmission (Ziegler et al., 2019). Unlike digital images or digital audio which is sensitive to distortions like compression, cropping, blurring or pixel-wise dropout, text can usually be transmitted

losslessly through different kinds of public channels. Nevertheless, text generally has low entropy and lacks sufficient redundancy for information hiding (Sharma et al., 2016), which often results in low embedding capacity of linguistic steganography. For example, in traditional modification-based methods (such as synonym substitution (Xiang et al., 2014, 2018) and spelling transformation (Shirali-Shahreza, 2008)), where secret information is encoded by slightly modifying an existing coverttext, the options for modification can be very limited to keep the text fluent enough so as not to arouse suspicions.

In recent years, powered by the advanced technology of deep learning and natural language processing, language models based on neural networks have made significant progress in generating fluent text (Radford et al., 2019; Brown et al., 2020), which bring new vitality to linguistic steganography and facilitate the investigation of generation-based methods (Fang et al., 2017; Yang et al., 2018a; Dai and Cai, 2019; Ziegler et al., 2019; Yang et al., 2020a; Zhou et al., 2021). The generative linguistic steganography directly transform secret information into innocuous-looking steganographic text (stegotext) without any coverttext. Using an off-the-shelf language model, secret information can be encoded in the selection of token at each time step autoregressively during the generation procedure, which greatly alleviates the drawback of low embedding capacity. However, previous methods inevitably introduce distortions during generation. The imperceptibility of generative linguistic steganography still needs further optimization.

In this paper, we aim to further improve the imperceptibility of generative linguistic steganography. The contributions of this work are the following:

1. We present **ADG** (**Adaptive Dynamic**

Grouping), a novel generative linguistic steganographic method based on off-the-shelf language models, which groups the tokens adaptively in accordance with their probability at each time step to embed secret information dynamically in the generated stegotext.

2. We discuss the security of ADG and give mathematical proof, which reveals that the proposed method is provably secure.
3. Through quantitative analysis, we derive satisfactory experimental results in terms of both imperceptibility and embedding capacity, which further verifies the effectiveness of ADG.

Our code is available at <https://github.com/Mhzzzzz/ADG-steganography>.

2 Formalism

2.1 Notation

We use lowercase letters in bold type (e.g. \mathbf{a}) to denote vectors, normal lowercase letters (e.g. a) to denote scalars and uppercase letters (e.g. A) to denote sets. We use the symbol $|A|$ to denote the size of a set. Calligraphic letters denote neural models (e.g. \mathcal{A}). Both English letters and Greek letters are adopted. We use $p(\cdot)$ and $q(\cdot)$ to denote distributions and $f(\cdot)$ to denote functions, which are usually shortened to p , q and f . Subscripts and superscripts are used to tell the different variables/distributions/functions apart.

2.2 Generative Linguistic Steganography

Language modeling is a task to estimate the joint distribution of serialized natural language $p_{LM}(\mathbf{w})$, where \mathbf{w} is a sequence of n tokens $[w_1, w_2, \dots, w_n]$ and each token belongs to the vocabulary Σ . For an autoregressive language model \mathcal{L} , the output is usually factorized as a product of conditional distribution of the current token

$$\begin{aligned} p_{LM}(\mathbf{w}) &= p_{LM}(w_1, w_2, \dots, w_n) \\ &= p_{LM}(w_1) \cdot \prod_{t=2}^n p_{LM}(w_t | w_1, \dots, w_{t-1}). \end{aligned} \quad (1)$$

According to Simmons (1984), it is usually supposed that Alice (sender) wants to send a secret message $\mathbf{m} \sim \text{Uniform}(\{0, 1\}^l)$ to Bob (receiver) through a public channel monitored by Eve (adversary). In generative linguistic steganography, they share an *embedding* algorithm f_{emb} which takes a

language model \mathcal{L} and the secret message \mathbf{m} as input and then outputs *stegotext* \mathbf{y} to transmit. They also share a corresponding *extraction* algorithm f_{ext} , which is the inverse mapping of f_{emb} that is able to recover the secret message \mathbf{m} according to the language model \mathcal{L} and the received stegotext \mathbf{y} .

2.3 Imperceptibility

In order to avoid raising Eve’s suspicions, stegotext \mathbf{y} is required to be fluent enough and statistically indistinguishable from natural innocuous text \mathbf{x} , which we call *coverttext*. Cachin (1998) proposed the *information-theoretic security* of steganography to measure the statistical imperceptibility quantitatively, which is defined as the Kullback-Leibler divergence (KL divergence) between the distributions of coverttext \mathbf{x} and stegotext \mathbf{y} . The distortion of generative linguistic steganography is two-fold: one is introduced by the bias of the language models, which is the gap between the true distribution of natural text $p_{true}(\mathbf{x})$ and the modeled distribution $p_{LM}(\mathbf{x})$; the other is introduced by f_{emb} . Instead of directly sampling from the modeled distribution, the embedding algorithm f_{emb} actually provides a special way to sample from $p_{LM}(\mathbf{y})$, which we call *steganographic sampling* (stegosampling). It is equivalent to sampling from a modified distribution $q(\mathbf{y})$ produced by an implicit language model \mathcal{L}' . In a word, the latter distortion is the gap between $p_{LM}(\mathbf{y})$ and $q(\mathbf{y})$, which can also be regarded as the gap between the conditional distributions $p_{LM}(y_t | y_{<t})$ and $q(y_t | y_{<t})$. We simply use p_{LM} and q to refer to the conditional distributions in the rest of this paper.

3 Related Work

In the early stage, some researchers investigated rule-based approaches or using Markov Chains to achieve generative linguistic steganography (Wayner, 1992; Chapman and Davida, 1997; Chapman et al., 2001; Chapman and Davida, 2002; Dai et al., 2010; Moraldo, 2014; Luo et al., 2016; Yang et al., 2018b). However, these methods followed a simplistic pattern and are hard to guarantee the grammatical correctness and the semantic fluency of the generated stegotext.

With the development of deep learning, language models based on neural networks show great performance on automatic text generation. The pattern of generating stegotext with neural language models has been widely accepted. Fang et al. (2017)

proposed a linguistic steganographic method that randomly partitioned the vocabulary Σ into 2^b bins $[B_1, B_2, \dots, B_{2^b}]$ and each one contained $|\Sigma|/2^b$ tokens. At each time step, they selected the token with the highest probability within the bin according to the b -bit secret information to be embedded. Yang et al. (2018a) improved the embedding algorithm by building the mapping from secret information to tokens dynamically at each time step rather than statically in advance. Concretely, the top 2^k tokens with the highest probability were encoded by Huffman coding algorithm. Then they took the token which has the same code as the secret information. Dai and Cai (2019) proposed *patient-Huffman*, which was an improved version of Yang et al. (2018a) that sacrificed embedding capacity for imperceptibility. They first calculated the distortion (total variation distance or KL divergence) between q and p_{LM} and then only used Huffman coding embedding algorithm to embed secret information when the distortion was less than a preset threshold δ . Otherwise they directly sampled a token to avoid high distortion occasions. Ziegler et al. (2019) employed arithmetic coding to embed secret information. They truncated the top h likely tokens and left out the low-probability long-tails. Then the tokens are encoded by arithmetic coding algorithm and selected according to the secret information. Compared with other coding algorithm, arithmetic coding has higher compression rate, which results in less damage to conditional probability distribution p_{LM} and helps to improve imperceptibility.

4 ADG Methodology

According to the analysis in Section 2.3, the distortion of generative linguistic steganography includes the bias of the language model \mathcal{L} and the damage to the conditional distribution caused by the embedding algorithm f_{emb} . The former is not our research priority. With the development of automatic text generation, the former distortion can be gradually minimized. In this paper, we mainly pay attention to the latter distortion. We aim to seek an optimal solution theoretically and experimentally.

Given an off-the-shelf language model, how can we embed secret information to the generated tokens? Unlike previous works that encoded the conditional distribution by lossless coding algorithm, we achieve this goal in a novel way by grouping. Through mathematical analysis and proof, we propose a provably secure method ADG, which does

little damage to the conditional distribution and is nearly equivalent to directly sampling from the full distribution. In this section, we investigate the security of steganography by grouping and give detailed descriptions of the proposed method.

4.1 Steganography by Grouping

Steganography by grouping is to group all tokens in the vocabulary into several groups, so that each group represents a unique secret message. E.g. we can Tokens belonging to the target group are able to make up the stegotext. In such a way, Bob reads each token in the sequence in turn and performs the same grouping operation to extrapolate which groups the current token belongs to, thereby extracting the corresponding secret information. The key question is: how to group the tokens at each time step to ensure an optimal imperceptibility? We have the following assumption.

Assumption 1. *For secret information in the form of uniformly distributed bitstream, adaptively grouping the vocabulary into u groups ($u = 2^r, r \in \mathbb{N}, r \leq \log_2 |\Sigma|$) with equal probability will ensure the optimal imperceptibility.*

Proof. Assuming that the discrete conditional probability distribution p_{LM} is arbitrarily partitioned into u groups to embed r -bit secret information. p_{ij} denotes the probability of the j -th token in the i -th group. η_i and n_i denote the total probability and the size of the i -th group respectively. Then we have

$$\sum_{j=1}^{n_i} p_{ij} = \eta_i, \quad \sum_{i=1}^u \eta_i = 1. \quad (2)$$

Our goal is to figure out the grouping algorithm to achieve the best imperceptibility, i.e. to minimize the gap between p_{LM} and q . First of all, starting from the modeled distribution $p_{LM} = [\dots, p_{ij}, \dots]$, we calculate the equivalent distribution q . The probability of each token is firstly normalized within its group ($1/\eta_i$) and then multiplied by the selected probability of the group, which is $1/u$ since secret information is uniformly distributed. Therefore, q has the following form

$$q = [\dots, p_{ij}/u\eta_i, \dots]. \quad (3)$$

We measured the gap between the two distributions with KL divergence, which is

$$D_{KL}(p_{LM}||q) = \sum p_{LM} \log \frac{p_{LM}}{q}$$

$$\begin{aligned}
&= \sum_{i=1}^u \sum_{j=1}^{n_i} p_{ij} \log \frac{p_{ij}}{p_{ij}/u\eta_i} \\
&= \sum_{i=1}^u \sum_{j=1}^{n_i} p_{ij} \log(u\eta_i) \\
&= \sum_{i=1}^u \log(u\eta_i) \sum_{j=1}^{n_i} p_{ij} \\
&= \sum_{i=1}^u \eta_i \log(u\eta_i). \tag{4}
\end{aligned}$$

Therefore, the KL divergence between the two distributions is a function of the vector $\boldsymbol{\eta} = [\eta_1, \eta_2, \dots, \eta_u]$.

Next, we will prove Assumption 1 in two steps.

[1]. Considering the auxiliary function $f_{aux}(\eta) = \eta \log(u\eta)$, ($0 \leq \eta \leq 1$), we firstly analyse its concavity and convexity on the domain of definition. For every $\eta_1, \eta_2 \in (0, 1)$ and $0 \leq \lambda \leq 1$,

$$\begin{aligned}
&f_{aux}(\lambda\eta_1 + (1-\lambda)\eta_2) \\
&- \lambda f_{aux}(\eta_1) - (1-\lambda)f_{aux}(\eta_2) \\
&= (\lambda\eta_1 + (1-\lambda)\eta_2) \log(u(\lambda\eta_1 + (1-\lambda)\eta_2)) \\
&- \lambda(\eta_1 \log(u\eta_1)) - (1-\lambda)(\eta_2 \log(u\eta_2)) \\
&= \lambda\eta_1 \log \frac{\lambda\eta_1 + (1-\lambda)\eta_2}{\eta_1} \\
&+ (1-\lambda)\eta_2 \log \frac{\lambda\eta_1 + (1-\lambda)\eta_2}{\eta_2} \\
&\leq \lambda(\eta_1 (\frac{\lambda\eta_1 + (1-\lambda)\eta_2}{\eta_1} - 1)) \\
&+ (1-\lambda)\eta_2 (\frac{\lambda\eta_1 + (1-\lambda)\eta_2}{\eta_2} - 1) \\
&= \lambda(\lambda\eta_1 + (1-\lambda)\eta_2 - \eta_1) \\
&+ (1-\lambda)(\lambda\eta_1 + (1-\lambda)\eta_2 - \eta_2) \\
&= \lambda(\lambda-1)\eta_1 + \lambda(1-\lambda)\eta_2 \\
&+ (1-\lambda)\lambda\eta_1 - (1-\lambda)\lambda\eta_2 \\
&= 0. \tag{5}
\end{aligned}$$

As a result, $f_{aux}(\eta)$ is convex over $(0, 1)$.

[2]. Then, when generalizing to u variables $\eta_1, \eta_2, \dots, \eta_u$, $\sum_{i=1}^u \eta_i = 1$, according to Jensen's inequality (Jensen et al., 1906), there is

$$\begin{aligned}
&\frac{\sum_{i=1}^u f_{aux}(\eta_i)}{u} = \frac{\sum_{i=1}^u \eta_i \log(u\eta_i)}{u} \\
&\geq f_{aux}(\frac{\sum_{i=1}^u \eta_i}{u}) = \frac{1}{u} \sum_{i=1}^u \eta_i \log(\sum_{i=1}^u \eta_i) = 0. \tag{6}
\end{aligned}$$

The equality sign holds if and only if

$$\eta_1 = \eta_2 = \dots = \eta_u. \tag{7}$$

It means that $D_{KL}(p_{LM}||q) = \sum_{i=1}^u f_{aux}(\eta_i)$ takes the minimum value 0 when each component of $\boldsymbol{\eta}$ is equal, in which case p_{LM} and q are equivalent and that achieves the optimal information-theoretic security defined by Cachin (1998). \square

Therefore, we basically construct the idea of our embedding algorithm, that is, to adaptively group the vocabulary into multiple groups at each time step, so that each group is assigned approximately the same probability. In practice, since the probability distribution is discrete, the probability of groups may not be absolutely equal. Firstly, we determine the number of groups u to be its maximum value $2^{\lfloor -\log_2 p_{max} \rfloor}$, where p_{max} is the highest probability in p_{LM} . Secondly, since the time complexity of solving the global optimal solution of equal grouping is unacceptable, we implement a suboptimal solution in ADG, as demonstrated in Algorithm 1. In line 10, we employ binary search algorithm to select the token that has the nearest probability of a given value. Our implementation enables us to obtain a unique grouping result for any p_{LM} , which ensures that the secret information can be extracted accurately and completely at the receiving end.

4.2 Recursion and Pruning

After obtaining the grouping results, we can select the group according to the next $\log u$ bits of secret information to be embedded and simply sample a token in the group to generate stegotext. As a matter of fact, we can also continue grouping the obtained groups to further enlarge the embedding capacity and recursively grouping the new groups until it is impossible to be equally participated (the normalized p_{max} of the current group is greater than 0.5). In order to improve the efficiency of the recursive grouping, we employ pruning strategy to remove the redundant grouping operations. We only need to recursively group the selected groups every time in accordance with the secret information to be embedded. In this manner, the amount of secret information embedded in each token is adjusted dynamically according to its probability distribution.

To sum up, at each time step, the proposed ADG embedding algorithm first conducts the equal

Algorithm 1: Suboptimal solution of equal grouping.

Data: vocabulary Σ , distribution p_{LM} **Result:** set of groups G

```
1 list of tokens = sorted ( $p_{LM}$ );
2  $p_{max}$  = probability of the first token;
3  $u = 2^{\lfloor -\log_2 p_{max} \rfloor}$ ;
4 mean =  $1/u$ ;
5 for ( $i = 1; i \leq u - 1; i ++$ ) do
6    $G_i =$  [the first token];
7   remove the first token;
8   while  $\sum$  probability of  $G_i < mean$  do
9      $\varepsilon = mean - \sum$  probability of  $G_i$ ;
10    select a token with the nearest
11    probability of  $\varepsilon$ ;
12    if probability of the token  $- \varepsilon < \varepsilon$ 
13      then
14        append the token to  $G_i$ ;
15        remove the token;
16      end
17    else
18      break;
19    end
20  end
21  mean =  $\frac{\text{probability of the rest tokens}}{u - i}$ ;
22 end
23 append the rest tokens to  $G_u$ ;
24  $G = [G_1, G_2, \dots, G_u]$ ;
```

grouping algorithm adaptively according to the conditional distribution, and then recursively repeats the operation on the selected group dynamically according to the secret information, until it is indivisible. At last, we normalize the probability of the last selected group and sample a token to generate the stegotext. We have proved the security of equal grouping algorithm. Obviously, it can also be extended to the recursive manner of ADG, which means the proposed method is provably secure.

4.3 Information Extraction

The extraction algorithm is basically the inverse process of the embedding algorithm. For an exactly successful extraction, Alice and Bob have to share the same language model, vocabulary and grouping algorithm. At each time step, Bob is supposed to recursively operate the same grouping algorithm as Alice do, and then select the group

contains the current token in the stegotext. The index of the selected groups reveal the embedded secret information.

5 Experimental Results and Analysis

In this section, we evaluate the performance of ADG in terms of both embedding capacity and imperceptibility. Details of our experiments and the analysis of the results are present in the following subsections.

5.1 Datasets

We evaluated the performance of ADG on three public corpora, namely “Large Movie Review Dataset” (**Movie**) (Maas et al., 2011), “All the News” (**News**)¹ and “Sentiment140” (**Tweet**) (Go et al., 2009). Large movie review dataset is originally built for binary sentiment classification, containing 100,000 movie reviews in total crawled from IMDb². “All the news” is a collection of publications of mainstream news media. Sentiment140 is also used in sentiment analysis tasks, which contains 1,600,000 tweets extracted from Twitter³.

We converted the raw text to lowercase and removed HTML tags and most punctuations, then segmented it into sentences with NLTK tools (Loper and Bird, 2002). We filtered out sentences with length below 5 or above 200. For the convenience of training and evaluation, any token occurring less than 10 times was mapped to a special token “_UNK”. We also added “_BOS” and “_EOS” at the beginning and end of each sentence to help training. Sentences in a batch were padded to the same length with a special padding token “_PAD”. Finally, we divided the preprocessed corpora into training set and test set according to the ratio of 9:1. Statistics are demonstrated in Table 3.

5.2 Implementation Details

In experiments, we utilized LSTMs (Hochreiter and Schmidhuber, 1997) for word-level generation. We stacked 2 LSTM layers and the model was implemented with Pytorch (Paszke et al., 2017). The dimension of word embedding was set to be 350. Hidden states in LSTM were set to be 512-dimensional vectors. In the training procedure, we applied SGD algorithm together with Adam

¹<https://www.kaggle.com/snapcrack/all-the-news>

²<https://www.imdb.com/>

³<https://twitter.com/>

Table 1: Results of ER, KLD_1 and KLD_2 .

METHOD	Movie			News			Tweet		
	\uparrow ER	\downarrow KLD ₁	\downarrow KLD ₂	\uparrow ER	\downarrow KLD ₁	\downarrow KLD ₂	\uparrow ER	\downarrow KLD ₁	\downarrow KLD ₂
Bins ($b = 1$)	1.000	2.497	27.595	1.000	2.742	26.331	1.000	2.431	19.519
Bins ($b = 2$)	2.000	2.338	33.206	2.000	2.593	35.207	2.000	2.421	17.604
Bins ($b = 3$)	3.000	2.319	29.778	3.000	2.592	55.781	3.000	2.429	21.286
Bins ($b = 4$)	4.000	2.439	54.155	4.000	2.550	87.441	4.000	2.314	27.230
Bins ($b = 5$)	5.000	2.503	73.075	5.000	2.500	116.857	5.000	2.482	29.171
Huffman ($k = 1$)	1.000	1.961	21.219	1.000	2.338	11.226	1.000	2.121	6.252
Huffman ($k = 2$)	1.824	1.433	13.199	1.824	1.751	8.793	1.841	1.586	5.208
Huffman ($k = 3$)	2.509	1.106	8.487	2.518	1.372	6.855	2.595	1.145	4.141
Huffman ($k = 4$)	3.145	0.819	6.334	3.224	1.084	5.419	3.266	0.880	3.197
Huffman ($k = 5$)	3.705	0.658	4.657	3.872	0.838	3.995	3.932	0.694	2.738
Patient-Huffman ($\delta = 1.0$)	1.125	0.327	0.767	0.809	0.256	0.441	0.988	0.298	0.545
Patient-Huffman ($\delta = 1.5$)	1.711	0.588	2.132	1.460	0.559	1.817	1.668	0.621	1.280
Patient-Huffman ($\delta = 2.0$)	2.129	0.819	4.564	1.905	0.808	3.497	2.201	0.908	2.445
Arithmetic ($h = 100$)	4.224	0.362	2.956	4.412	0.425	2.269	4.308	0.333	1.508
Arithmetic ($h = 200$)	4.651	0.240	2.321	4.908	0.295	1.688	4.805	0.253	1.749
Arithmetic ($h = 300$)	4.903	0.205	1.903	5.127	0.245	1.426	4.942	0.206	1.242
ADG	5.147	0.033	1.946	5.650	0.027	0.866	5.411	0.048	1.189

(Kingma and Ba, 2014) to train the language model. Learning rate was set to be 0.001. The SGD update direction was computed using a batch of 32 training samples. They were both trained for 30 epochs on one GeForce GTX 1080 GPU. In the generation procedure, we adopted the model performing best on test sets. All generated sentences must be longer than 5 and shorter than 200.

5.3 Baselines

We rebuilt Fang et al. (2017) (Bins), Yang et al. (2018a) (Huffman), Dai and Cai (2019) (Patient-Huffman) and Ziegler et al. (2019) (Arithmetic) as baselines. For fair comparison, we rebuilt all the baselines with the same language models. For Bins, we set b to be 1, 2, 3, 4, 5 and the corresponding number of bins was 2, 4, 8, 16, 32. For Huffman, we built Huffman tree with the top 2, 4, 8, 16, 32 likely tokens. For Patient-Huffman, we measured the distortion by KL divergence and restricted the threshold δ to 1, 1.5, 2 with top 8 tokens. For Arithmetic, we truncated the conditional distribution at $h = 100, 200, 300$. In each case, we generated 1,000 stegotext. We randomly chose same amount of coverttext from the test sets for further evaluation.

5.4 Metrics

The metrics we utilized to evaluate the performance on embedding capacity and imperceptibility are listed as follows.

Embedding Rate (ER): It is the average amount of information that one single token can carry, and is in unit of *bits per word (bpp)*. Embedding rate is a metric to indicate the embedding capacity. Higher is better.

KL Divergence between the implicit distribution q and the modeled distribution p_{LM} (KLD_1): It reflects the gap introduced by the embedding algorithm. Lower is better and the unit is *bit*.

KL Divergence between the statistical distributions of the sentence embedding of coverttext and stegotext (KLD_2): It indirectly reflects the overall information-theoretic security. We mapped all stegotext and coverttext to fixed length dense vectors \mathbf{v}_x and \mathbf{v}_y by third-party sentence vectorization tool (Le and Mikolov, 2014), and assumed that the resulting vectors of coverttext and stegotext both obey isotropic Gaussian distribution. Then KLD_2 is computed by

$$D_{KL}(p(\mathbf{v}_x)||p(\mathbf{v}_y)) \approx \sum (\log \frac{\sigma_y}{\sigma_x} + \frac{\sigma_x^2 + (\mu_x - \mu_y)^2}{2\sigma_y^2} - \frac{1}{2}), \quad (8)$$

where μ and σ are the mean and standard deviation of sentence vectors. We set the dimension of sentence vectors to be 100. Lower is better and the unit is *bit*.

Detection Accuracy: It reflects the anti-steganalysis ability of steganographic methods. Ste-

Table 2: Results of ER, Acc₁ and Acc₂.

METHOD	Movie			News			Tweet		
	↑ER	Acc ₁	Acc ₂	↑ER	Acc ₁	Acc ₂	↑ER	Acc ₁	Acc ₂
Bins ($b = 1$)	1.000	0.873	0.854	1.000	0.887	0.856	1.000	0.787	0.814
Bins ($b = 2$)	2.000	0.812	0.802	2.000	0.855	0.830	2.000	0.739	0.753
Bins ($b = 3$)	3.000	0.810	0.789	3.000	0.833	0.819	3.000	0.720	0.733
Bins ($b = 4$)	4.000	0.825	0.832	4.000	0.843	0.852	4.000	0.748	0.760
Bins ($b = 5$)	5.000	0.876	0.872	5.000	0.877	0.882	5.000	0.750	0.786
Huffman ($k = 1$)	1.000	0.891	0.891	1.000	0.891	0.885	1.000	0.785	0.806
Huffman ($k = 2$)	1.824	0.838	0.836	1.824	0.851	0.826	1.841	0.749	0.758
Huffman ($k = 3$)	2.509	0.796	0.760	2.518	0.816	0.785	2.595	0.684	0.702
Huffman ($k = 4$)	3.145	0.713	0.690	3.224	0.768	0.718	3.266	0.634	0.632
Huffman ($k = 5$)	3.705	0.673	0.645	3.872	0.710	0.664	3.932	0.602	0.593
Patient-Huffman ($\delta = 1.0$)	1.125	0.588	0.578	0.809	0.559	0.542	0.988	0.528	0.552
Patient-Huffman ($\delta = 1.5$)	1.711	0.654	0.683	1.460	0.674	0.683	1.668	0.589	0.581
Patient-Huffman ($\delta = 2.0$)	2.129	0.722	0.714	1.905	0.725	0.726	2.201	0.650	0.661
Arithmetic ($h = 100$)	4.224	0.601	0.582	4.412	0.630	0.608	4.308	0.547	0.554
Arithmetic ($h = 200$)	4.651	0.565	0.556	4.908	0.594	0.559	4.805	0.537	0.561
Arithmetic ($h = 300$)	4.903	0.571	0.562	5.127	0.558	0.566	4.942	0.532	0.534
ADG	5.147	0.548	0.544	5.650	0.543	0.519	5.411	0.496	0.497

Table 3: Statistics of the preprocessed datasets.

DATASET	$ \Sigma $	TRAINING	TEST
Movie	37,800	1,002,609	111,402
News	50,178	1,461,567	162,397
Tweet	30,152	1,572,599	174,734

ganalysis is the technology used by Eve to detect hidden information in stegocarriers, which is the opposite direction of steganography. In our experiment, we employed linguistic steganalysis approaches based on Fasttext (Yang et al., 2019) (Acc₁) and TextCNN (Yang et al., 2020b) (Acc₂). We took stegotext as positive samples and covertext as negative samples. We conducted 10-fold cross validation and reported the average accuracy. Closer to 50% is better.

Effective Embedding Rate: It is a new metric we proposed to evaluate the comprehensive performance of steganographic algorithms. It is defined to be calculated by

$$\text{EER} = 2 \times (1 - \text{Acc}) \times \text{ER}, \quad (9)$$

meaning that if the stegotext has a certain probability of being detected, the average amount of secret information actually transmitted should be discounted accordingly. For mathematical rigorosity and completeness, if $\text{Acc} < 0.5$, we assign $1 - \text{Acc}$ to Acc . In extreme cases where the stegocarriers are completely natural, the detection accuracy should be 50% and EER is equal to ER. On

the contrary, stegocarriers with 100% detection accuracy cannot carry a single bit. We calculated this metric with the accuracy results obtained by the two aforementioned steganalysis method (EER₁, EER₂). Higher is better and the unit is *bpp*.

5.5 Results and Analysis

The results of KLD₁ and KLD₂ are listed in Table 1. KLD₁ measures the distortion between q and p_{LM} , which is introduced by the embedding algorithm ADG. KLD₂ estimates the overall information-theoretic security that also considers the deviation of language models. In terms of KLD₁, we found that the proposed method ADG outperforms all baselines and it is very close to the optimal value 0 (stochastic sampling), which means generating stegotext by ADG is almost equivalent to normal generation with the language models. The results of KLD₂ are also advantageous, indicating that the generated stegotext is statistically consistent with the covertext. We noticed that some baselines can also perform well on KLD₂ (e.g. Patient-Huffman ($\delta = 1.0$)). However, they have a crucial flaw in embedding capacity.

Table 2 demonstrates the results of anti-steganalysis, where we found the tendency coheres with that of KLD₁ and KLD₂. The proposed method ADG outperforms all baselines on the three corpora and it is very close to the optimal value 0.5, which further confirms its imperceptibility. Besides, we also illustrated some examples of stegotext generated by ADG in Table 5 for qualitative

Table 4: Results of ER, EER₁ and EER₂.

METHOD	Movie			News			Tweet		
	↑ER	↑EER ₁	↑EER ₂	↑ER	↑EER ₁	↑EER ₂	↑ER	↑EER ₁	↑EER ₂
Bins ($b = 1$)	1.000	0.254	0.292	1.000	0.226	0.287	1.000	0.425	0.373
Bins ($b = 2$)	2.000	0.752	0.794	2.000	0.582	0.680	2.000	1.044	0.988
Bins ($b = 3$)	3.000	1.137	1.266	3.000	0.999	1.089	3.000	1.683	1.605
Bins ($b = 4$)	4.000	1.396	1.344	4.000	1.252	1.180	4.000	2.020	1.924
Bins ($b = 5$)	5.000	1.245	1.280	5.000	1.230	1.180	5.000	2.500	2.135
Huffman ($k = 1$)	1.000	0.218	0.219	1.000	0.219	0.231	1.000	0.430	0.387
Huffman ($k = 2$)	1.824	0.593	0.600	1.824	0.546	0.635	1.841	0.924	0.893
Huffman ($k = 3$)	2.509	1.024	1.202	2.518	0.927	1.083	2.595	1.638	1.549
Huffman ($k = 4$)	3.145	1.809	1.950	3.224	1.496	1.821	3.266	2.387	2.404
Huffman ($k = 5$)	3.705	2.427	2.627	3.872	2.249	2.602	3.932	3.133	3.200
Patient-Huffman ($\delta = 1.0$)	1.125	0.927	0.949	0.809	0.713	0.740	0.988	0.933	0.886
Patient-Huffman ($\delta = 1.5$)	1.711	1.182	1.083	1.460	0.952	0.925	1.668	1.369	1.400
Patient-Huffman ($\delta = 2.0$)	2.129	1.184	1.220	1.905	1.050	1.044	2.201	1.541	1.490
Arithmetic ($h = 100$)	4.224	3.371	3.527	4.412	3.269	3.459	4.308	3.908	3.843
Arithmetic ($h = 200$)	4.651	4.051	4.125	4.908	3.981	4.324	4.805	4.449	4.219
Arithmetic ($h = 300$)	4.903	4.207	4.290	5.127	4.532	4.450	4.942	4.630	4.606
ADG	5.147	4.648	4.699	5.650	5.164	5.435	5.411	5.373	5.384

Table 5: Examples of stegotext generated by ADG on the three corpora.

Movie	The supporting cast was also excellent. But I guess you 've seen the many silent movies along with his other films. And this movie was a precursor of val kilmer in the extreme. It 's a unique wonderful movie that deserves all the recognition it deserved. This is the worst movie I have ever seen.
News	The FBI estimated its total wealth on Thursday. Remember this is in part because of the actual policies of Donald Trump. He said he did not care about any counterintelligence investigation. Today however the process could not change even if he doesnt agree with Trumps rhetoric. More than 100 000 people have been detained and another 30 000 civilians have been wounded early on Sunday.
Tweet	Worst headache everrrr I dunno why but it was so scary. I had a blast today in the MTV Movie Awards. Ahhh some brothers do n't play sports! Sadly you will be missing so much. I do n't think the peach ice cream last night was good.

study. We found that the stegotext is fluent enough, with correct grammar and coherent semantics.

Finally, taking both embedding capacity and imperceptibility into account, we investigated effective embedding rate listed in Table 4. It can be concluded that our method has excellent comprehensive performance, which outperforms all baselines. In general, the experimental results indicate that the proposed method ADG is able to resist both perceptual and statistical steganalysis of Eve, meanwhile ensure a remarkable embedding rate, which reveals its effectiveness.

6 Conclusion

Previous works of generative linguistic steganography inevitably introduce distortions to the distribution estimated by off-the-shelf language models. In this paper, we attempted to achieve provably se-

cure generative linguistic steganography during the procedure of stegotext generation. We proposed ADG, which embeds secret information by adaptive dynamic grouping. According to the mathematical proof and extensive experiments conducted on three public corpora, we found that the proposed method is provably secure and capable of generating fluent stegotext with high embedding capacity and high imperceptibility. We hope our investigation of provably secure generative linguistic steganography can be leveraged as a building block for future research.

Acknowledgments

This work is supported in part by the National Key Research and Development Program of China under Grant U1936216, Grant 61862002, Grant 6200197 and Grant U1705261. The authors thank

anonymous reviewers for their insightful suggestions.

References

- Ross J Anderson and Fabien AP Petitcolas. 1998. On the limits of steganography. *IEEE Journal on selected areas in communications*, 16(4):474–481.
- Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.
- Christian Cachin. 1998. An information-theoretic model for steganography. In *International Workshop on Information Hiding*, pages 306–318. Springer.
- Mark Chapman and George Davida. 1997. Hiding the hidden: A software system for concealing ciphertext as innocuous text. In *International Conference on Information and Communications Security*, pages 335–345. Springer.
- Mark Chapman and George Davida. 2002. Plausible deniability using automated linguistic steganography. In *International Conference on Infrastructure Security*, pages 276–287. Springer.
- Mark Chapman, George I Davida, and Marc Rennhard. 2001. A practical and effective approach to large-scale automated linguistic steganography. In *International Conference on Information Security*, pages 156–165. Springer.
- Falcon Dai and Zheng Cai. 2019. Towards near-imperceptible steganographic text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4303–4308.
- Weihui Dai, Yue Yu, Yonghui Dai, and Bin Deng. 2010. Text steganography system using markov chain source model and des algorithm. *JSW*, 5(7):785–792.
- Tina Fang, Martin Jaggi, and Katerina Argyraki. 2017. Generating steganographic text with lstms. In *Proceedings of ACL 2017, Student Research Workshop*, pages 100–106.
- Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, 1(12):2009.
- Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.
- Mehdi Hussain, Ainuddin Wahid Abdul Wahab, Yamani Idna Bin Idris, Anthony TS Ho, and Ki-Hyun Jung. 2018. Image steganography in spatial domain: A survey. *Signal Processing: Image Communication*, 65:46–66.
- Johan Ludwig William Valdemar Jensen et al. 1906. Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta mathematica*, 30:175–193.
- Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- R Bala Krishnan, Prasanth Kumar Thandra, and M Sai Baba. 2017. An overview of text steganography. In *2017 Fourth International Conference on Signal Processing, Communication and Networking (IC-SCN)*, pages 1–6. IEEE.
- Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196.
- Yunxia Liu, Shuyang Liu, Yonghao Wang, Hongguo Zhao, and Si Liu. 2019. Video steganography: A review. *Neurocomputing*, 335:238–250.
- Edward Loper and Steven Bird. 2002. Nltk: the natural language toolkit. *arXiv preprint cs/0205028*.
- Yubo Luo, Yongfeng Huang, Fufang Li, and Chinchun Chang. 2016. Text steganography based on ci-poetry generation using markov chain model. *KSII Transactions on Internet and Information Systems (TIIS)*, 10(9):4568–4584.
- Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 142–150.
- Shilpi Mishra, Virendra Kumar Yadav, Munesh Chandra Trivedi, and Tarun Shrimali. 2018. Audio steganography techniques: A survey. In *Advances in Computer and Computational Sciences*, pages 581–589. Springer.
- H Hernan Moraldo. 2014. An approach for text steganography based on markov chains. *arXiv preprint arXiv:1409.0915*.
- Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.
- Shivani Sharma, Avadhesh Gupta, Munesh Chandra Trivedi, and Virendra Kumar Yadav. 2016. Analysis of different text steganography techniques: a survey. In *2016 Second International Conference on Computational Intelligence & Communication Technology (CICT)*, pages 130–133. IEEE.

- Mohammad Shirali-Shahreza. 2008. Text steganography by changing words spelling. In *2008 10th international conference on advanced communication technology*, volume 3, pages 1912–1913. IEEE.
- Gustavus J Simmons. 1984. The prisoners’ problem and the subliminal channel. In *Advances in Cryptology*, pages 51–67. Springer.
- Peter Wayner. 1992. Mimic functions. *Cryptologia*, 16(3):193–214.
- Lingyun Xiang, Yan Li, Wei Hao, Peng Yang, and Xiaobo Shen. 2018. Reversible natural language watermarking using synonym substitution and arithmetic coding. *Comput. Mater. Continua*, 55(3):541–559.
- Lingyun Xiang, Xingming Sun, Gang Luo, and Bin Xia. 2014. Linguistic steganalysis using the features derived from synonym frequency. *Multimedia tools and applications*, 71(3):1893–1911.
- Zhong-Liang Yang, Xiao-Qing Guo, Zi-Ming Chen, Yong-Feng Huang, and Yu-Jin Zhang. 2018a. Rnn-stega: Linguistic steganography based on recurrent neural networks. *IEEE Transactions on Information Forensics and Security*, 14(5):1280–1295.
- Zhong-Liang Yang, Si-Yu Zhang, Yu-Ting Hu, Zhi-Wen Hu, and Yong-Feng Huang. 2020a. Vae-stega: Linguistic steganography based on variational auto-encoder. *IEEE Transactions on Information Forensics and Security*, 16:880–895.
- Zhongliang Yang, Yongfeng Huang, and Yu-Jin Zhang. 2019. A fast and efficient text steganalysis method. *IEEE Signal Processing Letters*, 26(4):627–631.
- Zhongliang Yang, Yongfeng Huang, and Yu-Jin Zhang. 2020b. Ts-csw: text steganalysis and hidden capacity estimation based on convolutional sliding windows. *Multimedia Tools and Applications*, 79(25):18293–18316.
- Zhongliang Yang, Shuyu Jin, Yongfeng Huang, Yujin Zhang, and Hui Li. 2018b. Automatically generate steganographic text based on markov model and huffman coding. *arXiv preprint arXiv:1811.04720*.
- Xuejing Zhou, Wanli Peng, Boya Yang, Juan Wen, Yiming Xue, and Ping Zhong. 2021. Linguistic steganography based on adaptive probability distribution. *IEEE Transactions on Dependable and Secure Computing*.
- Zachary Ziegler, Yuntian Deng, and Alexander M Rush. 2019. Neural linguistic steganography. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1210–1215.