语义空间下基于情感表达的生成式文本隐写方法

刘玉玲[1]，王翠林[1]，付章杰[2]

（1. 湖南大学信息科学与工程学院，湖南 长沙 410082；
2. 南京信息工程大学计算机学院、软件学院、网络空间安全学院，江苏 南京 210044）

摘 要：针对现有生成式文本隐写方法存在的“过度优化”文本质量以及生成的隐写文本在语义表达上缺乏约束等问题，提出了一种在语义空间下基于情感表达的生成式文本隐写方法。该方法利用新媒体平台提供的情景融合的伪装场景，研究如何利用无监督抽取模型从原始数据集中抽取情感表达组合候选集合，并基于改进的二部图排序算法对情感表达组合候选集合进行排序，得到情感表达组合集合；然后将其映射到语义空间，实现基于情感表达组合生成用户观点的同时嵌入秘密信息。实验结果表明，与同类语义空间下生成式文本隐写方法相比，所提方法生成的含密商品评论的困惑度最低可达 10.536，且含密商品评论与主题具有较强相关性，进一步保证了隐写文本的认知隐蔽性，同时所提方法还可有效地用于安全保密通信领域，能够避免发送方被追踪溯源和关联分析。

关键词：生成式文本隐写；语义空间；无监督抽取模型；情感表达

中图分类号：TP309

文献标志码：A

DOI: 10.11959/j.issn.1000−436x.2023045

Generative text steganography method based on emotional expression in semantic space

LIU Yuling[1], WANG Cuilin[1], FU Zhangjie[2]

1. School of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, China
2. School of Computer Science, Nanjing University of Information Science and Technology, Nanjing 210044, China

Abstract: Aiming at the problems that “over optimizing” the quality of steganographic text and lack of constraints on the semantic expression of the generated steganographic text in existing generative text steganography methods, a generative text steganography method was proposed based on emotional expression in semantic space. In order to make use of the scene fusion provided by the new media platform to obtain many camouflage scenes, the focus was how to use the unsupervised extraction model to extract the emotional expression combination candidate set from the original data set, then sort the candidate set of emotional expression combinations based on the improved bipartite graph sorting algorithm to obtain the emotional expression combination set, map them to the semantic space, and then implement embedding secret information while generating the user’s opinions based on the emotion expression combinations. Experimental results show that, compared with the existing generative text steganography methods in semantic space, the product reviews generated by the proposed method have a minimum perplexity of 10.536, and have a strong correlation with the chosen product, which can further guarantee the cognitive concealment of steganographic texts. At the same time, the proposed method can also be effectively used in the field of secure and confidential communication, and can avoid the senders being traced and analyzed.

Keywords: generative text steganography, semantic space, unsupervised extraction model, emotional expression

收稿日期：2022−10−05；**修回日期：**2023−01−04

基金项目：国家自然科学基金资助项目（No.61872134）；教育部科技发展中心基金资助项目（No.2019J01020）；长沙市科技计划基金资助项目（No.Kh2004004）；湖南省交通运输厅科技计划基金资助项目（No.201935）

Foundation Items: The National Natural Science Foundation of China (No.61872134), The Science and Technology Development Center of the Ministry of Education (No.2019J01020), Science and Technology Project of Changsha City (No.Kh2004004), Science and Technology Project of Transport Department of Hunan Province (No.201935)

0　引言

隐写方法凭借其不可感知性的优势，现已成为网络安全领域的一个重要研究方向[1]。多年来，研究者一直致力于将秘密信息嵌入各种公共文本载体中[2-4]以实现隐蔽通信。受益于深度学习和自然语言处理技术的发展，近年来生成式文本隐写逐渐成为文本隐写领域的研究热点。

早在 1992 年，Wayner[5]就提出了基于 Mimic 的生成式文本隐写方法，生成的隐写文本中字符的统计分布规律近似正常文本，但生成的文本序列却不具备完整的语义信息，实用性较差。为了生成语义完整的隐写文本，Chapman 等[6]引入了句法结构模板，但这类方法生成的隐写文本模式单一，无法广泛适用于多个场景。Desoky[7]结合不同的使用场景，提出了多种特殊文本生成式隐写方法。然而这些方法通常要求特定的使用场景，因此不具有普遍性。Luo 等[8]和 Yi 等[9]提出利用马尔可夫链（MC, Markov chain）生成宋词，利用长短期记忆（LSTM, long short-term memory）模型生成唐诗，并将秘密信息嵌入文本生成过程中。但是，诗词毕竟是一种特殊体裁的文本，在日常生活中并不经常使用。在此基础上，Yang 等[10]提出基于条件概率编码（Conproc, conditional probablity coding）的生成式文本隐写 RNN-Stega 模型。该模型不仅实现了自然文本的生成，而且嵌入率可达 20%以上，吸引了国内外大量研究人员的注意。他们在此后的两年中提出了多种基于 Conproc 框架的改进方法。例如，Ziegler 等[11]将语言模型替换为 GPT-2 模型，将哈夫曼编码改为算术编码，生成了更加符合统计语言模型的自然文本。Xiang 等[12]将自然语句建模成字符序列，利用 Char-RNN 模型获取字符级的条件概率分布。Nakajima 等[13]提出 Syndrome-Trellis 的动态符号编码方法。Zhou 等[14]采用生成式对抗网络（GAN, generative adversarial network）模型进行隐写文本生成，并将基于 Top-K 的候选池构建方式改为动态候选池构建，这些方法进一步提升了生成的隐写文本的感知隐蔽性。

虽然上述方法有效地提高了生成式文本隐写方法的安全性，但其生成的隐写文本主要强调“过度优化”文本质量，在语义表达上缺乏约束，而且其情感和主题都是不可控的，因此存在被第三方检测识别的风险。尤其在实际应用场景中，生成文本的内容和主题还应该符合特定的上下文语境。

为了解决这一问题，本文聚焦生成式文本隐写的隐蔽性和安全性，利用新媒体平台实现情景融合的协同隐蔽通信。近年来，各种新媒体平台被人们广泛使用，如电商平台、短视频平台、微博或社区论坛等，导致多媒体数据生成量日益剧增。由于新媒体平台上互联网用户本身会产生大量评论文本，在这些场景下检测其生成的文本显然不能作为隐写存在的证据，这为生成式文本隐写提供了一个合理的情景融合的应用场景。另一方面，针对传统的基于点对点通信的隐蔽通信协议容易引起攻击者怀疑等问题，利用新媒体平台场景，能够保证生成式文本隐写的行为隐蔽性，从而避免发送方被追踪溯源和关联分析。同时，针对现有的基于符号空间的生成式文本隐写方法的隐蔽性和安全性难以满足实际要求的问题，本文提出了一种在语义空间下基于情感表达的生成式文本隐写方法，并在此基础上借用某电商平台的商品评论生成作为应用场景。消费者如果想要网购一件商品，电商平台上该商品的评价往往成为其最直接、最有效的意见来源。因此利用新媒体平台上的商品评论生成实现隐蔽通信具有极广泛的适用场景，并且有较好的行为隐蔽性和较高的安全性。

本文主要的研究工作如下。

1) 提出了一种基于情感表达的生成式文本隐写方法，通过将秘密信息映射到情感空间以实现语义空间的隐蔽通信，与传统的基于符号空间的方法相比，本文方法具有较好的感知隐蔽性以及认知隐蔽性。

2) 结合当今新媒体平台迅猛发展的时代背景，提出了一种基于无监督模型抽取情感表达组合以生成商品评论的方法，实现了协同隐蔽通信。该方法不仅增强了生成式文本隐写方法的适用性，还可以避免发送方被追踪溯源和关联分析，有效提高了隐写方法的安全性。

3) 引入了生成式文本隐写领域的常用评价指标，从感知隐蔽性、认知隐蔽性以及安全性等多方面验证了本文方法的有效性。

1　相关技术

为了更好地抽取新媒体平台上用户评论的情感表达组合，自动生成语序完备且语义自然的用户评论。本节首先介绍 jieba_DSIE（decentralize solidification information extraction）方法以提高分词

质量，找出网络词汇中的“新词”，然后详细介绍无监督抽取模型的具体步骤及收敛方式。

1.1 适应商品评论的词库构建方法

常用分词工具 jieba[15]不适合新媒体平台上的用户评论等不规范的文本，因此本文使用凝固程度–自由运用程度的 DSIE 方法找出“新词”，构建自定义词库。具体过程如下，首先，提取出超过阈值的文本片段，通过计算其凝固程度和自由运用程度选择符合定义 1 和定义 2 的文本片段；然后，与已有词库进行对比，以此确定“新词”。假设一个文本片段能够成词，那么它出现的概率应当与文本片段拆分概率乘积的差值较小，且能够灵活地出现在各种不同的语境中，具有非常丰富的左邻字集合和右邻字集合。本文给出如下 2 个定义。

定义 1 文本片段 X 的凝固程度 NG(X)。令文本片段 $X=(X_1, X_2, X_3, \cdots, X_m)$，其中，$m$ 表示文本片段的长度，NG(·)表示凝固程度的计算函数。枚举文本片段 X 的凝固方式 $X=S+E$，其中，文本片段 $S=(X_1, X_2, X_3, \cdots, X_t)$，文本片段 $E=(X_{t+1}, X_{t+2}, X_{t+3}, \cdots, X_m)$，$t\in[1, m-1]$，$t$ 的不同取值表示不同的凝固方式。$p(X)$、$p(S)$、$p(E)$分别表示文本片段 X、S、E 在整个语料中出现的概率，具体计算方式为文本片段出现的频次与语料库所有词数之比，则文本片段 X 的凝固程度为

$$\mathrm{NG}(X)=\frac{p(X)}{p(S)p(E)} \tag{1}$$

NG(X)的值越大，意味着 S 和 E 这 2 个文本片段经常出现在一起，其组成新文本片段的可能性也就越大。因此本文选定所有 NG(X)的最小值作为文本片段 X 的凝固程度。

定义 2 文本片段 Y 的自由运用程度 FS (Y)，即文本片段 Y 的左邻字集合和右邻字集合的丰富程度。首先需要统计出词语 Y 在语料库中的邻字集合 $Y\_L=(U_1, U_2, U_3, \cdots, U_n)$，对应邻字出现概率为 $p_1, p_2, p_3, \cdots, p_n$，假设各个邻字的出现彼此独立。通过式(2)可计算出 Y 的左右信息熵，那么文本片段 Y 的自由运用程度为

$$\mathrm{FS}(Y)=\min\left(H(U_{\text{left}}),\ H(U_{\text{right}})\right)$$
$$H(U)=\mathrm{E}\left[-\log p_i\right]=-\sum_{i=1}^{n}p_i\log p_i \tag{2}$$

1.2 基于无监督抽取模型的情感表达方法

由于人工标注情感极性的数据集难以获取，因此本文采用一种无监督抽取的方法获取情感表达组合。首先通过分词工具获取情感表达组合候选集合，然后基于改进的排序算法对情感表达组合候选集合进行排序。对于排序靠后的集合利用语义相似度算法再次进行选择，最终得到情感表达组合。

1.2.1 情感表达组合候选集合的获取

利用词性信息可以获取到对象和观点词的候选集合[16]。抽取时，首先创建对象列表 Object_list[]和观点词列表 Opinion_list[]，计算列表长度函数 Len(·)，将抽取出的对象以及观点词添加到对应的列表中，情感表达组合的最大组合数为 Len (Object_list[])×Len(Object_list[])，然后将对象和观点词的连接词 conjunction 添加到对应情感表达组合的 combination 列表里。

为了更准确地反映情感表达组合 combination 和 conjunction 之间的映射关系，需要对 conjunction 做更细致的区分，如例 1 和例 2 所示，在抽取时，用(+, –)代表语序方向，将“观点词 + conjunction + 对象”形式的 conjunction 记作 conjunction +，而“对象 + conjunction + 观点词”形式的 conjunction 记作 conjunction – 。

例 1 好看（观点词）的（连接词）书本（对象）

例 2 商品（对象）的（连接词）使用方法（观点词）

在整个抽取过程中，还需要统计 combination 被 conjunction 匹配的具体次数，即“二元对–conjunction–匹配次数”形式，如例 3 所示。

例 3 [运行速度 – 快]——— 挺 –———1
[性价比 – 高]——— 非常 +———6

1.2.2 情感表达组合的排序

为了给后续排序提供大量的候选集合，本文采用一种改进的二部图排序算法用于挑选情感表达组合。其核心思想基于以下 4 个假设，为了叙述简洁，在假设中，本文用函数 grade(·)表示计算某一个 combination 或 conjunction 的分数，similar(·)表示采用余弦相似度算法计算相似度值的公式，num(·)表示统计某一个 combination 或 conjunction 的具体使用次数。

假设 1 if Function_matching(num(combination) == i, num(conjunction) == j) (i == 1 && $j > i$)
then grade(combination) ↑
end if

假设 2 if Function_matching(num(combination)

== i, num(conjunction) == j) (j == 1 && $i > j$)

then grade(conjunction) ↑

end if

基于上述假设，可以得到 combination 与 conjunction 之间的映射关系，如图 1 所示。为了使模型收敛，本文将得到的二部图结构转化成矩阵形式，其中，count 表示 combination 被 conjunction 匹配的具体次数，如图 2 所示。

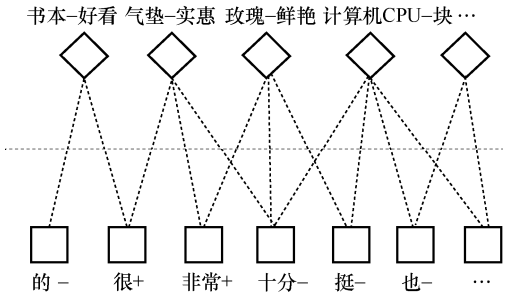


图 1　情感表达组合与连接词的二部图模型

$$\begin{array}{c} \quad\quad \text{conjunction}_1 \quad \cdots \quad \text{conjunction}_2 \quad \cdots \quad \text{conjunction}_r \\ \begin{matrix}\text{combination}_1\\ \vdots \\ \text{combination}_2 \\ \vdots \\ \text{combination}_n\end{matrix} \begin{bmatrix} \text{count}_{11} & \cdots & \text{count}_{12} & \cdots & \text{count}_{1r} \\ \vdots & & \vdots & & \vdots \\ \text{count}_{21} & \cdots & \text{count}_{22} & \cdots & \text{count}_{2r} \\ \vdots & & \vdots & & \vdots \\ \text{count}_{n1} & \cdots & \text{count}_{n2} & \cdots & \text{count}_{nr} \end{bmatrix} \end{array}$$

图 2　结构关系转化为矩阵

二部图排序算法的矩阵迭代计算式为

$$\begin{cases} \boldsymbol{N}_i = \boldsymbol{M}\boldsymbol{F}_i \\ \boldsymbol{N}_i' = \text{norm}(\boldsymbol{N}_i) \\ \boldsymbol{F}_{i+1} = \boldsymbol{M}^{\mathrm{T}}\boldsymbol{N}_i' \\ \boldsymbol{F}_{i+1}' = \text{norm}(\boldsymbol{F}_{i+1}) \end{cases} \tag{3}$$

其中，$\boldsymbol{M}$ 是由图 2 得到的二部图关系矩阵；$\boldsymbol{N}_i$ 和 $\boldsymbol{F}_i$ 均为一维矩阵，$\boldsymbol{F}_i$ 是 combination 的分数矩阵，初始化矩阵向量为 **1**，$\boldsymbol{N}_i$ 是 conjunction 对应的分数矩阵。为了保证不同类型商品的语料库之间具有可比性，使矩阵的每一维按照所占的比例重新分配分数值，本文在每一次矩阵运算后对 $\boldsymbol{N}_i$ 和 $\boldsymbol{F}_i$ 分别进行标准化处理，得到矩阵 $\boldsymbol{N}_i'$ 和 $\boldsymbol{F}_i'$。标准化处理的计算示例如式(4)所示。

假设 3 if the function similar(conjunction_1, conjunction_2) > 0.5 of conjunction_1

then choose conjunction_2

end if

假设 4 if the function similar(combination_1, combination_2) > 0.5 of combination_1

then choose combination_2

end if

对结果进行采样分析，发现仅采用二部图排序算法在召回率方面有一定的缺陷。排序完成后，依然存在一些正确的 combination 被排在了靠后的位置。其原因为选择的语料库不同，导致同一个 conjunction 在不同的语料库中出现的次数具有明显的差异。因此本文基于假设 3 和假设 4 进一步提出改进方法：事先利用 word2vec[17]模型将语料库训练成模型文件，计算得分较低的 combination_2 与得分较高的 combination_1 的相似度分数 grade，选择分数超过阈值的情感表达组合，得到最终的情感表达组合集合。表 1 显示了本文对几个主流电商平台评论抽取的结果统计。

表 1　　**数据来源与抽取数量**

数据集	对象属性列表/个	连接词/个
网易考拉	73 321	84 923
淘宝	13 215	20 329
京东	24 312	34 578

2　基于情感表达的信息隐藏方法

本节以某电商平台的商品评论生成为例，首先对本文方法进行全面概述，然后详细介绍信息嵌入与提取算法，最后通过一个实例给出算法的适用场景。

2.1　方法概述

本文利用无监督抽取模型从自然商品评论中抽取 Top-K 个情感表达组合，将其映射到不同的语义空间，并采用索引下标对语义空间进行编码，最后根据秘密信息选择特定的语义空间生成商品评论。图 3 显示了基于情感表达的信息隐藏方法的流程。

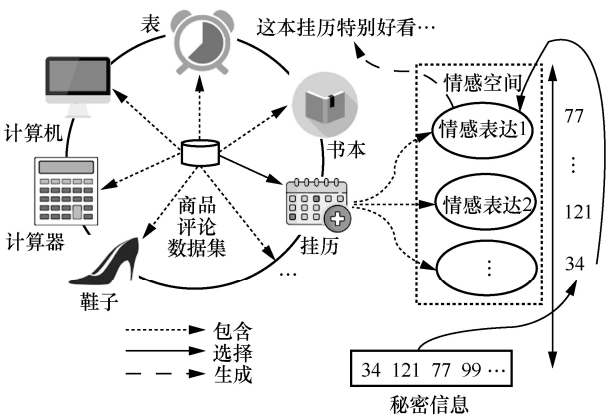


图 3　基于情感表达的信息隐藏方法的流程

大量商品的自然评论都是用户通过特定的电商平台发布对某商品的个人评价，因此任何人都可以在该平台上进行访问、浏览、发布等操作，这为隐蔽通信提供了极大的应用场景。隐蔽通信的消息发送方不需要将模型生成的含密商品评论点对点地发送给接收方，只需要将含密商品评论发布在公开的电商平台上，并事先与接收方约定好指定商品，接收方即可浏览相应的商品评论，通过调用本文方法提取出秘密信息。这种方式隐藏了一对一的隐蔽通信行为，能够有效地阻止隐蔽通信行为被攻击者发现，从而避免发送方被追踪溯源。

为了更好地阐述本文的核心算法，表 2 给出了算法使用的符号说明及其定义。

表 2　　符号说明及其定义

符号	定义
K	通信双方选定的商品
O	商品列表集合
Size	商品列表的长度
P	选定的商品属性
Q	选定的情感组合
Extraction_model	无监督提取函数
dict	数据字典
Secret_information	待嵌入的秘密信息
GetIndex	获取秘密信息索引下标函数
I	秘密信息的索引下标集合
L	索引下标集合的长度
C	商品评论列表集合
C'	含密商品评论

2.2　信息嵌入与提取算法

信息嵌入与提取算法具体可分为如下步骤，分别是 1) 对秘密信息进行分词操作，得到分词索引下标；2) 运行无监督抽取模型，得到情感表达组合并将其映射到语义空间；3) 根据秘密信息选择特定语义空间下的情感表达组合生成含密评论。具体描述如下。

步骤 1　消息发送方输入秘密信息后，需要对秘密信息进行分词操作，得到其索引下标，这里详细给出调用 DSIE 算法计算词语“商品”左右信息熵的实例。

假设句子为“这个商品的商品盒比未升级之前商品的商品盒还要丑”，统计得到“商品”的左邻字集合和右邻字集合，如表 3 所示。

表 3　　“商品”的左右邻字集合

左邻字集合	右邻字集合
个	的
的	盒
前	的
的	盒

根据式(2)，其左邻字的信息熵为

$$-\left(\frac{1}{2}\right)\log\left(\frac{1}{2}\right)-\left(\frac{1}{4}\right)\log\left(\frac{1}{4}\right)-\left(\frac{1}{4}\right)\log\left(\frac{1}{4}\right)$$

右邻字的信息熵则为

$$-\left(\frac{1}{2}\right)\log\left(\frac{1}{2}\right)-\left(\frac{1}{2}\right)\log\left(\frac{1}{2}\right)$$

步骤 2　根据消息发送方与接收方事先约定好的指定商品，调用无监督抽取模型，随机抽取概率最高的对象候选项以及情感表达组合集合。在计算情感表达组合的分数时，每一次迭代完成后都要对结果进行标准化处理，直至式(3)中 $\boldsymbol{F}_i$ 和 $\boldsymbol{F}_{i+1}$ 近似收敛，此时可得到每一个情感表达组合的分数以及连接词的分数，标准化处理的计算式为

$$\boldsymbol{M}'_j=\frac{\boldsymbol{M}_j}{\sum_{j=1}^{D}\boldsymbol{M}_j}\times D,\boldsymbol{M}\in R^{D\times 1} \tag{4}$$

其中，$\boldsymbol{M}$ 是需要标准化处理的矩阵；D 是 $\boldsymbol{M}$ 矩阵的维度，即 $\boldsymbol{M}$ 矩阵标准化后得到的总分数。根据式(4)使每一维分数能够重新分配总分数 D。

步骤 3　该步骤是信息嵌入算法中最关键的一步。首先，本文需要对秘密信息转化得到的索引下标进行求余操作。然后，根据步骤 2 中得到的对象候选项列表，选择余数列表下标对应的对象。最后，在指定对象抽取的情感表达组合中，将索引下标与 Size 做除操作，选择对应的商品评论组合，重复上述步骤，直到处理完所有秘密信息。

为了更直观地了解隐藏过程，本文设计了如算法 1 所示的信息嵌入算法和如算法 2 所示的信息提取算法。

算法 1　信息嵌入算法

输 入　商品信息 K，待嵌入秘密信息

Secret_information = (m_1, m_2, m_3,⋯, m_L)

输出 含密商品评论 C'

1) GetIndex(Secret_information) → I = (I_1, I_2, I_3,⋯, I_L)//对 Secret_information 进行分词操作后得到索引下标集合 I

2) Extraction_model(K)→ 商品列表集合 O = { 'O_1','O_2','O_3', ⋯}，其 Size 等于 Num；商品评论列表 C ={"O_1": ['C_{11}','C_{12}','C_{13}',⋯]，"O_2": ['C_{21}','C_{22}','C_{23}',⋯]}//运行无监督抽取模型，得到对象列表集合 O 以及商品评论列表集合 C

3) for each index containing I

4) select O_P by count(I_i, Num) == P//对索引下标 I_i (i∈[1, L])进行取余操作，选中对象属性 O_P

5) select C_{PQ} by count (I_i, Num) == Q//对索引下标 I_i (i∈[1, L]) 进行除操作，选中情感组合 C_{PQ}

6) end for

7) if len(combination) > 3//判断当前观点表达数是否超过 3 个

8) continue 步骤 2)

9) end if

10) output (C'= { C_{PQ_1}， C_{PQ_2}， C_{PQ_3} ,⋯})

算法 2 信息提取算法

输入 商品信息 K，含密商品评论 C'

输出 秘密信息 Secret_information

1) Extraction_model(K)→ 商品列表集合 O = { 'O_1', 'O_2', 'O_3', ⋯}，其 Size 等于 Num；商品评论列表集合 C ={'O_1": ['C_{11}', 'C_{12}', 'C_{13}', ⋯]，"O_2": ['C_{21}','C_{22}','C_{23}',⋯]}//运行无监督抽取模型，得到商品列表集合 O 以及含密商品评论 C'

2) for each C_{PQ_i} containing C'

3) get I_i by count(P, Q)//通过 P 和 Q 得到秘密信息索引下标

4) end for

5) by dict(I) → Secret_information//通过数据字典查找索引下标对应的词语提取秘密信息

6) output(Secret_information)

2.3 适用场景

图 4 以实例展示了 Alice 和 Bob 之间的秘密通信流程，Alice 想要通过某电商平台的评论给 Bob 传递秘密消息，如“星期天在咖啡厅见面”。Alice 与 Bob 需要事先约定选择某一种商品，如“氨基酸洗面奶”；Alice 运行信息嵌入算法后生成一段针对该商品的含密评论“我觉得商品洗完脸后非常舒服，不干燥，棒棒棒”，并将该评论发表在某电商平台上。Bob 通过浏览事先约定的该商品的评论，运行信息提取算法后可提取其秘密信息，从而实现秘密信息的隐蔽传输。

3 实验及结果分析

在生成式文本隐写领域，评价指标主要包括 3 个方面：感知隐蔽性[18]（隐写文本达到语义完整、语法正确且足够自然的要求）、统计隐蔽性[19]（隐写文本符合自然文本的统计分布特征）以及认知隐蔽性[20]（隐写文本贴合相关主题或者满足上下文约束条件）。本文方法并没有操作文字符号，所生成的含密评论与自然评论的统计分布特征理论上是一

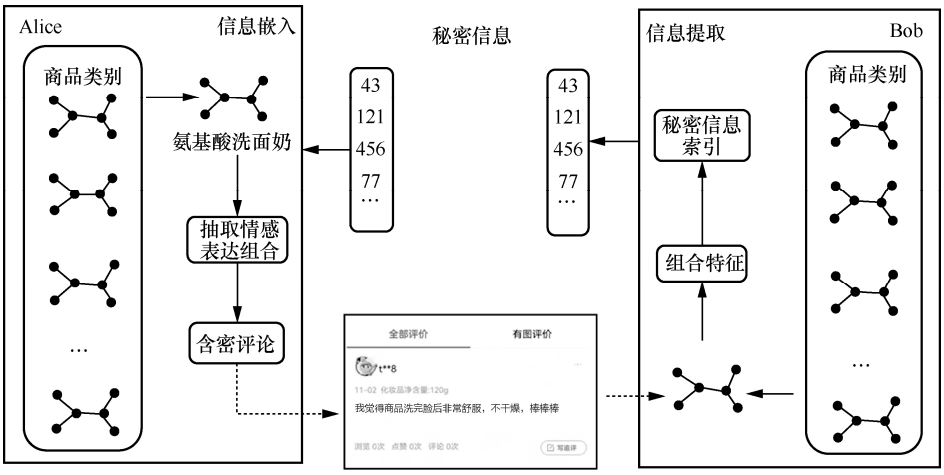


图 4　Alice 与 Bob 之间的秘密通信流程

致的，因此本文聚焦感知隐蔽性以及认知隐蔽性这 2 个方面进行实验验证。

1) 感知隐蔽性

本文采用生成式文本隐写常用的感知隐蔽性评价指标困惑度（PPL, perplexity）衡量生成含密评论的文本质量，其具体数学定义如式(5)所示[19]。考虑到评论文本具有网络化、口语化、多样化等语言特点，PPL 并非越小越好。

$$\mathrm{PPL}=2^{-\frac{1}{N}\sum_{i=1}^{N}\log P_i(v_{i_1},v_{i_2},\cdots,v_{i_n})} \tag{5}$$

其中，N 表示生成含密评论的数量，单词序列 $\{v_{i_1}, v_{i_2},\cdots,v_{i_n}\}$ 表示第 i 个句子，$\log P_i(\cdot)$ 表示第 i 个句子的概率。

2) 认知隐蔽性

在评价认知隐蔽性时，不仅需要判断含密评论文本与选定的商品具有相关性，同时还需要判断评论的情感极性是否符合特定情感表达。基于此，本文采用的相关评价指标包括 t-SNE 分布以及改进后的 BLEU（bilingual evaluation understudy），其核心思想是使用不同长度的候选短语（n 元组）将生成的含密评论与原始商品评论相匹配，从而衡量两者之间的相关程度。一般来说，BLEU 值越高，表明生成的评论文本与原始评论文本的相似程度越高，效果也就越好，其计算式为

$$\mathrm{BLUE}=\frac{\sum_i\sum_j \mathrm{Count}_{\mathrm{clip}}(n\text{-gram})}{\sum_i\sum_j \mathrm{Count}(n\text{-gram})} \tag{6}$$

其中，$\mathrm{Count}(n\text{-gram})$ 表示生成的评论文本中 n 元模型出现的次数，$\mathrm{Count}_{\mathrm{clip}}(n\text{-gram})$ 表示自然评论中 n 元模型出现的次数。

针对生成的评论比自然评论短导致 BLEU 分数相对较高从而影响实验结果的问题，本文引入了惩罚机制。当模型生成的评论低于自然评论时乘以 BP 惩罚因子（一个小于 1 的比例），以便更加客观地反映生成含密评论的质量。

$$\mathrm{BP}=\begin{cases}1, l_{\mathrm{c}}>l_{\mathrm{s}}\\ \exp\left(1-\dfrac{l_{\mathrm{s}}}{l_{\mathrm{c}}}\right), l_{\mathrm{c}}\leqslant l_{\mathrm{s}}\end{cases} \tag{7}$$

其中，l_{c} 是自然评论中句子的长度，l_{s} 是模型生成的评论中最接近自然评论的长度。

3.1 实验设置

本文以电商平台广泛存在的商品评论作为应用场景，需要大量的商品评论数据集用于训练模型。但目前还未有机构发布商品评论的公开数据集，所以本文通过爬虫框架从网易考拉、淘宝以及京东获取了约 9.3 万句用户商品评论作为实验的数据集。数据集的商品种类为 25 种（其中包含护肤品、计算机以及图书等），每条商品评论的平均长度约为 48 B。实验中，首先将原始数据集 raw-comment 进行繁体转简体，清洗无意义字符，分词、词性标注等操作，得到了 clean-comment 数据集，并据此构建了索引字典。数据集概况如表 4 所示。

表 4 数据集概况

数据集	商品评论句子数/句	字符数/B	平均每句字符数/B	商品类别/种
网易考拉	55 456	2 661 954	48.01	25
淘宝	14 232	687 690	48.32	25
京东	23 911	1 182 160	49.44	25

本文基于 NVIDIA geforce RTX 2080 Ti 显卡搭建了 GPT-2-python3.6.10-Jieba0.42.1 环境训练模型。在抽取情感表达组合时，采用滑动窗口的方式，根据词性获取选定商品的特征属性，再根据不同属性出现的概率分配权重并进行排序，在排序的过程中采用余弦算法对相似的观点表达进行合并，输出最终的情感表达组合二元对。

通过对自然评论的数据分析，本文发现自然评论含有的情感表达组合一般不会超过 3 个，为了保证商品评论的可读性，在每次输出含密评论前，模型需要判断当前观点表达是否合法，即含有的情感表达组合是否超过 3 个，若模型判定不合法，将重新生成含密评论。

3.2 实验结果及分析

3.2.1 感知隐蔽性

PPL 是文本生成领域评价生成文本质量的常用指标。表 5 比较了当前主流的几种基于语言模型和 Conproc 框架的生成式文本隐写方法[21-23]（AC 和 HC 分别表示基于算术编码和基于霍夫编码的条件概率编码方法）在不同嵌入容量（单位为 bit/word）下生成的 PPL 的实验结果。由表 5 可知，随着嵌入容量的增加，基于语言模型和 Conproc 的嵌入模型生成文本的质量逐渐变差。而本文方法由于并没有显式地操作文字符号，因此嵌入容量并不会影响生成商品评论文本的质量。

表 5　同类隐写方法在不同嵌入容量下生成的 PPL 对比

方法	嵌入容量/(bit·word^{-1})	PPL
LSTM	1.000	30.665
	2.000	10.027
	3.000	74.543
VAE-Stega (LSTM-LSTM) (HC)	1.000	45.115
	1.863	49.511
	2.577	59.532
VAE-Stega (BERT-LSTM) (AC)	1.000	30.266
	1.866	36.349
	2.596	40.832
VAE-Stega (BERT-LSTM)(HC)	1.000	30.266
	1.866	36.349
	2.596	40.832
RNN-Stega (HC)	1.000	20.915
	1.845	24.839
	2.565	29.187
本文方法	1.000	**20.341**
	2.000	20.933
	3.000	20.431

针对商品不同特征生成的含密评论的 PPL 值对比如表 6 所示，其中，t 表示针对某一商品抽取出的特征数量。由实验结果可知，本文方法生成的商品评论的困惑度最高可达 14.432，低于文献[24]采用的 CTRL 模型所生成的文本困惑度。

表 6　针对商品不同特征生成的含密评论的 PPL 值对比

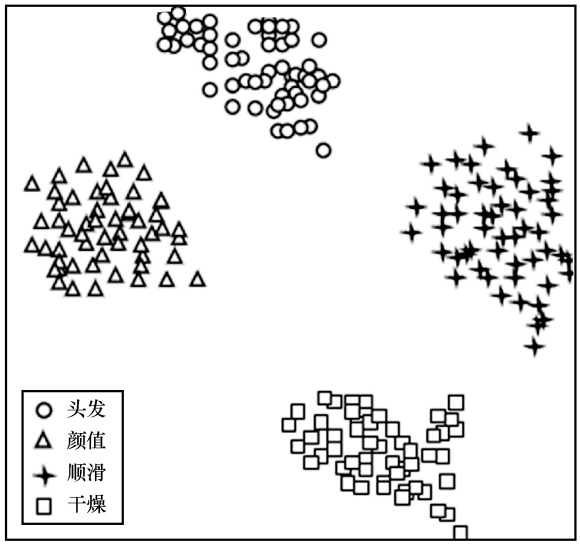
t	本文方法	CTRL
4	10.536～10.900	16.003～16.541
8	11.453～11.543	18.661～18.983
10	13.334～13.454	18.901～19.345
16	14.342～14.432	19.494～19.834

事实上，在电商评论中，如“痘痘肌”“炒鸡棒”等词是符合大众思维且容易理解的网络词汇，但在预训练模型中会认定这样的词语是不符合语法规则的，这会极大地影响实验的准确性，因此在实际使用中，需要用到一些常用的网络词汇去预训练模型，尽可能地确保生成文本的质量和隐蔽性。

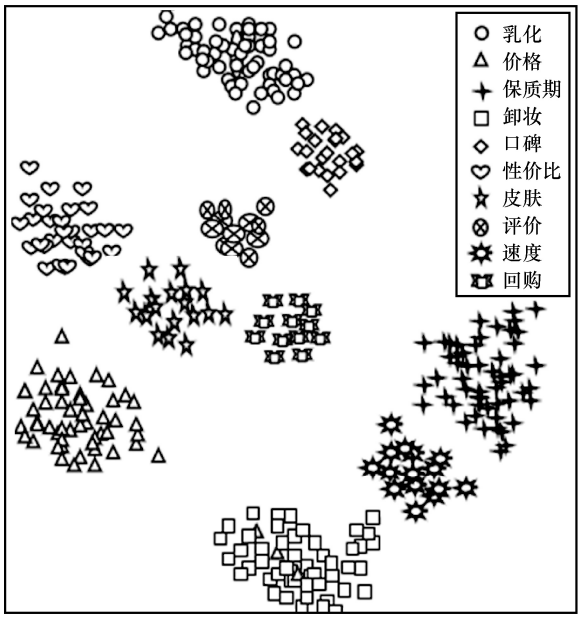
3.2.2　认知隐蔽性

为了评价模型生成的含密商品评论是否满足相关主题的约束，以商品“洗发水”和“卸妆油”为例进行分析。模型抽取商品“洗发水”的特征列表为['头发', '颜值', '顺滑', '干燥']，抽取商品“卸妆油”的特征列表为['乳化', '价格', '保质期', '卸妆', '口碑', '性价比', '皮肤', '评价', '速度', '回购']。

本文以特征列表的下标索引代表特征向量，并通过 t-SNE[25]简化为二维向量，然后将其映射到语义特征空间。图 5 显示了在隐写方案中针对洗发水（图 5(a)）和卸妆油（图 5(b)）选择不同特征（以不同图形区分）生成的含密评论在语义空间中的分布。从图 5 中可以看出，选择相同特征生成的含密评论空间分布集中，这进一步说明了生成的商品评论与主题具有相关性。



(a) 洗发水



(b) 卸妆油

图 5　含密评论在语义空间下的分布

借鉴机器翻译领域评价文本质量的常用指标 BLEU[26]，本文将该指标的结果表示为含密评论与自然评论的相似程度。针对不同对象生成含密评论的 BLEU 对比如表 7 所示。由表 7 可知，当模型为 1-gram 时，其含密评论的 BLEU 均值最高。

表 7 针对不同对象生成含密评论的 BLEU 对比

t	1-gram	2-gram	3-gram	4-gram
4	0.834	0.764	0.642	0.554
10	0.778	0.704	0.597	0.441

考虑到 BLEU 指标的评价方法主要是计算 n 元模型（n-gram）的匹配数量，而含密评论经常会出现不同特征词，导致评分较低。因此该指标对评估含密评论的质量有一定的局限性。本文抽取部分隐写结果进行了人工分析，结果显示生成的含密评论与自然评论的语义及语句流畅性相差较小。

3.2.3 嵌入效率

嵌入效率（ER, embedding rate）是评估隐写算法性能的一个重要指标，它反映了隐藏算法嵌入秘密信息的效率。本文将嵌入效率定义为嵌入在评论中的秘密信息二进制位数（Binary）与评论本身中的单词总数（Total）之间的比率，其计算式为

$$\mathrm{ER}=\frac{\mathrm{Binary}}{\mathrm{Total}} \tag{8}$$

图 6 显示了不同方法针对相同商品嵌入不同长度秘密信息的嵌入效率。图 7 显示了本文方法针对不同商品嵌入不同长度秘密信息的嵌入效率，具体选择了防晒霜、口红以及喷雾这 3 种商品。从图 7 中可以看出，语义空间下本文方法的嵌入效率都不高。因此在下一步的研究工作中，需要将语义空间下的隐写方法与传统的符号隐写方法相结合以提高隐藏容量。

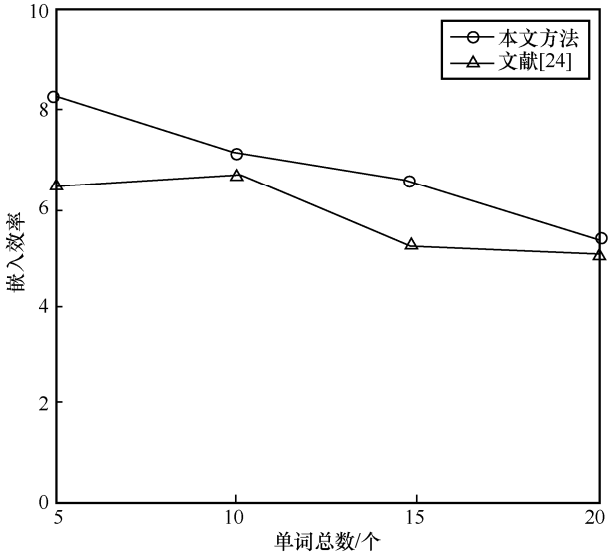


图 6 不同方法针对相同商品嵌入不同长度秘密信息的嵌入效率

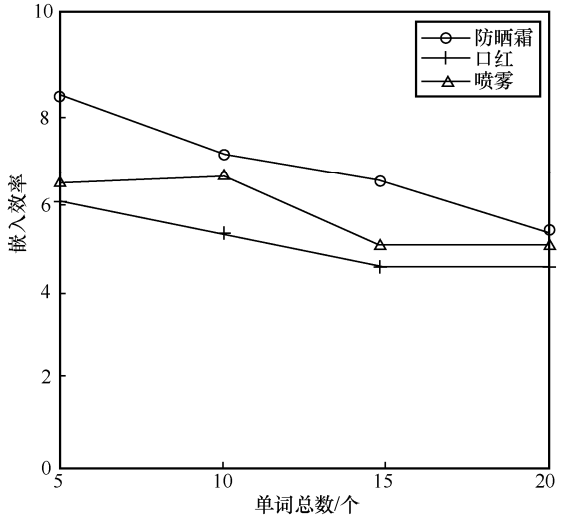


图 7 本文方法针对不同商品嵌入不同长度秘密信息的嵌入效率

3.2.4 安全性分析

安全性是隐写算法重要的评价指标之一。是否能够抵抗隐写分析的检测是评价隐写算法是否安全的重要体现。不少学者提出了一些隐写分析算法来检测隐写文本的安全性。例如，文献[14]提出了一种基于 BERT 的文本隐写分析方法，实验结果表明该方法相比其他检测方法具有更高的检测效率，因此本文采用该方法进行隐写分析的检测实验。在实验中，本文将模型学习率设置为 0.001，权重衰减默认为 0。结果显示，该方法对本文所生成的含密评论检测正确率只有 0.52，进一步验证了本文方法的有效性。

当秘密信息长度较长时，为了进一步保证本文方法的安全性，发送方可以对秘密信息进行有效切分，将秘密信息嵌入多个指定商品的评论中实现协同隐蔽通信。这样，即使隐蔽通信方式被破解，网络攻击者也无法提取出完整的秘密信息，并且部分秘密信息引导生成的短评论文本也更加符合新媒体平台上用户产生的评论文本特性。

当接收方浏览电商平台的含密商品评论时，本文利用通信双方在电商平台构建好友关系后，便可即时查看对方的商品评论的特点，因此可以避免“虚警”问题的产生。

3.2.5 jieba_DSIE 分词方法的讨论

本文利用词性信息获取情感表达组合候选集合，因此分词效果越好，得到的情感表达组合越准确。本文主要采用如下 4 种评价指标衡量商品评论数据集的分词效果：准确率 P_{acc}、未登录词召回率 R_{oov}、召回率 R_{cal} 和 F_1-评测值。4 种指标的定义式分别为

$$P_{\text{acc}}=\frac{\text{正确切分出的词的数目}}{\text{切分出的词的总数}}$$

$$R_{\text{oov}}=\frac{\text{正确切分出的未登录词的数目}}{\text{标准答案中未登录词的数目}}$$

$$R_{\text{cal}}=\frac{\text{正确切分出的词的数目}}{\text{应切分出的词的数目}}$$

$$F_1=\frac{2P_{\text{acc}}R_{\text{cal}}}{P_{\text{acc}}+R_{\text{cal}}}\times 100\% \tag{9}$$

jieba 作为一种强大的分词工具，提供了精确模式、全模式以及搜索引擎模式 3 种分词模式。本文在自己的数据集上对比了这 3 种分词模式，发现这 3 种分词模式在未登录词召回率上都无法令人满意，因此本文采用了 jieba_DSIE 的分词方法。由图 8 可知，与仅使用 jieba 相比，jieba_DSIE 方法在未登录词召回率上增加了 30%，提高了分词模型的效果。究其原因，jieba 分词属于传统的词库分词，新词识别能力受制于词库容量的大小，利用 DSIE 方法得到新词并载入词库后可以帮助其获得更好的性能。

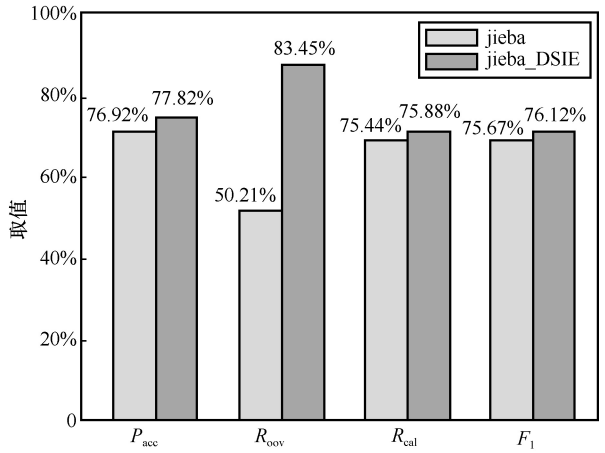


图 8　2 种方法的性能对比

4　结束语

本文提出了一种在语义空间下基于情感表达的生成式文本隐写方法，可用于在各新媒体平台上实现隐蔽通信。通过利用无监督抽取模型抽取情感表达组合生成含密评论的同时嵌入秘密信息。在抽取情感表达组合的过程中，充分考虑实体之间和情感词之间的约束。本文以某电商平台的商品评论为例进行了实验，实验结果表明，本文方法具有更高的安全性以及适用性，但嵌入效率有待提高。在今后的工作中，一方面可尝试将基于语义空间的隐写方法与传统的基于符号空间的隐写方法相结合，进一步提高本文方法的嵌入效率，另一方面也可以设计更合理的协同隐蔽通信协议以提高隐写安全性。

参考文献：

[1] PETITCOLAS F A P, ANDERSON R J, KUHN M G. Information hiding-a survey[J]. Proceedings of the IEEE, 1999, 87(7): 1062-1078.

[2] 吴国华, 龚礼春, 袁理锋, 等. 中文文本信息隐藏研究进展[J]. 通信学报, 2019, 40(9): 145-156.
WU G H, GONG L C, YUAN L F, et al. Review of information hiding on Chinese text[J]. Journal on Communications, 2019, 40(9): 145-156.

[3] 张祯, 倪嘉铭, 姚晔, 等. 基于同义词扩展和标签传递机制的文本无载体信息隐藏方法[J]. 通信学报, 2021, 42(9): 173-183.
ZHANG Z, NI J M, YAO Y, et al. Text coverless information hiding method based on synonyms expansion and label delivery mechanism[J]. Journal on Communications, 2021, 42(9): 173-183.

[4] WANG C L, LIU Y L, TONG Y J, et al. GAN-GLS: generative lyric steganography based on generative adversarial networks[J]. Computers, Materials & Continua, 2021, 69(1): 1375-1390.

[5] WAYNER P. Mimic functions[J]. Cryptologia, 1992, 16(3): 193-214.

[6] CHAPMAN M, DAVIDA G I, RENNHARD M. A practical and effective approach to large-scale automated linguistic steganography[C]//Proceedings of the 4th International Conference on Information Security. New York: ACM Press, 2001: 156-165.

[7] DESOKY A. Nostega: a novel noiseless steganography paradigm[J]. Journal of Digital Forensic Practice, 2008, 2(3): 132-139.

[8] LUO Y B, HUANG Y F, LI F F, et al. Text steganography based on ci-poetry generation using Markov chain model[J]. KSII Transactions on Internet and Information Systems, 2016, 10(9): 4568-4584.

[9] YI X Y, LI R Y, SUN M S. Generating Chinese classical poems with RNN encoder-decoder[C]//International Symposium on Natural Language Processing Based on Naturally Annotated Big Data, China National Conference on Chinese Computational Linguistics. Berlin: Springer, 2017: 211-223.

[10] YANG Z L, GUO X Q, CHEN Z M, et al. RNN-Stega: linguistic steganography based on recurrent neural networks[J]. IEEE Transactions on Information Forensics and Security, 2019, 14(5): 1280-1295.

[11] ZIEGLER Z, DENG Y T, RUSH A. Neural linguistic steganography[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Stroudsburg: Association for Computational Linguistics, 2019: 1210-1215.

[12] XIANG L Y, YANG S H, LIU Y H, et al. Novel linguistic steganography based on character-level text generation[J]. Mathematics, 2020, 8(9): 1558.

[13] NAKAJIMA T V, KER A D. The syndrome-trellis sampler for generative steganography[C]//Proceedings of 2020 IEEE International Workshop on Information Forensics and Security (WIFS). Piscataway: IEEE Press, 2021: 1-6.

[14] ZHOU X J, PENG W L, YANG B Y, et al. Linguistic steganography based on adaptive probability distribution[J]. IEEE Transactions on Dependable and Secure Computing, 2022, 19(5): 2982-2997.

[15] 石凤贵. 基于 jieba 中文分词的中文文本语料预处理模块实现[J]. 电脑知识与技术, 2020, 16(14): 248-251, 257.
SHI F G. Realization of Chinese text corpus preprocessing module

based on jieba Chinese word segmentation[J]. Computer Knowledge and Technology, 2020, 16(14): 248-251, 257.

[16] 卢奇, 陈文亮. 大规模中文实体情感知识的自动获取[J]. 中文信息学报, 2018, 32(8): 32-41.
LU Q, CHEN W L. Automatically building a large scale dictionary of Chinese entity sentiment expressions[J]. Journal of Chinese Information Processing, 2018, 32(8): 32-41.

[17] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[J]. arXiv Preprint, arXiv: 1301.3781, 2013.

[18] YANG Z L, GONG B T, LI Y M, et al. Graph-Stega: semantic controllable steganographic text generation guided by knowledge graph[J]. arXiv Preprint, arXiv: 2006.08339, 2020.

[19] ZHANG S Y, YANG Z L, YANG J S, et al. Provably secure generative linguistic steganography[C]//Proceedings of Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Stroudsburg: Association for Computational Linguistics, 2021: 3046-3055.

[20] YANG Z L, ZHANG S Y, HU Y T, et al. VAE-Stega: linguistic steganography based on variational auto-encoder[J]. IEEE Transactions on Information Forensics and Security, 2021, 16: 880-895.

[21] JELINEK F, MERCER R L, BAHL L R, et al. Perplexity—a measure of the difficulty of speech recognition tasks[J]. The Journal of the Acoustical Society of America, 1977, 62(S1): S63.

[22] FANG T N, JAGGI M, ARGYRAKI K. Generating steganographic text with LSTMs[C]//Proceedings of ACL 2017, Student Research Workshop. Stroudsburg: Association for Computational Linguistics, 2017: 100-106.

[23] MAAS A L, DALY R E, PHAM P T, et al. Learning word vectors for sentiment analysis[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. New York: ACM Press, 2011: 142-150.

[24] ZHANG S Y, YANG Z L, YANG J S, et al. Linguistic steganography: from symbolic space to semantic space[J]. IEEE Signal Processing Letters, 2021, 28: 11-15.

[25] ZHOU H Y, WANG F, TAO P. T-distributed stochastic neighbor embedding method with the least information loss for macromolecular simulations[J]. Journal of Chemical Theory and Computation, 2018, 14(11): 5499-5510.

[26] PAPINENI K, ROUKOS S, WARD T, et al. BLEU: a method for automatic evaluation of machine translation[C]//Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. New York: ACM Press, 2002: 311-318.

[作者简介]

刘玉玲（1980– ），女，湖南宁乡人，博士，湖南大学副教授、博士生导师，主要研究方向为多媒体内容安全、保密技术、自然语言处理等。

王翠林（1999– ），女，土家族，湖南湘西人，湖南大学硕士生，主要研究方向为文本内容安全、自然语言处理等。

付章杰（1983– ），男，河南南阳人，博士，南京信息工程大学教授、博士生导师，主要研究方向为人工智能安全、区块链安全、数字取证等。