面向中文用户评论的自动化众包攻击方法

王丽娜[1,2]，郭晓东[1,2]，汪润[1,2]
（1. 武汉大学空天信息安全与可信计算教育部重点实验室，湖北 武汉 430072；
2. 武汉大学国家网络安全学院，湖北 武汉 430072）

摘 要：面向文本的自动化众包攻击具有攻击成本低、隐蔽性强等特点，这种攻击可以自动生成大量虚假评论，影响用户评论社区的健康发展。近些年来，有学者研究面向英文评论社区的文本自动化众包攻击，但是鲜有针对中文评论社区的自动化众包攻击的研究，针对这一不足，提出了基于汉字嵌入 LSTM 模型的中文文本自动化生成攻击方法。通过训练由汉字嵌入网络、LSTM 网络和 Softmax 全连接网络组成的多层网络模型，并引入温度参数 *T* 构建攻击模型。实验中，从淘宝网的在线用户评论中抓取了超过 5 万条真实的用户评论数据，验证所提攻击方法的有效性。实验结果表明，生成的虚假评论可以有效地欺骗基于语言学分析的分类检测方法和基本文本拷贝检测等方法, 并且通过大量的人工评估实验发现所生成的文本具有真实性强、类型广等特点。

关键词：用户评论社区；自动化众包攻击；汉字嵌入网络；长短期记忆网络
中图分类号：TP309.1
文献标识码：A
doi: 10.11959/j.issn.1000−436x.2019149

Automated crowdturfing attack in Chinese user reviews

WANG Li’na[1,2], GUO Xiaodong[1,2], WANG Run[1,2]
1. Key Laboratory of Aerospace Information Security and Trusted Computing Ministry of Education, Wuhan University, Wuhan 430072, China
2. School of Cyber Science and Engineering, Wuhan University, Wuhan 430072, China

Abstract: The text-oriented automated crowdturfing attack has a series of features such as low attack cost and strong concealment. This kind of attack can automatically generate a large number of fake reviews, with harmful effect on the healthy development of the user review community. In recent years, researchers have found that text-oriented crowdturfing attacks for the English review community, but there was few research work on automated crowdsourcing attacks in the Chinese review community. A Chinese character embedding LSTM model was proposed to automatically generate Chinese reviews with the aim of antomated crowdturfing attacks, which model trained by a combination with Chinese character embedding network, LSTM network and softmax dense network, and a temperature parameter *T* was designed to construct the attack model. In the experiment, more than 50 000 real user reviews were crawled from Taobao's online review platform to verify the effectiveness of the attack method. Experimental results show that the generated fake reviews can effectively fool linguistics-based classification detection approach and texts plagiarism detection approach. Besides, the massive manually evaluation experiments also demonstrate that the generated reviews with the proposed attack approach perform well in reality and diversity.

Key words: user review community, automated crowdturfing attack, Chinese character embedding network, LSTM

收稿日期：2019−03−22；**修回日期**：2019−05−31
基金项目：国家自然科学基金资助项目（No.61876134，No.U183610015）；中央高校基本科研业务费专项基金资助项目(No.2042018kf1028)
Foundation Items: The National Natural Science Foundation of China (No.61876134, No.U183610015), The Central University Basic Business Expenses Special Funding for Scientific Research Project (No.2042018kf1028)

1 引言

在自媒体时代，用户评论社区中的每一个网络用户都可以自由地发布自己的观点，其中包括用户对产品的质量或服务的评论。例如美国最大的用户评论平台 Yelp，平均每天有三百多万的用户访问，拥有超过 1.71 亿条评论。在国内，随着淘宝、京东、大众点评等电子商务平台的发展，越来越多的消费者选择通过商品评论社区来分享自己的购物体验、评价商品质量和服务。用户在发布评论的同时也倾向于依赖其他用户的评论做出消费决策。文献[1-2]指出用户评论的倾向性（积极或消极）是影响商品口碑的一项重要因素，此外，文献[3-4]的调查研究表明商品的口碑会从多个方面影响消费者的消费决策，比如消极评论的增加会导致餐馆减少 5%~9%的收入，另一项研究[5]表明 80%的美国用户承认会参考用户评论来选择购买的产品和服务。面对自由开放的用户评论社区，针对某一件商品或某一款服务，发布大量相同倾向性的用户评论，比如大量好评或大量差评，从而提升或降低商品或服务的口碑，能产生巨大的商业利益，所以恶意组织或个人选择多种方式生成大量的虚假评论来攻击竞争对手或提升自身产品的影响力，其中，面向文本的众包攻击是一种被广泛应用的虚假评论生成方法。面向文本的众包攻击，简称为众包攻击(crowdturfing attack)，即采用众包的方式雇佣数个网络写手，捏造大量不符合事实的评论[6]，诋毁竞争对手或提高自己的口碑。恶意的众包攻击会对公平竞争产生非常负面的影响，不仅对商家造成伤害，也会破坏用户评论社区的健康环境，导致用户评论的可信度下降，消费者将失去主要的获取网络商品信息的渠道[7-9]。

传统基于人工的众包攻击方法具有以下不足。1) 攻击成本高、效率低，基于人工的众包攻击方法需要支付大量的酬劳，并且效率低，难以形成大规模的攻击行动；2) 攻击效果差，在面对大规模的真实攻击场景中，需要在限定的时间内生成大量的虚假评论，导致了人工生成的评论质量低，效果差。因此针对传统众包攻击中的诸多不足，一种新型的攻击方式逐渐被关注，即自动化众包攻击(automated crowdturfing attack)[10]，两者的比较如图 1 所示。在人工众包攻击中，攻击者将任务分发给网络写手，网络写手撰写大量的虚假评论。而自动化众包攻击方法使用计算机程序取代网络写手，由程序自动撰写大量虚假评论。自动化众包攻击利用机器学习模型在特征表示方面的巨大优势，自动地生成大量高质量、逼真的用户评论。这种攻击方式可以有效地满足大规模、真实场景的攻击要求，能够快速地生成大量的高质量用户评论。因此，自动化众包攻击成为目前用户评论社区攻击的重要手段。

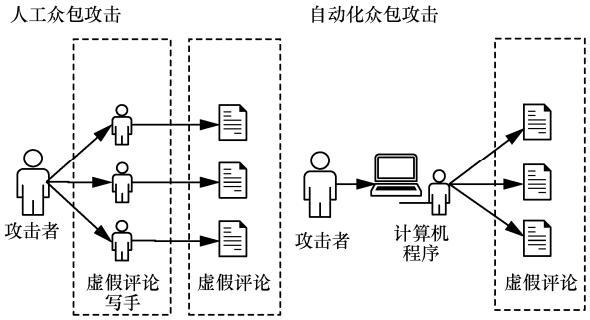

图 1 传统众包攻击与自动化众包攻击的比较

近些年来，针对文本自动化众包攻击的研究大多针对英文文本。文献[10-18]提出了面向英文用户评论的虚假评论生成方法和自动化众包攻击方法。这类方法采用独热向量（该向量的索引位置为 1，其他位置为 0）表示英文字母，任意 2 个字母之间都是孤立的，无法表示语义层面上的相关信息，这使面向英文文本的方法并不适用于面向中文用户评论的自动化众包攻击。英文中的每个单词都由空格或标点隔开，不同长度的单词也都由 26 个基础字母组成，这种简单的语言结构更利于深度网络的学习；而中文句子中的词汇没有明显的分界，且组成每个词汇的汉字数量庞大，汉字与汉字之间本身又具有一定的语义联系，使中文句子的结构比英文句子的结构复杂，这给中文的处理造成了困难。

针对上述问题，本文利用嵌入网络学习汉字间上下文语义关系的特性，解决英文方法中独热向量表示方法不能捕捉汉字之间语义联系的问题，提出了一种面向中文文本的自动化众包攻击方法，流程如图 2 所示，具体如下。1) 利用汉字嵌入网络方法学习汉字间的语义距离，其中语义接近的汉字所表示的向量距离相近，语义较远的汉字则向量距离较远。同时汉字嵌入网络本质上是一种低维映射，可以有效减少后续网络体积，包括神经元个数和训练参数个数，加快训练速度。2) 基于长短期记忆网络(LSTM，long short-term memory)模型进行句法学习，并解决中文评论生成中的长依赖问题。例如中文评论“衣服收到，做工精细，没有色差…衣服

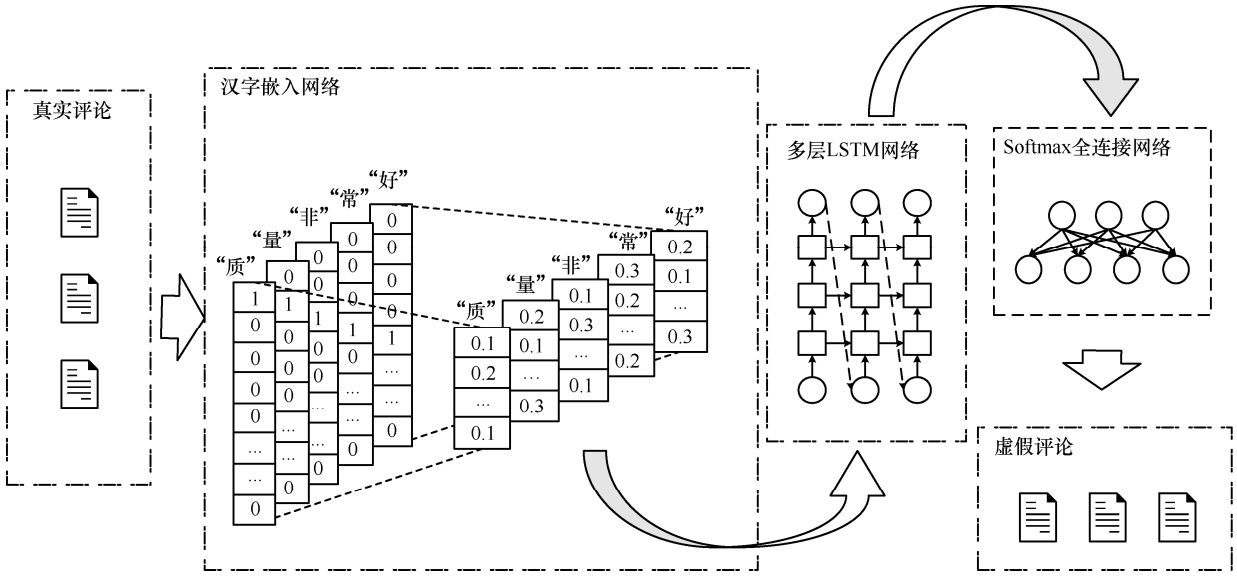

图 2　面向中文文本的自动化众包攻击流程

上身效果特别好”中，2 个汉字的组合“衣服”相互长依赖，但与评论中间省略的汉字依赖性很小，拥有记忆长依赖特性的LSTM网络在该类问题上表现效果优异，通过多层 LSTM 网络可以进一步学习中文评论中汉字的排列规律。3)采用 Softmax 全连接网络来生成大量多类型中文评论，满足自动化众包攻击的需求。Softmax 全连接网络基于之前两层网络的计算生成一组汉字概率分布，Softmax 函数不同的分层对应不同汉字的生成概率，基于生成概率进行随机抽样可以得到多个可选择的高概率汉字，生成评论中每个位置上的汉字都有更多的选择，则可以组合成多种内容，从而获得大量多类型的生成评论，但是评论的内容质量会下降。所以本文进一步使用温度参数 T 来平衡生成评论的质量和多样性。参数 T 对分层的汉字生成概率进行缩放，T 越大，分层间隔的差异越小，挑选每个汉字的概率趋同，随机抽样会抽取更多的低概率汉字，使模型生成评论类型更多但质量下降；T 越小则结果相反。

为了验证所提文本攻击方法的有效性，本文通过以下 3 个角度进行验证。1)基于语言学的文本相似性分析。对评论进行语言学分析是鉴别评论内容虚假的主要手段，例如词袋分析、LIWC 分析等。这里使用基于语言学分析的分类器来对比生成评论和真实评论的相似性，并验证生成评论的真实性。2)基于文本拷贝检测的文本生成多样性评估。生成句型单一的虚假评论并不能满足自动化众包攻击的需求，本文使用拷贝检测工具来对比本文攻击方法和传统基于名词替换的攻击方法，并验证生成评论的多类型。3)基于人工的文本真实性测试。机器手段只能分析生成评论的句子构成和特征，句子的可读性仍然需要端到端的用户调查方式来评估。实验结果表明，本文提出的面向中文文本的自动化众包攻击方法，能够快速、高效地生成大量的高质量文本，可以有效地应用于大规模、真实的用户评论社区攻击场景中。本文的主要贡献如下。

1) 提出了一种基于汉字嵌入 LSTM 模型的自动化众包攻击方法，该方法面向中文用户评论社区，可以自动地生成大量逼真的虚假评论。

2) 提出了一种基于汉字的嵌入方法，该方法可以有效地学习汉字之间的语义关系，能够生成更加真实的中文虚假评论。

3) 提出了一种基于汉字的随机抽样方法，通过在抽样分层函数中引入温度参数 T 来控制每一个汉字的采样概率，可以有效地平衡本文攻击方法生成虚假评论的质量和数量，极大地提高了自动化众包攻击的效果。

2　相关工作

2.1　文本生成方法

传统的自动文本生成技术，比如 n-gram 模型和基于文本模板的模型[12-14]，在模仿真实文本时都有局限性，这些技术所生成的文本存在语法错误、词不达意等问题，很可能会被用户判断为虚假内容。随着统计模型和神经网络的发展和普及[15-16]，

基于LSTM[17]模型的文本生成技术逐渐引起了研究者的关注。Yu 等[18]的研究表明基于 LSTM 模型的文本生成技术可以生成更加连贯的文本。Auli 等[19]的研究工作表明，基于 LSTM 网络的语言生成模型已经成为一种结构良好且极具前景的文本生成方法。LSTM 网络模型被广泛用于各个领域，Kannan[20]将真实用户内容聚类，使用 LSTM 网络模型构建效果良好的自动电子邮件应答器；Karpathy 等[21]使用双向 LSTM 网络模型来生成图像的文本描述，并在诸多图片检索实验中取得了良好的效果；Serban 等[22]使用一种扩展和分层递归编码的 LSTM 模型来生成电影场景中的对话文本，能够高度还原电影对话的连贯性；Shang 等[23]通过搜集大量社交网络中的对话文本，使用 LSTM 模型实现了社交软件中的自动问答。虽然 LSTM 模型在文本生成领域已经取得了多个成果，但是这些研究旨在提高生成文本的质量。对于自动化众包攻击，不仅需要文本质量足以欺瞒用户，还要产生足够多的不同类型文本来改变评论区的褒贬比例，达到影响消费者的消费决策的目的，这是传统文本生成技术不能满足的。

2.2 英文众包攻击方法

针对传统众包攻击方法存在的成本高、规模小、效果低等不足，自动化众包攻击方法逐渐吸引了研究者的关注。近些年，研究者主要围绕面向英文用户评论的虚假评论生成和自动化众包攻击等展开了一系列的研究工作。Bartoli 等[11]在2016年提出了自动生成虚假评论的可能性，并使用神经网络模型实现生成方法，通过真实用户调查研究，验证了自动生成虚假评论的可行性。Yao 等[10]在 2017 年提出自动化众包攻击的方法，针对最大的英文评论平台 Yelp，研究者使用神经网络模型学习意大利餐厅评论，生成一套虚假评论模板，然后通过关键词替换的方式生成针对日本餐厅的虚假评论，结果表明这些虚假评论足够真实且可以欺骗母语为英语的用户。Yu 等[18]在 2018 年提出了一种针对社交网络里恶意竞争的自动化众包攻击方法，研究者以 Twitter 平台为例，将评论文本按情感倾向聚类成不同极性的评论集，使用神经网络有针对性地学习并生成对应极性的虚假评论，实验表明 85.2%的生成评论具有正确的语法并具有一定的语义。这些研究成果表明自动化众包攻击方法是一种十分有研究意义的新型攻击方式，会给各类评论系统带来巨大的威胁。虽然面向英文用户评论的自动化众包攻击方法已经取得了比较好的效果，但直接用于进行中文自动化众包攻击的效果并不好[24]。原因主要在于，面向英文的自动化众包攻击使用基于字符的循环神经网络(char-level RNN)模型，Graves[25]和 Lebret[26]等的研究表明，在英文自然语言生成的问题上，基于字符的循环神经网络模型比基于词汇的循环神经网络（word-level RNN）效果更好。但常用汉字比常用英文字母庞大得多，字符集数量的差距导致每个字符对应的字向量也成倍增长，进一步使攻击模型的参数、占用的内存和计算开销都成倍增长。同时由于英文和中文本身的区别，例如英文字母间没有语义关系，而汉字之间存在语义关系，英文词汇由空格天然分割，汉字词汇没有明显分割，所以上述面向英文用户评论表现良好的自动化众包攻击方法并不适用于中文用户评论。

3 攻击方法概述

3.1 问题描述

面向中文的自动化众包攻击方法使用计算机程序自动地生成大量中文虚假评论，这些评论混杂在真实评论中不容易被分辨，且能够在一定程度上影响用户的消费决策。本文只关注生所成评论的文本内容能否接近真实评论且不易被鉴别，而其他鉴别虚假评论的因素，如评论与发布者本身的关系、发布者的影响力、发布者的评论历史、评论的发布日期、发布者的 IP 地址等不在本文考虑范围内。众包攻击方法的假设前提如下。1)攻击者有获取大量评论语料的能力。知识丰富的攻击者可以从用户评论社区上获得足够多的真实评论语料来训练攻击模型。2) 攻击者可以利用比较容易获得的计算资源来训练攻击模型，比如个人电脑、办公电脑、小型服务器、租用的云计算服务器等进行训练。

本文攻击模型使用的数学模型如下：给定一个含有 n 条真实评论的训练语料集 $R=\{r_1, r_2, r_3,\cdots, r_{n-1}, r_n\}$，语料集中包含 v 个不同的汉字，组成汉字集 V。对于每条评论 r，按顺序将 t 位置上的汉字表示为 w_t。攻击模型通过学习 w_i 之间的语义距离和 r_i 对应的句法，并接收生成长度参数 s，生成高质量的虚假评论集 $R'=\{r_1', r_2', r_3', \cdots\}$，其中，$R'$包含所有汉字的数量即为参数 s，虚假评论 r_1'、r_2'、r_3'等由攻击模型输出的评论划分标记分开。则攻击达到

的效果为：R 和 R'在语言学分类上大致相同，即 Tag(r_i')=Tag(r_j)，真实评论集 R 中任意 2 条评论 r_i 和 r_j 不会高度相似，同时 R'中的任意评论 r_i 都不与 R 中任意评论 r_j 高度相似，即 sim(r_i',r_j) ≤ sim(r_i, r_j)。

3.2　方法描述

攻击方法的流程如图 3 所示。首先抓取真实评论获得原始语料，对原始语料预处理后得到训练语料，然后经计算依次得到汉字嵌入向量、中间隐向量、汉字概率分布。最后基于汉字概率分布随机抽样生成评论。

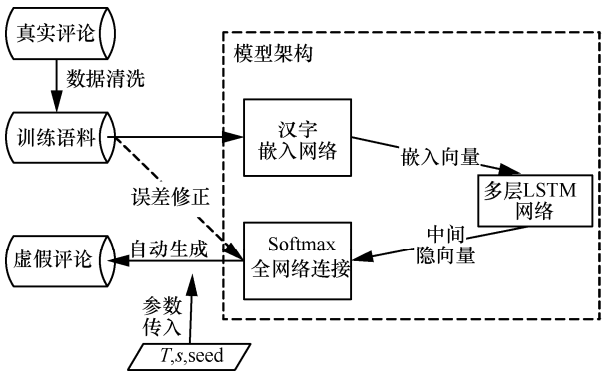

图 3　攻击方法流程

1) 原始语料预处理。从淘宝抓取的原始语料包含大量汉字文本内容之外的元素，所以数据清洗删除这类元素包括 emoji 表情、颜文字和外文字符。然后去除影响嵌入向量计算重复叠词和极少使用的汉字，将文本统一编码为 unicode 格式方便模型训练。为了在语料中划分每条评论，本文在每条评论的起始位置添加起始标记<SOR>(start of review)，评论末尾添加结束标记<EOR>(end of review)。

2) 汉字嵌入向量生成。在预处理之后，训练语料只包含与中文相关的语素。对比英文语素，中文汉字之间包含丰富的语义相关性，学习这种语义信息可以使攻击模型生成更加逼真的虚假评论。嵌入网络是一种有效且应用广泛的神经网络结构，将训练语料输入汉字嵌入网络，经过多轮迭代训练后得到关于每个汉字的嵌入向量(embedding vector)，其中，嵌入向量的距离关系表示汉字之间的语义关系。

3) 中间隐向量生成。嵌入向量仅包含语义距离关系，句子构成和排列方式是生成评论中更重要的信息。LSTM 网络具有序列敏感特性，被广泛应用于文本生成领域，本文中使用 LSTM 网络从训练语料中学习句子构成和排列方式，得到包含句法信息的中间隐向量，用于获得汉字概率分布。

4) 获得汉字概率分布。LSTM 网络学习到的写作方式只是用一些难以琢磨的中间隐向量(hidden vector)表示，所以需要一层全连接网络将隐向量所代表的概率分类到每一个汉字元素上。Softmax 函数是应用广泛的多分类函数，其简单且高效的表现是许多输出网络激活函数的首选。隐向量输入 Softmax 全连接网络，经过前向传播算法和 Softmax 分层函数计算后输出汉字概率分布，其中把概率最大的汉字作为预测汉字，对比预测汉字与训练汉字的误差，经反向传播算法计算并更新模型参数，训练完毕后 Softmax 网络得到一种隐向量与汉字概率的映射关系。

5) 随机抽样获得生成评论。汉字嵌入网络、LSTM 网络和 Softmax 网络整体构成一个攻击模型，给定一个起始字符、生成长度参数 s 和温度参数 T。起始字符作为评论生成的种子，生成长度参数 s 作为模型输出的迭代轮次，每一轮次输入之前已经生成的汉字，基于当前输入汉字序列计算输出汉字概率分布，然后对概率分布进行随机采样即可生成下一个汉字。温度参数 T 控制 Softmax 的抽样概率，将在第 4 节详细说明。这里将评论起始标记<SOR>送入攻击模型，记为 w_0，模型计算 $P(w_1|w_0)$，然后对所有汉字进行随机抽样得到 w_1，然后再将 w_0、w_1 输入攻击模型计算 $P(w_2|w_0,w_1)$，从而得到汉字 w_2，以此类推，最后得到生成评论集$\{w_0, w_1, w_2, \cdots, w_s\}$。

攻击模型经上述流程后生成的虚假评论如表 1 所示，其中每条评论都是从样本中随机抽取的，ID 表示各评论在表 1 中的顺序。温度参数 T 越大，生成的评论灵活性越高，攻击模型可以生成多种多样丰富的表达，但用词不当、语序错误的风险也会增加。

4　攻击方法的设计与实现

在本文的攻击方法实现中，首先利用汉字嵌入网络从预处理后的训练语料中学习汉字之间的语义关系，包括汉字共现统计和汉字嵌入权重计算；然后使用 LSTM 网络在嵌入向量基础上学习句子组成和顺序关系，生成中间隐向量，通过 Softmax 全连接层将隐向量映射到每一个汉字的概率，按汉字概率分布随机抽样获得生成评论。

表 1 不同温度参数 T 下生成的虚假评论样例

T	ID	文本评论内容
0.3	1	“衣服收到了，质量很好，穿上很舒服，也很好看。很喜欢，喜欢的小仙女可以下单哦。”
	2	“很厚实。穿上也很舒服，很喜欢，很喜欢，下次还会来的哦。”
0.5	3	“裤子很好看，穿在身上很暖和，客服推荐的尺码也很准，非常满意。”
	4	“质量很好，面料也舒服。穿上很舒服，颜色也很正，大小也合适，很喜欢。”
0.8	5	“物流快，几天就到手了，因为喜欢宽松的，穿上刚刚好。是我想要的款式，颜色也很正，好评。”
	6	“裤子真的太好看了，少女心满满了。大小合适，特漂亮，已推荐朋友们购买哦。”
1.0	7	“宝贝收到了炒鸡舒服，袖子是松紧的又给朋友买个小一号，没有色差价钱也不贵。好惊喜店家非常认真，物流贼快，下次再来。”
	8	“衣服收到，做工精细，没有色差。款式很美，小妹妹看正合适，价格也不高，穿起来显瘦，总之挺喜欢的，以后还会再来哒。”
1.3	9	“裤子版型款式很潮，胖妹子拍照，换洗小脚裤都很好看。款式新颖大方跟图片描述一模一样。耶哈哈太值了。”
	10	“太稀饭了上身效果好，很厚重，颜色也美观。冬天穿美感，颜色还显皮肤白。领裤都很好看，朋友再入下这种一个小店，合适，袖子太满意啦。”

4.1 汉字嵌入网络

本文采用向量空间模型(VSM) [27]，即嵌入网络来学习汉字之间的语义距离。嵌入网络分为字嵌入网络和词嵌入网络。通常词嵌入网络比字嵌入网络能更好地表示语义，因为现代汉语的基本语义元素是词汇，词汇中的汉字拆分后可能和原来的语义不同，但是词嵌入网络有一个中文分词的预处理过程，这个分词过程往往需要一个由语言学专家总结的中文词典[28]。而中文用户评论是一种非常新颖的文体且更新速度很快，里面存在很多的网络新词汇，例如表 1 例句中的“版型”(指服装轮廓)、“炒鸡”（超级）等，中文词表是不随网络新词汇的增加随时更新的，且更新耗费大。同时中文分词算法上存在一定的偏差，不能解决交集型歧义，例如“天真的你”会被划分成“天/真的/你”，不能解决组合型歧义，例如“网球拍卖完了”会被划分为“网球/拍卖/完/了”，考虑嵌入网络本身也存在着映射上的偏差，2 个计算过程带来的误差累加也是不可忽视的。而汉字直接嵌入不存在这些问题，统计语言模型直接从语料中学习成词的规则和句法，按照规则和句法自动造词，极大地保证了生成文本的灵活性，让生成的虚假评论看起来更加接近网络语言，更加真实。基于以上特点，本文使用汉字嵌入网络来构建自动化众包攻击模型。汉字嵌入网络首先计算窗口内 2 个汉字的联合概率分布，例如评论“质量非常好”，如图 4 所示。

给定一个大小为 2 的窗口，通过滑动窗口统计每 2 个汉字的共现频次。然后分别对于每一个汉字，统计所有与之共现的汉字频次，然后按频次占比转化为共现概率值，没有共现则概率为 0。所以每一个汉字 w_t 都对应一个长度为 V 的概率向量 $\boldsymbol{p}$，其中向量 $\boldsymbol{p}$ 位置 i 上概率值 p_i 表示汉字 w_t 与汉字 w_i 的共现概率，此概率向量作为嵌入网络的输出损失判别依据。汉字嵌入网络结构如图 5 所示。

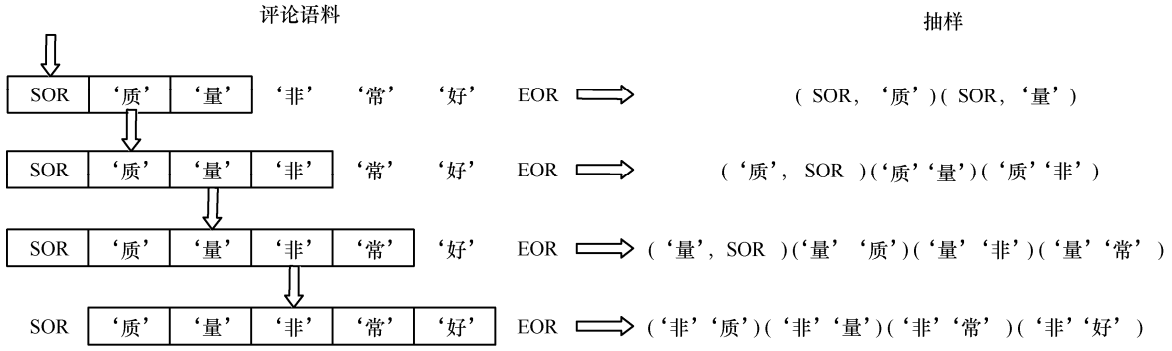

图 4 统计汉字共现概率

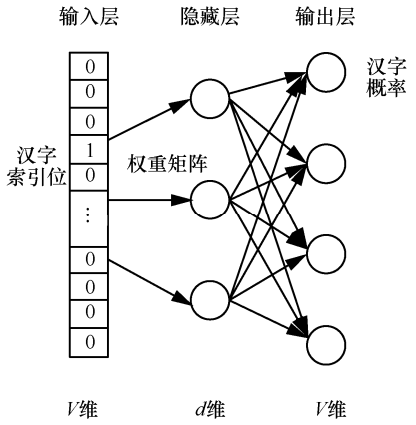

图 5　汉字嵌入网络

如图 5 所示，汉字嵌入网络包括一个节点数量为 V 的输入层，一个节点数量为嵌入向量长度 d 的线性隐含层和一个节点数量为 V 的概率输出层。其中输入层依次接收每个汉字的独热向量，输出层依次以对应汉字的概率向量为损失判别依据。嵌入网络通过如式(1)所示的最大化对数似然函数，反复迭代使输入汉字通过前向传播计算能够正确得到对应概率向量。其中输入层与隐层之间维度为 $V\times d$ 的全连接权重矩阵 $\boldsymbol{M}$ 为所需要的嵌入矩阵，$\boldsymbol{M}$ 对应行 i 所包含的权重向量即是索引为 i 汉字的嵌入向量。

$$L(w_t)=\frac{1}{n}\sum_{t=1}^{n}\sum_{-c\leqslant j\leqslant c,j\neq 0}\log p(w_{t+j}\mid w_t) \tag{1}$$

其中，c 表示汉字 w_t 的上下文环境，$P(w_{t+j}|w_t)$表示 w_t 与 w_{t+j} 的汉字共现概率。计算对数似然函数的误差后，使用反向传播算法更新权重矩阵，然后通过反复迭代，最终得到一个关于汉字集 V 的汉字嵌入向量矩阵。

4.2　多层 LSTM 网络

本文利用LSTM网络学习真实评论中的句子构成和排列规律。在处理序列数据的问题上，LSTM是一种被广泛应用的神经网络模型，例如短文本分类[29]、自动问答[30]、图片标注[31]、自动摘要[32]等。LSTM 模型通过记忆单元来存储和利用那些句子中长距离的依赖元素，这使 LSTM 能够学习较长的句子且保证良好的效果。多层 LSTM 网络使用多层非线性函数的组合来学习句子中每个汉字之间的概率分布，即当 2 个或多个汉字经常按序出现在一起时，网络可以捕捉到这种统计特征。由于在线评论往往是长短不一的，本文使用一种窗口方法来解决这个问题，窗口方法假设一个汉字的出现只与它前面窗口内的汉字排列顺序有关，和窗口外的汉字无关。具体为将训练语料中的所有评论首尾拼接在一起组成一个长文本，由于每条评论都有起始标记和结束标记，所以每条评论在长文本中是容易被划分的，也容易被 LSTM 模型捕捉到。然后设定一个固定窗口，窗口在长文本上以步长为 1 来滑动抽取汉字段交由 LSTM 模型学习，这个窗口通常设置为 100，这对生成短评论文本来说是足够的。

4.3　Softmax 全连接网络

在真实的攻击场景中，需要兼顾虚假评论的真实性及多样性，本文在 Softmax 全连接网络中引入了温度参数 T 来平衡评论真实性和多样性。在多分类问题上，Softmax 全连接网络是一种表现优异的神经网络结构，能够将 LSTM 网络输出的低维中间向量重新映射到汉字集中的每一个汉字。网络包括 d 个输入神经元和 v 个输出神经元，激活函数使用 Softmax 函数。因为攻击模型不仅需要生成逼真的虚假评论，而且需要虚假评论在数量上具有一定的规模，同时具有各种各样的类型，在训练语料的基础上可以产生新的词汇或句子的组合。所以本文使用 Softmax 函数时引入了一个重要的温度参数 T，具体如式(2)所示。

$$P(\mathrm{Softmax}(\boldsymbol{e}_t)=k)=\frac{\exp\{\frac{\boldsymbol{e}_t^j}{T}\}}{\sum_{j=1}^{v}\exp\{\frac{\boldsymbol{e}_t^j}{T}\}} \tag{2}$$

其中，k 表示汉字集中汉字的索引，$\boldsymbol{e}_t$ 表示全连接网络的输出向量，$\boldsymbol{e}_t^j$ 表示输出向量 $\boldsymbol{e}_t$ 在索引 j 的分量。当 T 小于 1 时，降低了攻击模型选择出现可能性较低字符的概率，所以攻击模型在生成文本序列时就会有较大概率选择可能性较大的一些字符，这样生成的评论质量较高，但也限制了文本生成的种类。随着温度参数 T 升高，攻击模型选择低概率字符的机率增大选择高概率字符的机率减少，这样就可以生成丰富多样的文本，当然随着多样性的丰富，生成的文本也会伴随着更多的错误，比如词汇组合不匹配、上下文不一致等问题。

5　实验与结果分析

在中文评论社区的自动化众包攻击中，生成的虚假文本具备真实性强、类型广泛等特点。因此，本文从以下几个角度来对比和分析攻击方法的有效性：1) 将传统面向英文文本的生成方法与本文方

法进行对比；2) 从语言学角度分析本文方法生成的虚假评论与真实评论的相似性；3)验证本文攻击方法在拷贝检测上优于传统基于词替换的攻击方法；4)验证本文生成的虚假评论能够欺骗用户并影响用户的消费决策。为了防御文中所提的攻击方法，本文中给出了一种可行的防御方法。

5.1 数据集

淘宝网的商品分类广泛，具有如电子产品、家居用品、衣物用品等多种分类。本文针对一种最常见的自动化众包攻击场景，从淘宝网的衣物评论社区中抓取了 51 121 条好评数据，来模拟攻击者提高商家口碑的攻击方式。原始评论集中的所有评论随机打乱顺序后，按 3:2 的比例划分，其中 30 672 条评论作为攻击模型训练数据集，20 449 条评论作为实验测试数据集。

5.2 英文生成方法对比

传统针对英文环境的自动化众包攻击方法通常采用 Char-LSTM 模型，已有实验[10]表明该模型可以生成高质量的英文评论。但 Char-LSTM 用于生成中文虚假评论有诸多缺点。1) 特征稀疏。直接将中文汉字映射为独热向量，则训练语料中包含大量的零值，稀疏的特征使模型训练缓慢。2) 参数数量多。Char-LSTM 模型训练所需的参数大约为本文模型的 10 倍。模型的参数往往决定模型训练所需的内存，显然更小的内存消耗使经济成本和时间成本都有所降低。3) 缺少汉字之间语义表示。在英文中，英文字母之间并没有语义联系。学习汉字之间的语义联系可以使生成中文评论更具灵活性，例如表 1 中的新词汇“炒鸡”（超级）和“稀饭”（喜欢），灵活使用汉字使生成的中文评论看起来更逼真。

针对以上问题，对比实验设置如下。对于 Char-LSTM 模型和汉字嵌入 LSTM 模型同样采用 CPU-i7-6700K、GPU-TiTanX 和内存 32 GB 的硬件环境，为了突出汉字嵌入 LSTM 模型的改进部分，这里设置 2 种模型相同部分的 LSTM 网络同样为 3 层，每层 1 000 个神经单元，激活函数采用 ReLU，初始学习率设置为 0.01，训练 400 轮。实验结果如表 2 所示。

如表 2 所示，在模型结构相同的情况下，Char-LSTM 模型所需要的参数约为 54 MB，汉字嵌入 LSTM 模型所需训练参数约为 4.3 MB，在模型训练中，参数的数量往往决定模型的内存和时间开销，进一步决定模型成本开销。并且，在迭代 400 轮后，Char-LSTM 的损失仍然远高于汉字嵌入 LSTM 模型，训练所需时间也明显多于汉字嵌入 LSTM 模型。对比实验表明汉字嵌入 LSTM 在处理中文评论上性能优于 Char-LSTM 模型。

5.3 生成评论与真实评论对比

基于语言学分析的分类器也是目前主要的虚假评论检测手段[6,33-34]，攻击方法生成的虚假评论文本要欺骗分类器检测，就要在这些语言学细节上接近真实评论文本。对于虚假评论文本的语言学分析主要分为 3 个维度[34]：基于语法分析维度、基于语义分析维度、基于文体元数据分析维度。本文从这 3 个维度中提取出 5 组共 113 个分类特征用于训练软间隔 SVM 线性分类器。

1) 词袋特征。词袋特征是一种不考虑词序将句子表示为一个词集合的方法。词袋特征在文本分类、情感极性分析、图像识别等领域中应用广泛。在恶意众包论坛中的虚假评论数据集检测上，词袋特征能够达到 89.6%的准确率[35]。本文使用 jieba 分词工具[36]将一条评论中的每个句子表示为一个词集合，然后将不同词集合的最大余弦相似度来作为该条评论的一个分类特征。

2) 结构特征。评论的结构特征用来表示一条评论的基本结构，包括评论中的词汇数量、句子数量、句子的平均词汇数量和词汇的平均汉字数量，一共 4 个分类特征。

3) 词性特征。通过对评论中的词汇进行词性标注，并统计词频的方式表示为评论的词性特征。Li 等[37]的研究表明真实评论和虚假评论在语法特征上呈现不同的特点：真实评论包含更多的名词和形容词，而虚假评论包含更多的动词、副词和代词。

表 2 模型实验对比

模型	结构	参数数量/MB	误差					训练时长/h
			迭代 0 次	迭代 10 次	迭代 50 次	迭代 100 次	迭代 400 次	
Char-LSTM	3 层 LSTM	54	100%	83.14%	63.57%	40.39%	39.42%	24
汉字嵌入 LSTM	3 层 LSTM	4.3	100%	60.13%	42.72%	31.2%	28.93%	5

本文使用 jieba 中的 POS(part of speech)工具，将评论进行分词并标注词性，然后分别统计名词、形容词、动词、副词和代词的数量，作为 5 个分类特征。

4) 极性特征。情感极性分析是一种语义分析方法，通过统计不同情感词的数量来分析一条评论所代表的情感倾向。研究表明虚假评论比真实评论包含更多的情感词，表现为更积极或更消极[37]。本文使用 snownlp 工具[38]给出评论的情感分数，分数在[0,1]之间，然后将情感分数距中性分数(0.5)的距离作为一个分类特征。

5) 文体特征。LIWC(linguistic inquiry and word count)工具[39]被广泛应用于英文文本的文体特征提取。本文使用基于 LIWC 开发的中文文体特征提取工具 TextMind[40]，TextMind 将简体中文常用的 5 000 个词汇划分为 102 个类别，例如社交、情感、认知、感知、生理、空间、时间、金钱、宗教等。本文使用 TextMind 工具对评论进行词汇标注，然后统计每类词的数量作为 102 个分类特征。

实验使用测试集中随机抽取的 1 000 条真实评论和攻击模型生成的 1 000 条虚假评论用于训练软间隔 SVM 线性分类器，并采用 10 折交叉验证来计算平均精确率和召回率。精确率(precision)表示为由分类器成功标记为虚假评论和所有由分类器标记为虚假评论数量的百分比，召回率(recall)表示为由分类器标记为虚假评论和数据集中所有虚假评论数量的百分比，精确率和召回率随温度参数 T 的变化如图 6 所示。

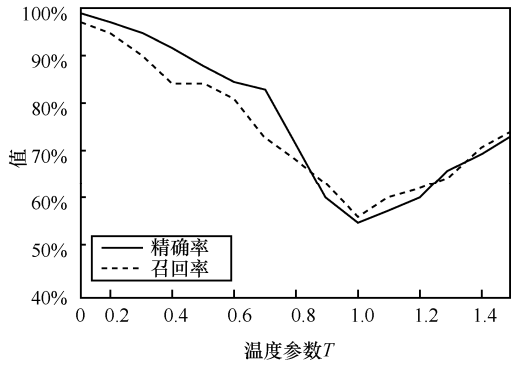

图 6　语言学分析−对比真实评论和生成评论

在温度参数 T 为 0.1 时，精确率和召回率都接近 100%，这表示 SVM 分类器很够准确地标记出虚假评论。随着温度参数 T 增大，分类器的分类能力下降。在温度参数为 1.0 时攻击效果最好，精确率为 55%，召回率为 56%，这表示分类器接近随机分类（50%分类精确率），分类器并不能很好地辨别出数据集中的虚假评论。随着温度参数 T 的继续增大，攻击模型生成的虚假评论质量下降。对比实验表明，当温度参数 T 在 0.8 到 1.2 时，攻击模型生成的虚假评论具有和真实评论一样的真实性。

5.4　生成评论与训练语料对比

在语言学分析的角度上，虽然攻击模型生成的虚假评论可以欺骗分类器的检测，但是传统基于名词替换的攻击方法也可以产生接近真实的虚假评论。这是因为基于名词替换的方法采用数个通用的模板，这些模板基本来自某些典型的真实评论，然后通过把模板中物品或服务的名词，替换为与攻击目标相关的名词来生成虚假评论。这种传统攻击方法生成的虚假评论除了部分名词不同于模板外，其他文本部分和模板完全相同，这就使基于语言学分析的分类器并不能检测出这种攻击方式。为了进一步对比和分析本文攻击方法优于传统名词替换的攻击方法，这里使用拷贝检测来验证。拷贝检测通过计算 2 个文本的内容相似度来判断文本是否存在剽窃行为。这里采用 Winnowing[41]工具来进行拷贝检测实验，Winnowing 是一种已被证明效果很好且被广泛应用的方法。Winnowing 方法使用散列函数将文本映射为一组与之对应且唯一的指纹信息，然后对于待检测文本和待比较数据集都生成这样的指纹信息，最后采用 Jaccard 指数计算指纹信息之间的相似度。本文采用待检测文本与待比较数据集中每条文本相似度的最大值作为拷贝检测分数，分数范围为 0~1，分数越高则代表待检测文本越有可能是拷贝而来的虚假评论。

实验假设用户评论社区的防御工作者可以获取到包含攻击模板的评论数据集，实际上这很容易做到，防御工作者可以查找所有已经存在于评论社区的评论，而基于名词替换的虚假评论一旦出现在评论社区中就可以被用来检测其他新发布的评论。

对于本文攻击方法，选取训练集中的所有真实评论作为待比较数据集，选取攻击模型生成的 1 000 条虚假评论作为待检测数据集，本文方法的拷贝检测分数为待检测中所有虚假评论拷贝检测分数的平均值；对于名词替换方法，同样选取训练集中的所有真实评论作为待比较数据集，然后基于该数据集抽取所有标注为名词词性的词汇组成名词集，并随机从待比较数据集中抽取 1 000 条真实评论作为模板，用名词集中的名词随机替换掉模板中的名词，

最后将这 1 000 条评论作为待检测数据集，实验结果如图 7 所示。同时选取测试集中所有评论和训练集中的评论作为参照，实验结果如图 8 所示。

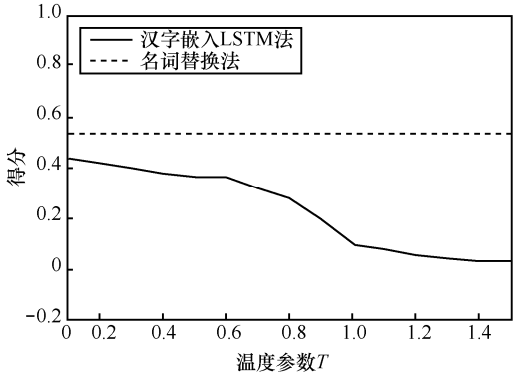

图 7 拷贝检测−对比名词替换方法和汉字嵌入 LSTM 方法

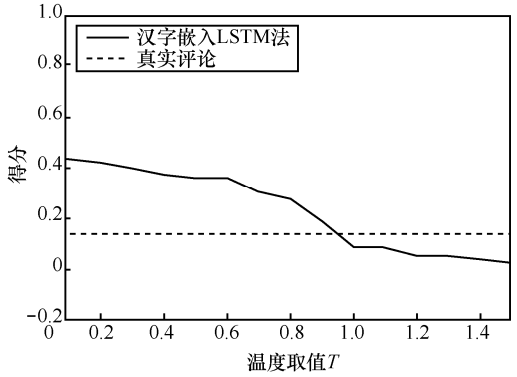

图 8 拷贝检测−对比汉字嵌入 LSTM 方法和真实评论

如图 7 所示，名词替换法的拷贝得分为 0.536 分，高于所有温度下的汉字嵌入 LSTM 方法，即 Winnowing 拷贝检测工具更容易检测出名词替换方法生成的虚假评论，随着温度参数 T 增大，汉字嵌入 LSTM 方法的拷贝得分逐渐变小，生成的虚假评论更不容易被检测。同时，图 8 说明本文攻击方法生成的虚假评论不是简单地复制训练集中的真实评论，而是在训练集的基础上生成了新的评论。在温度参数 T 为 0.9 时，虚假评论有最接近真实评论的拷贝得分。实验表明本文方法比传统基于名词替换的攻击方法有更好的实用性。

5.5 生成评论的用户调查

本文使用用户调查的方式，验证攻击方法生成的虚假评论对用户的影响。考虑到具体的攻击场景，实验邀请 50 名具有丰富安全背景知识的参与者，填写调查问卷。50 名参与者被分为 5 组，每组 10 人，每组的调查问卷分别对应温度参数 T 为 0.2、0.5、0.8、1.0 和 1.3。每个参与者被分配 30 条评论，包括测试集中随机抽取的 20 条真实评论和 10 条攻击模型生成的虚假评论。问卷要求参与者判断 30 条评论中每条评论是否有提高商品口碑的作用，并标记出 10 条认为是模型生成的虚假评论。实验统计标准采用召回率和有效率，召回率(recall)定义为成功标记出虚假评论和问卷中所有虚假评论数量的百分比，有效率(helpfulness)定义为问卷中所有虚假评论中被标记为有效的百分比。统计所有调查结果取平均值后结果如表 3 所示。

随着温度参数 T 逐渐变大，参与者识别虚假评论的召回率在逐渐下降，这表明越来越多的虚假评论被参与者遗漏，且更多的虚假评论被标记为有作用。当温度参数为 1.0 时，攻击模型表现最好，参与者遗漏掉了 43%的虚假评论并将 48%的虚假评论标记为有作用。调查表明本文攻击方法产生的虚假评论具有欺骗用户和并影响用户消费决策。

表 3 用户调查的实验结果

T	召回率	有效率
0.3	97%	0
0.5	66%	24%
0.8	52%	37%
1.0	43%	48%
1.3	63%	31%

5.6 防御方法

针对本文的自动化众包攻击方法，这里给出一种可行的防御方法。在理想情况下，防御模型用于辨别的任何标准都可以由攻击模型来对应解决，所以使用一种机器学习模型防御另一种机器学习模型的攻击是不可行的。但是对于同一数据集，无论训练多少种机器学习模型，每个模型所能学习到的句子概率分布是相同的。根据数据集所含信息量一定的这个特点，本文给出了一种可行的防御方法，如图 9 所示。

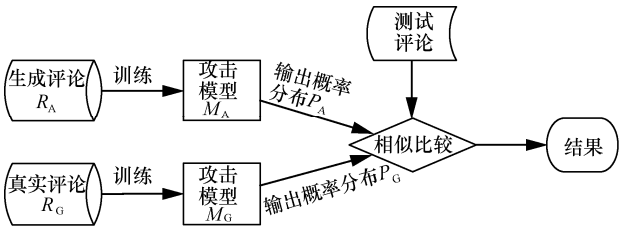

图 9 中文自动化众包攻击的防御方法流程

本文防御方法假设：1) 评论社区的防御工作人

员可以获得用于训练攻击模型的评论训练集 R_A，并且可以获取一个和训练集 R_A 不相同的真实评论数据集 R_G；2) 评论社区的防御工作人员了解攻击模型的结构。防御方法流程如下：针对训练评论数据集 R_A 和真实评论数据集 R_G 这 2 个数据集，分别使用本文攻击方法训练攻击模型 M_A 和 M_G，训练好的攻击模型可以根据输入汉字预测下一个汉字的概率分布。然后将待检测评论逐字输入到 2 个攻击模型中，得到 2 个不同的概率分布 P_A 和 P_G，通过对数似然比(log-likelihood ratio)计算概率分布 P_A 与概率分布 P_G 的距离。最后将待检测评论的中每个汉字得到的对数似然比求和取平均得到平均距离，若平均距离接近攻击模型 P_A 则判定该条评论是虚假评论，否则判定评论为真实评论。防御方法流程的伪代码如算法 1 所示。

算法 1　防御方法

输入

攻击评论语料集 R_A;

真实评论语料集 R_G;

测试评论 T;

输出

评论是否虚假

procedure defense(A, B):

$N \leftarrow$评论 T 的长度

$\text{LSTM}_A \leftarrow$训练语料集 R_A

$\text{LSTM}_G \leftarrow$训练语料集 R_B

for t = 1 to N-1 do

将 X_t 输入 LSTM_A 模型

$P_A \leftarrow P_{\text{LSTM}}(X_{t+1}=x_{t+1} \mid x_1, \cdots, x_t)$

将 X_t 输入 LSTM_G 模型

$P_G \leftarrow P_{\text{LSTM}}(X_{t+1}=x_{t+1} \mid x_1, \cdots, x_t)$

$d \leftarrow \log \dfrac{P_A}{P_G}$

$D \leftarrow \sum \dfrac{d}{N-1}$

if $D > 0$ then

返回虚假评论

else

返回真实评论

end if

end for

针对防御方法和基于语言学分类器方法的对比实验，采用精确率作为对比标准，精确率(precision)定义为成功标记为虚假评论和所有标记为虚假评论数量的百分比，实验结果对比如图 10 所示。

如图 10 所示，随着温度参数 T 的增大，防御方法检测虚假评论的精确率稳定在 97%左右，语言学分类器的精确率逐渐降低，最低处为 53%。实验表明防御方法能够有效检测出攻击模型生成的虚假评论，但是该防御方法的不足也是明显的，需要假设训练数据集已知，因此如何设计更加实用的防御方法是未来研究的重点。

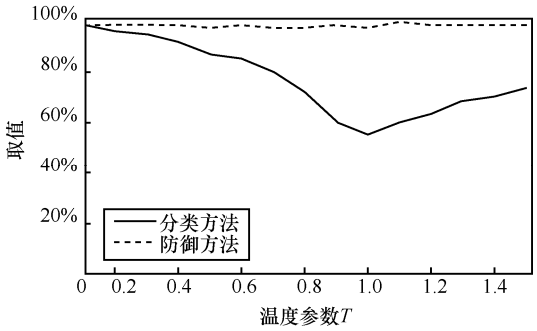

图 10　防御方法与语言学分类器的对比

6　结束语

本文面向中文用户评论社区，提出了一种有效的自动化众包攻击方法。利用汉字嵌入方法，LSTM 网络、Softmax 全连接网络和引入温度参数 T 的随机抽样方法，设计并实现了一种自动化众包攻击方法，该方法可以生成大量具有真实性、多类型及实用性的虚假评论，能够有效地满足大规模、真实场景下的攻击要求。实验中，本文将传统面向英文文本的生成方法与本文方法进行对比。在语言学角度分析文中方法生成的虚假评论与真实评论的相似性。验证本文攻击方法在拷贝检测上优于传统基于词替换的攻击方法。验证本文生成的虚假评论能够欺骗用户并影响用户的消费决策。实验表明本文提出的攻击模型和方法是可行且有效的。本文中提出了一种可能的防御方法，但是该方法假设模型训练的数据集已知。因此在未来的研究中，需要重点研究针对这类攻击的有效防御方法，增强用户评论社区抵御针对自动化众包攻击的安全威胁。

参考文献：

[1]　AKOGLU L, CHANDY R, FALOUTSOS C. Opinion fraud detection

in online reviews by network effects[C]//The International AAAI Conference on Weblogs and Social Media. AAAI, 2013:1-12.

[2] LACKERMAIR G, KAILER D, KANMAZ K. Importance of online product reviews from a consumer's perspective[J]. Advances in Economics and Business, 2013, 1(1):1-5.

[3] 金立印. 网络口碑信息对消费者购买决策的影响：一个实验研究[J]. 经济管理, 2007(22):36-42.
JIN L Y. The impact of internet word-of-mouth information on consumers' purchase decisions: an experimental study[J]. Economic Management, 2007(22): 36-42.

[4] LUCA M, ZERVAS G. Fake it till you make it: reputation, competition, and yelp review fraud[J]. Management Science, 2016, 62(12): 3412-3427.

[5] LIPSMAN A. Online consumer-generated reviews have significant impact on offline purchase behavior[C]// Industry Analysis, comScore Inc. 2007:2-8.

[6] JINDAL N, LIU B. Opinion spam and analysis[C]// International Conference on Web Search and Data Mining. ACM, 2008:219-230.

[7] WANG G, WILSON C, ZHAO X, et al. Serf and turf: crowdturfing for fun and profit[C]//The International Conference on World Wide Web. ACM, 2012: 679-688.

[8] LEE K, WEBB S, GE H. Characterizing and automatically detecting crowdturfing in Fiverr and Twitter[J]. Social Network Analysis and Mining, 2015, 5(1):1-6.

[9] 孟美任, 丁晟春. 虚假商品评论信息发布者行为动机分析[J]. 情报科学, 2013(10):100-104.
MENG M R, DING S C. Analysis of behavioral motivation of publishers of false commodity commentary information[J]. Information Science, 2013(10):100-104.

[10] YAO Y, VISWANATH B, CRYAN J, et al. Automated crowdturfing attacks and defenses in online review systems[J]. ACM, 2017：1143-1158.

[11] BARTOLI A, LORENZO A D, MEDVET E, et al. "Best dinner ever!!!": automatic generation of restaurant reviews with LSTM-RNN[C]// IEEE/WIC/ACM International Conference on Web Intelligence. IEEE, 2017:721-724.

[12] MIKOLOV T. Statistical language models based on neural networks[M]. Mountain View, Presentation at Google, 2012:2-80.

[13] WEN T H , GASIC M , MRKSIC N , et al. Semantically conditioned LSTM-based natural language generation for spoken dialogue systems[J]. Computer Science, 2015:1711-1721.

[14] ZHANG Q, WANG D Y, VOELKER G M. DSpin: detecting automatically spun content on the web[C]//The Network and Distributed System Security Symposium. 2014:1-11.

[15] 杨钊, 陶大鹏, 张树业, 等. 大数据下的基于深度神经网的相似汉字识别[J]. 通信学报, 2014, 35(9):184-189.
YANG Z, TAO D P, ZHANG S Y, et al. Similarity of Chinese character recognition based on deep neural network under big data[J]. Journal on Communications, 2014, 35(9):184-189.

[16] 张蕾, 章毅. 大数据分析的无限深度神经网络方法[J]. 计算机研究与发展, 2016, 53(1):68-79.
ZHANG L, ZHANG Y. Infinite depth neural network method for big data analysis[J]. Journal of Computer Research and Development, 2016, 53(1): 68-79.

[17] MURPHY K P. Machine learning: a probabilistic perspective[J]. MIT press, 2012, 27(2):62-63.

[18] TAI Y, HE H, ZHANG W Z, et al. Automatic generation of review content in specific domain of social network based on RNN[C]// IEEE International Conference on Data Science in Cyberspace. IEEE, 2018:601-608.

[19] AULI M, GALLEY M, QUIRK C, et al. Joint language and translation modeling with recurrent neural networks[C]//The 2013 Conference on Empirical Methods in Natural Language Processing. 2013:1044-1054.

[20] KANNAN A. Smart reply: automated response suggestion for email[C]//The ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2016:955-964.

[21] KARPATHY A, FEI-FEI L. Deep visual-semantic alignments for generating image descriptions[J]. IEEE Conference on Computer Vision and Pattern Recognition, 2015:3128-3137.

[22] SERBAN IV, SORDONI A, BENGIO Y, et al. Hierarchical neural network generative models for movie dialogues[J]. Cornell University, arXiv:1507.04808.

[23] SHANG L, LU Z, LI H. Neural responding machine for short-text conversation[C]//The Annual Meeting of the Association for Computational Linguistics and The International Joint Conference on Natural Language Processing, Association for Computational Linguistics. 2015(1): 1577-1586.

[24] SUN Y, LIN L, TANG D, et al. Radical-enhanced Chinese character embedding[J]. Lecture Notes in Computer Science, 2014:279-286.

[25] Graves A. Generating Sequences With Recurrent Neural Networks[J]. Cornell University, arXiv:1308.0850.

[26] LEBRET R, GRANGIER D, AULI M. Neural text generation from structured data with application to the biography domain[C]//The Conference on Empirical Methods in Natural Language Processing. 2016:1203-1213.

[27] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality[J]. Advances in Neural Information Processing Systems, 2013(26):3111-3119.

[28] 曾谁飞, 张笑燕, 杜晓峰, 等. 基于神经网络的文本表示模型新方法[J]. 通信学报, 2017, 38(4):86-98.
ZENG S F, ZHANG X Y, DU X F, et al. A new method of text representation model based on neural network[J]. Journal on Communications, 2017, 38(4):86-98.

[29] WANG X, LIU Y, CHENGJIE S U, et al. Predicting polarities of tweets by composing word embeddings with long short-term memory[C]//The Annual Meeting of The Association for Computational Linguistics, The International Joint Conference on Natural Language Processing. 2015(1): 1343-1353.

[30] SUTSKEVER I, VINYALS O, LE Q V. Sequence to sequence learning with neural networks[J]. In Advances in Neural Information Processing Systems. 2014: 3104-3112.

[31] XU K, BA J, KIROS R, et al. Show, attend and tell: neural image

caption generation with visual attention[C]//International Conference on Machine Learning. 2015:2048-2057.

[32] RUSH A M, CHOPRA S, WESTON J. A neural attention model for abstractive sentence summarization[C]//The Conference on Empirical Methods in Natural Language Processing. 2015:379-389.

[33] CRAWFORD M , KHOSHGOFTAAR T M , PRUSA J D , et al. Survey of review spam detection using machine learning techniques[J]. Journal of Big Data, 2015, 2(1):1-24.

[34] 李璐昄, 秦兵, 刘挺. 虚假评论检测研究综述[J]. 计算机学报, 2018, 41(4): 946-968.
LI W, QIN B, LIU T. A review of false comment detection research[J]. Chinese Journal of Computers, 2018, 41(4): 946-968.

[35] OTT M, CHOI Y, CARDIE C, et al. Finding deceptive opinion spam by any stretch of the imagination[C]//The Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics. 2011(1):309-319.

[36] LI J, OTT M, CARDIE C, et al. Towards a general rule for identifying deceptive opinion spam[C]//The Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics. 2014(1):1566-1576.

[37] CHEN C, CHEN J, SHI C. Research on credit evaluation model of online store based on snow NLP[C]// In E3S Web of Conferences, EDP Sciences. 2018(53): 3-39.

[38] HUANG C L, CHUNG C K, HUI N, et al. The development of the Chinese linguistic inquiry and word count dictionary[J]. Chinese Journal of Psychology, 2012, 54(2):185-201.

[39] GAO R , HAO B , LI H , et al. Brain and health informatics: developing simplified Chinese psychological linguistic analysis dictionary for microblog[M]. Berlin: Springer, 2013:359-368.

[40] SCHLEIMER S, WILKERSON D S, AIKEN A. Winnowing: local algorithms for document fingerprinting[C]//The ACM SIGMOD International Conference on Management of Data. ACM, 2003: 76-85.

[作者简介]

王丽娜（1964– ），女，辽宁营口人，博士，武汉大学教授、博士生导师，主要研究方向为软件和系统安全、信息隐藏、人工智能安全等。

郭晓东（1994– ），男，山西大同人，武汉大学硕士生，主要研究方向为自然语言处理、人工智能安全等。

汪润（1991– ），男，安徽安庆人，武汉大学博士，主要研究方向为软件和系统安全、人工智能安全等。