# Miti通过对比学习遏制人类与计算机意见欺诈

YULIYA TUKMACHEVA, 斯科尔科沃科学技术研究院，俄罗斯IVAN OSELEDETS*，斯科尔科沃科学技术研究院，俄罗斯EVGENY FROLOV,斯科尔科沃科学技术研究院，俄罗斯

我们提出了一种用于在协同过滤推荐系统中检测虚假文本评论的新方法。现有算法专注于检测由语言模型生成的虚假评论，却忽略了不诚实用户撰写的文本，这些用户撰写评论主要是为了获取经济利益。我们提出了一种基于对比学习的架构，该架构利用用户的人口统计特征以及文本评论，作为识别虚假评论的额外依据。这样，我们便能应对两种不同类型的虚假评论刷评行为，并增强推荐系统抵御偏颇评论的能力。

**ACM参考文献格式：**
Yuliya Tukmacheva、Ivan Oseledets 和 Evgeny Frolov。2022。通过对比学习缓解人为和计算机意见欺诈。1，1（2022年1月），15页。https://doi.org/XXXXXXX.XXXXXXX

arXiv:2301.03025v1 [cs.AI] 8 Jan 2023

## 1 引言

目前，推荐系统已深度融入许多热门的在线商业平台。这些平台收集用户和商品的属性数据，并积累产品和服务的评分与反馈，以开发能够显著提升用户市场体验的算法。这些算法能够通过以下方式影响用户的购买行为：（1）为用户提供最相关的个性化商品选择；（2）降低个人搜索成本；（3）通过反馈机制，缓解拥有同质化卖家和商品的大型商业平台中的信息不对称。由于推荐系统具有影响用户营销决策的能力，因此它们已成为评分和评论操纵（也称为攻击）的诱人目标。具体而言，这些攻击旨在提高或降低某些商品的排名和文本评论，或单纯破坏商业平台整体的效率与可信度。

本研究致力于解决在拥有文本评论的平台上过滤欺骗性意见和检测异常行为的任务。之所以着重关注文本评论，是因为与容易被操纵的星级评分系统相比，文本是更具信息量、也更可靠的产品和卖家质量信息来源（见 [19], [14], [27], [28]）。欺骗性意见可以通过两种方式生成：（1）雇用专业人类写手为获取报酬而评论产品；（2）利用最先进水平（SOTA，以下同）的自动语言模型模仿人类言语。这一区别至关重要，因为

*以及人工智能研究院。

作者地址：Yuliya Tukmacheva，斯科尔科沃科学技术研究院，121205，莫斯科，俄罗斯；Ivan Oseledets，斯科尔科沃科学技术研究院，121205，莫斯科，俄罗斯；EvgenyFrolov，斯科尔科沃科学技术研究院，121205，莫斯科，俄罗斯，evgeny.frolov@ skoltech.ru。

允许为个人或课堂使用而免费制作本作品全部或部分内容的数字或纸质副本，前提是副本不得为营利或商业利益而制作或分发，且副本须附有本声明及首页的完整引文。对于本作品中由 ACM 以外的其他方拥有版权的部分，必须尊重其版权。注明出处后可以摘录。若要以其他方式复制、重新发表、发布到服务器或分发至列表，则须事先获得具体许可和/或支付费用。请向 permissions@acm.org 申请许可。© 2022 计算机协会。稿件已提交至 ACM

现有算法并未考虑这一区别，只专注于过滤人工文本评论 [22]，而没有将雇用专业写手的问题纳入考量。虚假评论写手市场的规模不容低估，因为所有在线评论中近 4% 是真实人类发表的不诚实意见，这导致在线市场的收入损失超过 1520 亿美元[1][2]。

本研究旨在开发一种机器分类器，供推荐平台用于：（1）检测由最先进的生成式语言模型（如基于 Transformer 的生成式预训练 Transformer（GPT）模型（见 [21], [5]））生成的观点欺诈攻击；（2）过滤不诚实的人类写手撰写的偏见文本，从而确保系统对这两类攻击策略的稳健性。一般而言，推荐系统，尤其是基于协同过滤的推荐系统[3]，面临冷启动问题：缺乏交互历史、行为轨迹或既定偏好的新用户，无法与新创建的档案自动且准确地区分开来。本研究弥补了以往研究的局限，并针对冷启动场景下的这两类攻击。为此，我们利用用户人口统计特征和文本评论，将其作为识别虚假观点的额外证据 [22], ；这一做法基于 [1], 论文的前提，该论文利用用户属性检测针对星级评分系统的攻击。其背后的逻辑如下：具有相同属性的用户最有可能拥有相同的偏好。由于用户的人口统计信息并非公开信息，只有系统管理员才能获取，攻击者会创建元数据随机分配的用户档案。由此，我们假设真实文本评论与用户属性之间存在潜在联系，而虚假文本（无论由人类还是语言模型撰写）则不呈现这种联系。

本文的主要贡献如下：

- 将文本评论与用户元数据关联起来，以筛除虚假文本评论，这一设定是新的，现有研究尚未涉及。利用文本评论与人口统计元数据之间潜在关联的想法尚未经过检验，而推荐系统可以采用所提出的方法，比以往更高效地筛除虚假文本评论。因此，本研究为后续相关领域的其他研究奠定了基准。

- 本研究旨在解决检测不诚实用户以及生成式语言模型使用情况的问题。现有算法和方法只专注于检测人工生成文本，却忽视了不诚实人类为了经济利益而撰写偏见文本的情况。所提出的方法兼顾了这两类攻击，使推荐系统更加稳健可靠。

- 本研究为推荐系统中的欺诈检测数据库新增了两个数据集。第一个数据集是经过处理和扩充的 Google 本地商家评论与元数据，并加入了从用户元数据中精心构建的工程化特征，以及通过所提出的抽样流程提取的标签。该数据集可进一步用于训练和测试针对真实用户不诚实评论的算法。第二个数据集由 GPT-J-6B模型 [25]生成的虚假文本评论构成，并为其随机分配了用户人口统计属性，可进一步用于训练和

[1]https://theprint.in/opinion/almost-4-of-all-online-reviews-are-fake-their-impact-is-costing-us-152-billion/715689/. [2] https://www.cs.uic.edu/ liub/FBS/fake-reviews.html. [3]基于协同过滤的推荐系统利用具有相似偏好的用户群体之间的关联，提供自动推荐。

测试用于检测人工生成文本评论的模型。第 3 节和第 3 节将更详细地介绍这些数据集的处理和特征工程。

本文其余部分安排如下。第 2 节介绍推荐系统中欺诈检测的现有研究，并讨论其不足。第 3 节介绍方法，并阐述本研究的理论框架。第 4 节介绍实证应用，报告估计结果并开展对比实验。第 5 节讨论所得结果，回顾研究局限并作出总结。

## 2 相关工作

业界对稳健且有效的虚假内容检测算法的需求，推动了这一研究领域的广泛研究。推荐系统攻击检测的一个有前景的研究方向，是在动态纳什强化学习框架下模拟攻击（见 [9], [17], [24]）。这种方法通过模拟攻击环境并定义恶意用户的成本函数，推导其最优策略并加以应对。然而，在大型在线市场中，这种方法并不适用，因为其中可能存在成本函数和动机各异的异质攻击者。本文提出的方法克服了这一局限，能够处理异质恶意智能体，且不依赖于对攻击者策略的近似。此外，该方法也适用于这样的情况：同一用户可能既被判定为恶意用户，又被视为可信用户；例如，真实用户可能针对某类商品充当垃圾信息发送者，而针对另一类商品则是普通顾客。我们的模型将每对属性与文本视为独立于此前交互，并能够涵盖此类（尽管实际中罕见的）情况。

另一类用于检测欺诈性用户档案的模型，利用多种统计指标来过滤异常。已有研究提出了如下统计指标：(1) 商品推荐集合中评分与中位数的偏差 [14], (2) 对评分和情感值极端的商品进行评分 [27] 以及 (3) 在狭窄时间窗口内给出的评分 [28]。然而，所提出的统计指标很容易被规避，因此应采用更先进的检测指标。

另一个研究方向是采用聚类方法（见 [15], [3]）。这些模型试图找出高度连接的用户团，并利用聚类方法识别异常用户档案。用于虚假内容过滤的最先进聚类模型是基于支持向量机（SVM）的模型。在 [2] 中，研究提出了支持向量机与高斯混合模型（SVM-GMM）+；在 [23] 中，则采用了支持向量机与主成分分析（SVM-PCA）+方法。然而，这些方法都面临冷启动问题。所提出的方法将每个观测都视为新的观测，因此能够在冷启动条件下检测虚假内容。

用于推荐系统攻击检测的更先进技术包括潜变量模型，这类模型能够对用户进行真实与虚假类型的编码。在 [16], 中，潜变量模型通过求解最大化星级评分系统中评分分布熵的问题来揭示用户类型。在 [26], 中，潜变量模型通过抵御对抗性投毒攻击发挥作用；这类攻击是指旨在最小化经验风险的虚假用户档案生成过程。[1] 中描述的算法也采用了生成式因子模型，将用户元数据用作识别虚假评分的额外证据来源。本研究采用了该模型的基本思路，不过上述所有变分推断模型都是为星级评分平台开发的，并且忽略了文本评论操纵。正如上文所述，星级评分系统本身提供给用户的信息并不充分，而文本评论被认为是对顾客而言更可靠的信息来源。因此，

能够处理由人类和语言模型撰写的文本评论的方法，在社会层面更受欢迎。因此， [1] 中提到的假设促使我们使用用户元数据来生成模型，使其能够处理文本评论并过滤两种攻击策略。现有论文尚未研究这一问题，因此也没有可供比较的基线方法。 [22]提出了目前最先进的、仅基于评论本身判断文本评论真伪的方法。该方法介绍了使用 GPT-2 模型 [21], 生成虚假评论的流程，并对稳健优化的 BERT 预训练方法（RoBERTa） [18] 模型进行微调，以将文本评论分类为真实或虚假。我们将该模型作为基准，与所提出的方法进行比较。这样，我们也可以检验用户元数据是否确实有助于判断评论的真伪。

## 3 方法

### 3.1 虚假评论生成流程

*3.1.1 GPT-J-6B*模型。 为了模拟虚假评论攻击，我们计划使用最先进的无监督、基于 Transformer 的多任务语言模型，具体而言，即生成式预训练 Transformer（以下简称 GPT）模型（见 [21],[5]）。

这些模型旨在执行多任务学习，例如机器翻译、文本摘要和理解，以及问答。本研究采用 GPT-J-6B [25] 模型——GPT-3 模型 [5] , 最新一代 GPT 模型的民主化开源版本——根据处理后的问答历史生成答案。

GPT-J-6B模型是由 EleutherAI 开发的自回归文本生成模型，由多个 Transformer 解码器块堆叠而成，每个块又由掩码自注意力层、编码器-解码器自注意力层和前馈神经网络组成。它拥有 60 亿个参数，并在大型 Pile 数据集合 [11],上进行训练；该数据集包含 800 GB 的文本，用于训练语言模型。给定文本上下文后，模型旨在预测序列中的下一个词。由于接受了大量文本数据的训练，该模型在不同的零样本下游任务中表现相对出色 [10][6],，能够生成简洁且语法通顺、足以模仿真人言语的文本。GPT-J-6B模型依赖迁移学习，因此无需额外训练或调整数据集即可完成任务。因此，我们可以构建由 GPT-J-6B模型处理的问答提示，从而大规模生成虚假评论。

*3.1.2* 生成流程。将文本评论集作为真实文本示例后，恶意用户便可利用最先进水平的生成模型来模拟人类撰写的文本。我们根据真实评论制定了虚假评论生成流程，具体步骤如下：

(1) 对文本评论进行标准预处理；(2) 使用TextRank4Keywords算法提取每条文本评论的关键词；(3) 选择部分关键词发送至提示中；(4) 将关键词随机替换为同义词/反义词；(5) 构建问答提示，供GPT-J-6B模型用于虚假评论生成；(6) 对用户属性进行采样，并将其与虚假文本随机匹配，以生成攻击者画像数据集；(7) 将虚假数据集与原始数据集的随机子集合并，以获得用于虚假内容分类任务的平衡数据集。

步骤 1. 对文本评论进行标准预处理，包括删除冗余符号和停用词。

步骤 2. 接下来，我们应打乱数据集，并按类别分组。对于每个类别，我们迭代地抽取 4 条评论，并使用TextRank4Keywords算法从每条评论中提取关键词。TextRank4Keywords算法基于 PageRank算法，通过以下公式计算网站的权重，并将网页表示为有向图：

$$S(V_i) = (1-d) + d \cdot \sum_{j \in \mathrm{In}(v_i)} \frac{1}{|\mathrm{Out(V_j)}|} S(V_j),$$

其中，$S(V_i)$ 是第 $i$ 个节点（网页）的权重，$d$ 是用于没有出边的节点的阻尼因子（我们设为 $d = 0.85$），$\mathrm{In}(V_i)$ 是节点 $i$ 的入边集合，$\mathrm{Out}(V_j)$ 是节点 $j$ 的出边集合。

要将 PageRank 用于文本，我们将每条评论拆分为句子，并指定窗口大小 $k$（我们设为 $k = 4$）。因此，窗口中的任意两个词组成一条无向边，然后使用上述公式计算该图的权重。之后，我们选取文本评论中最重要的 10 个词。为简便起见，我们标注词语的词性，并仅分析名词、动词、副词和形容词。

步骤 3. 我们从某个类别中 4 条经过处理的文本评论各自包含的唯一关键词集合里，随机选取四分之一，并将其作为提示生成的基础词。

步骤 4. 由于基于GPT的模型似乎对先前查询中使用的词语很敏感，因此，对于见过的词语，它们会返回示例答案，所以我们应将所有词语替换为随机选取的同义词。因此，如果该子集中存在同义词，我们就将其中所有词语替换为对应的同义词。此外，值得注意的是，攻击者意图抬高或压低产品排名，因此会生成具有预期情感倾向的文本。为此，我们以 0.5 的概率将每个选定的关键词替换为其反义词，以考虑所有评论整体上都偏正面或偏负面的情况。这样，我们就能获得情感倾向多样的评论，并通过不同文本丰富数据收集。

步骤 5. 我们按以下方式生成问答提示。从每条评论中提取关键词，并将其视为查询。这些查询的答案将是与这些关键词对应的真实文本评论。然后，对于虚假评论，我们将包含同义词或反义词的随机子集作为查询发送，并要求生成式Transformer模型生成评论。

步骤 6。 有了虚假评论文本后，可以将这些文本与随机用户属性匹配。由于用户元数据的分布信息并不广为人知，攻击者只能随机分配用户特征。为此，我们从真实元数据中随机抽取用户属性，并将其分配给虚假文本评论。

步骤 7。 我们从原始数据集中选取随机子集，并将其与创建的数据集合并。这样便得到一个平衡数据集，其长度是虚假数据集的两倍，可用于虚假内容分类任务。

### 3.2 人工垃圾信息发送者检测的采样过程

在获得包含真实文本评论及相应用户人口统计特征的数据集后，我们可以修改该数据收集过程，以模拟人工垃圾信息发送者攻击。为了在不诚实者中创建用于虚假内容过滤的数据集，

真实用户应执行算法1所述的以下流程。采样前，应使用多语言双向变换器表示（以下简称 BERT）（区分大小写） [8]，获取数据集中每条文本评论的嵌入。然后，针对每条文本评论，我们计算其平均文本嵌入，大小为 (1，768)。接着，应选择随机观测 $k$，以及其文本评论和属性，并从用户 $i$处获取这些信息。以 2/34 的概率，我们保留文本评论—属性对不变，并将其标记为正对。以 1/3 的概率，我们选择任意用户 $j$（$j \neq i$）与第 $k$条观测进行匹配。然后，在用户 $j$撰写的所有评论中，找出与第 $k$条文本评论最不相似的一条（根据嵌入比较）。为此，我们计算第 $k$条评论的嵌入向量与矩阵 $Rev_j$ 之间的点积；该矩阵由用户 $j$撰写的所有评论的嵌入向量组成。之后，我们选择点积最大的文本评论 $m$，并交换用户属性：将用户 $j$的元数据分配给第 $k$条文本评论，将用户 $i$的元数据分配给第 $m$条文本评论。我们将这些打乱后的配对标记为负对。这样，我们就得到了一个平衡数据集，其中一半用户属性经过打乱，以确保文本评论与用户人口统计特征之间的关联被消除（负对）；另一半数据则保持不变（正对）。接下来，我们应训练基于对比学习的模型，以区分用户属性嵌入与文本嵌入的正对和负对，使负对的嵌入在欧几里得距离上尽可能远，而正对的嵌入尽可能近。当前任务的对比学习设置详见第2章第2.3节。

[4]概率的选择确保了最终数据集平均而言是平衡的。

**算法1**人类垃圾信息发送者检测数据集的采样过程

**输入：** 经过打乱、包含文本评论嵌入 D 的数据集。

**输出：** 包含负样本对和正样本对的平衡数据集。

```
unprocessed_idx ← range(0, len(D))
对 k ← 0 至 len(D) 执行
    如果 k 属于 unprocessed_idx 则
        coin_flip ~ 伯恩(p = 2/3)
        如果 coin_flip = 1 则
            从 unprocessed_idx中移除 k
            将标签 = 0 to k添加到 D 中的第 th 行
        否则如果 coin_flip = 0 则
            user_i ← D.username[k]
            rev_ik ← D.review[k]
            从 D.usernames \ user_i 中采样 user_j
            Rev_j = [D.reviews[D.username == user_j][, 如果 idx 位于 unprocessed_idx]] 中
            max_emb_idx = 取最大值（Rev_j × rev_ik）
            将文本嵌入与索引 k 和 max_emb_idx互换
            从 unprocessed_idx中移除 k
            从 unprocessed_idx中移除 max_emb_idx
            为 D 中第 k行添加标签 = 1
            为 max添加标签 = 1 _emb_D 中第 idx 行
        结束 if
    结束 if
结束 for
```

*3.2.1* 多语言 *BERT*（区分大小写）。为了获取评论文本的嵌入，我们使用多语言 BERT（区分大小写） [8], ，因为该数据集包含以不同语言编写的文本评论。

该 Transformer 模型采用自监督方式，并以掩码语言建模为目标，在包含 104 种语言文本的大型维基百科语料库上进行了训练。BERT 的架构与基于GPT的模型相似，采用堆叠的 Transformer 编码器块，由一系列自注意力层和前馈神经网络组成。这种方法会随机掩盖句子中的一些词语（将它们替换为 MASK 标记），并训练模型预测这些被掩盖的词语，从而确保模型输出双向文本表示。我们使用预训练多语言 BERT（区分大小写）来提取文本评论中每个词元的上下文表示（向量嵌入）。首先，我们对每条评论中的句子进行分词，将其转换为评论中的词语列表，并填充至最大长度 512；如果句子长度超过 512，则进行截断。然后，我们从分词器的输出中获取索引，从而得到每个词元在 BERT 词汇表中的索引，并获取注意力掩码，其中未填充的词元对应 1，填充的词元对应 0。接下来，我们将索引和注意力掩码输入 BERT 模型，以获取文本评论中每个句子的上下文嵌入。之后，我们为每条文本评论计算平均文本嵌入向量，用于对比学习设置。

### 3.3 对比学习设置

我们遵循 [12]中描述的方法。将相似（正）和不相似（负）向量对作为由$W$、 $G_W(X_1)$和 $G_W(X_2)$参数化的 $G_W$ 深度神经网络的输出后，我们将它们之间的点间距离定义为欧几里得距离（“语义相似度”的近似值）：

$$D_W(X_1, X_2) = ||G_W(X_1) - G_W(X_2)||_2 .$$

接下来，我们按如下方式定义对比损失：

$$\mathcal{L}(W) = \sum_{i=1}^{P} L\left((W, (Y, X_1, X_2)^i\right),$$

$$L\left(W, (Y, X_1, X_2)^i\right) = (1 - Y)L_S\left(D_W^i(X_1, X_2)\right) + YL_D\left(D_W^i(X_1, X_2)\right),$$

其中， $X_1$ 和 $X_2$ 是成对的向量， $Y = 0$ 是相似（正）对的目标标签， $Y = 1$ 是不相似（负）对的目标标签，$(Y, X_1, X_2)^i$ 是第 $i$个样本对， $L_S$ 是正对的部分损失函数， $L_D$ 是负对的部分损失函数， $P$ 是样本数。

应使用此类函数 $L_S$和 $L_D$，使得相对于 $W$最小化 $L$后，正对的距离 $D_W$较小，而负对的 $D_W$较大。相应部分损失函数的选择满足这一条件：

$$L_S(W, X_1, X_2) = \frac{1}{2}(D_W(X_1, X_2))^2$$

$$L_D(W, X_1, X_2) = \frac{1}{2}(\max(0, m - D_W(X_1, X_2)))^2,$$

其中， $m > 0$ 是间隔，对应于 $G_W(\cdot)$周围的半径，因此只有位于间隔内的负对才会被计入。通过计算这些函数的梯度可以看出：

$$\frac{\partial L_S}{\partial W} = D_W \frac{\partial D_W}{\partial W}$$

$$\frac{\partial L_D}{\partial W} = \begin{cases} 0, \text{ if } D_W > m, \\ -(m - D_W)\frac{\partial D_W}{\partial W}, \text{ otherwise,} \end{cases}$$

因此，在 $L_S$ 情况下， $D_W$ 的较小值会产生梯度，使 $D_W$ 减小；而由于负号的存在，在 $L_D$ 情况下则相反。

当前方法解决了深度神经网络环境下所定义对比损失的最小化问题，下一小节将对此进行介绍。

*3.3.1* 深度神经网络架构。 我们采用与 [7], [4] 中介绍的孪生网络类似的多层感知机（MLP）架构。该架构包含两个参数集相同的函数分支 $G$，分别用于处理（1）平均文本嵌入向量（$X_1$）和（2）类别特征与数值特征的嵌入（$X_2$），如图 15 所示。两个输入（$X_1$ 和 $X_2$）分别经过相应的分支，得到第一分支的输出 $G^1$（$X_1$）和第二分支的输出 $G^2$（$X_2$）。随后，我们计算两个分支输出之间的$W$ 成对距离，这些输出代表向量；然后使用该距离标量值计算部分损失（$L_S$和 $L_D$）以及总对比损失 $L$。两个分支的输出均表示

[5]基于 https://yashuseth.wordpress.com/2018/07/22/pytorch-neural-network-for-tabular-data-with-categorical-embeddings 中所述的方法。稿件已提交至 ACM

向量，而样本对的标签（$Y$）随后也会与这些向量一同传入对比损失函数。接着，通过 Adam优化算法更新网络参数 $W$，并通过对比损失值的反向传播计算梯度。总梯度是两个分支输出所产生贡献的总和。

我们还采用了权重初始化、丢弃法和批归一化等正则化技巧来应对过拟合。使用生成数据集进行实验的结果见第 3 章。

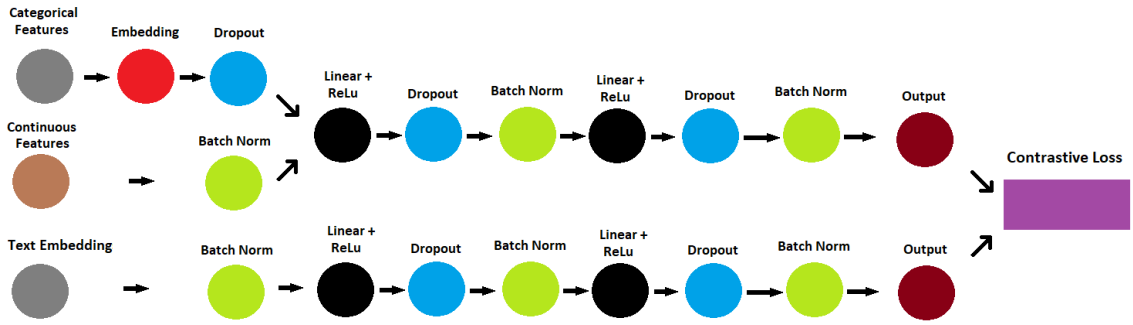


图 1. 多层感知机（MLP）架构，该架构已用于所述的对比损失设置中。

### 3.4 fakeRoBERTa基准

如上所述，推荐系统领域尚未研究当前研究的问题陈述。因此，没有可用于根据质量指标与我们的模型进行比较的有效算法。虚假评论检测领域的最先进方法由 [22] 提出，并创建了 fakeRoBERTa 语言模型，用于将文本集合分类为虚假文本（由上文讨论的 GPT-2 [8] 模型生成）和真实文本。我们将采用这一方法作为基准，并检验利用用户元数据是否有助于检测虚假评论。

作者收集了由 GPT-2 模型生成的虚假评论数据集，随后对 RoBERTa 模型进行微调，以完成指定的分类任务。他们使用不同架构的语言模型，以考虑数据泄漏的可能性，以及分类模型对依赖的生成模型语言风格过拟合的可能性。结果表明，fakeRoBERTa 的准确率达到 96%，在文本分类任务中，其性能得分高于所有参与比较的 OpenAI 模型。此外，作者还开展了众包实验，比较人类与计算机检测虚假评论的能力。他们发现，在生成的数据集上，人类的表现仅略好于随机分类，这表明采用生成式语言模型的攻击已日益先进。然而，作者关注的是检测生成式虚假评论这一问题，却忽略了为获取经济利益而撰写评论的人类灌水者。

本章介绍本研究开展的数值实验。首先，概述研究中使用的语言和程序。接着，详细介绍用于训练和验证所提出的基于对比学习的方法的数据集。之后，介绍用于评估所提方法性能并将其与最先进水平的 RoBERTa 进行比较的质量指标

用于判定文本评论真实性的基准。最后，给出了实验结果以及所提方法的优势。

## 4 数值实验

### 4.1 数据集

*4.1.1* 人工偏见评论过滤数据集。

数据集描述。符合本研究设定的数据集应同时包含用户的人口统计特征和相应的文本评论。目前唯一可用于训练和测试所提出方法的开源数据集是 Google 本地商家评论与元数据[6] [13] [20]。这个来自谷歌的数据集规模庞大，收录了五大洲 310 万家本地商家（涵盖餐馆、商店、学校、酒店、美容院、医院等 4.8 万个类别）的约 1150 万条文本评论和五星评分，来自 450 万名用户。该数据集还包含顾客的人口统计特征，例如用户名和 ID、当前位置的 GPS坐标、曾居住地的坐标、教育和工作经历（地点以及职位/专业），以及评论时间戳。

我们仔细处理了定义的数据集，并为其扩充了若干工程化特征。最终扩充后的 Google 本地商家评论与元数据包含以下特征：

(1) 当前国家（类别型：200 个类别），(2) 曾居住地数量（数值型），(3) 学历（类别型：5 个类别），(4) 曾就读地点数量（数值型），(5) 教育专业（类别型：101 个类别），(6) 当前职业（类别型：101 个类别），(7) 以往工作数量（数值型），(8) 失业状态（类别型：2 个类别），(9) 退休状态（类别型：2 个类别），(10) 本地商家类别（类别型：51 个类别），(11) 文本评论长度（数值型），(12) 评论中的不重复词语数量（数值型）。

我们还针对第 2 章所述的对比学习设置执行了抽样流程，并添加了二元标签目标变量：当文本与属性构成正对时，取值为 0；否则取值为 1。这样，数据中一半被标记为正对，表示用户元数据与评论文本之间存在潜在关联；另一半被标记为负对，其中用户属性经过打乱，并以非随机方式分配给其他用户的文本评论。由此得到的数据集体现了人类垃圾信息发送者实施的攻击，他们撰写有偏见的文本评论。

*4.1.2* 生成式语言模型文本检测数据集。我们还整理了一个新的由语言模型生成的虚假评论数据集。我们依据第2章所述流程，基于谷歌本地商家评论和元数据文本，为 GPT-J-6B模型构建了问答提示 [25]。根据 [1], 的前提，我们假设恶意用户无法获取用户属性的分布，因此会随机填写自己的人口统计信息；因此，如果使用虚假个人资料，用户属性与文本评论之间的关联就会被破坏。此后，为了模拟虚假

[6] https://cseweb.ucsd.edu/ jmcauley/datasets.html#google_本地版本。稿件已提交至 ACM

个人资料攻击，我们为每条虚假评论文本随机抽取用户属性，并由此获得了一个数据集，用于模拟生成式语言模型利用虚假随机用户元数据发起的攻击。

### 4.2 评估指标

在当前研究中，我们打算训练第2章中定义的多层感知机（MLP）：若输入对为正对，即用户人口统计特征与所写文本相匹配，则缩短两个分支输出之间的距离；否则，则增大距离。为此，我们求解对比损失最小化的优化问题。

由于生成的两个数据集均为平衡数据集，因此我们可以使用准确率作为质量指标。训练好模型后，该模型会拉开负对的属性嵌入与平均文本嵌入之间的距离，并使正对的这些嵌入彼此靠近。在评估期间，我们可以将模型输出向量之间的成对距离与阈值进行比较。在当前实验中，我们将阈值设为 0。5。

fakeRoBERTa基准也使用准确率作为质量指标。它还会最小化负似然损失，其定义为

$$\log \mathbb{P}(\mathcal{D}|\theta) = \sum_{i=1}^{n} \left( y_i \log \hat{y}_{\theta,i} + (1 - y_i) \log \left(1 - \hat{y}_{\theta,i}\right)\right),$$

其中 D 为观测到的数据，$\theta$表示待优化的参数，而第 $i$个观测值被标记为正类的概率为 $\hat{y}_{\theta,i}$。

### 4.3 实验结果及与现有算法的比较

我们在两种分类设置下比较了所提出算法的性能：

(1) 区分用户元数据保持不变（即用户人口统计特征与文本评论嵌入之间的关联保持原样）的真实人类所写的文本评论，以及用户元数据经过打乱（即所定义的关联遭到破坏）的真实人类所写的偏颇评论；(2) 区分真实诚实人类所写的文本评论与生成式语言模型所写的文本评论。

诚实用户与人类垃圾评论者。我们按照第2章中定义的流程，在一个生成的数据集（基于 Google 本地商家评论与元数据，包含 496,320 个观测值）上采用了所提出的对比学习设置：一半数据保持原样（正对），另一半数据经过打乱（负对）。我们将数据按 80-20% 划分为训练集和验证集。所提出的架构训练了 30 个训练轮次。结果见表 1，训练和验证指标的变化情况见图 2。

准确率指标并不算很高，训练表现也不够稳定，但它们为后续算法的开发奠定了良好的基线。这些数值也提出了一个值得研究的问题：人们撰写的文本与其人口统计属性之间是否存在紧密联系？这种联系可能过于微妙，机器模型难以捕捉。

诚实用户与语言模型。我们还将所提出的方法用于另一个生成的数据集，其中包含从原始谷歌数据集中随机选取的 $14k$ 条真实文本评论及其用户属性（正对），以及 $14k$ 条人工生成的文本评论，并随机分配了用户元数据。我们还将数据分别按 80-20% 的比例划分为训练集和验证集，并对模型进行了 30 个训练轮次。训练集和验证集的对比损失值与准确率指标见表 1 和图 3。

表 1. 所提出方法和 fakeRoBERTA 基准在诚实人类与人类灌水者、诚实人类与语言模型分类任务的训练集和验证集上的损失和准确率。

| 设置 | 人类与人类（对比-学习） | 人类 vs 语言模型（对比学习-ing) | 人类 vs 语言模型（fakeRoBERTA） |
|---|---|---|---|
| 训练对比损失 | 1213.82 | 22393.25 | - |
| 验证对比损失 | 1112.41 | 22466.42 | - |
| 训练负对数似然损失 | - | - | 1.01 |
| 验证负对数似然损失 | - | - | 0.96 |
| 训练准确率 | 0.73 | 0.62 | 0.49 |
| 验证准确率 | 0.66 | 0.60 | 0.5 |

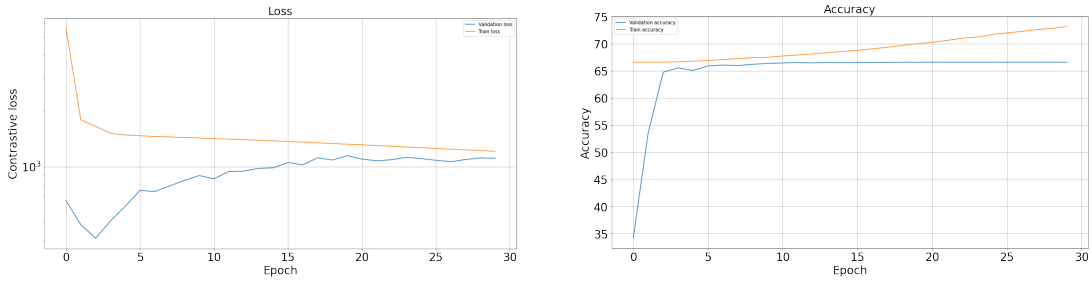


图 2. 诚实人类与人类灌水者分类任务中，训练集（橙色）和验证集（蓝色）的对比损失与准确率指标曲线图。

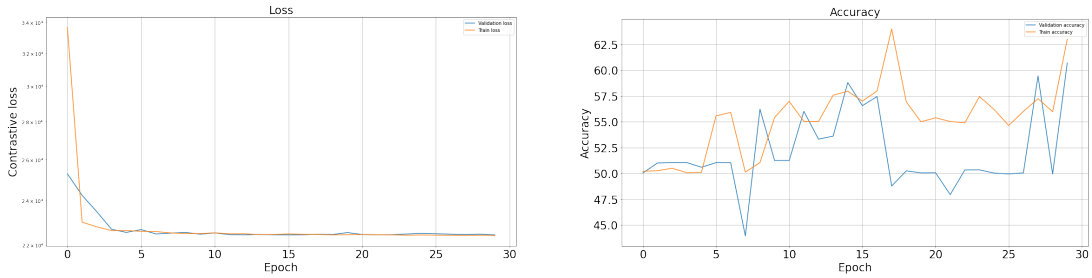


图 3. 诚实人类与语言模型分类任务的训练集（橙色）和验证集（蓝色）的对比损失与准确率指标图。

可以观察到，与人类对人类的设置相比，这两项指标都更差。这或许可以解释为：语言模型是在真实评论上训练的，因此可能生成了过拟合的文本评论；此外，训练数据集规模也不够大（仅有 $28k$ 个观测值）。

**与 *RoBERTa* 的比较。**如前所述，本研究提出的方法具有创新性，目前尚无有效的基准可供比较。尽管如此，我们仍能将所提出的算法与虚假评论检测领域的最先进方法 fakeRoBERTa 进行比较。该方法经过训练，可区分真实评论与 GPT-2 模型生成的评论。我们遵循作者提供的训练流程，论文已提交至 ACM

并对 RoBERTa 进行了微调，使其能够将真实评论与 GPT-J-6B 生成的评论区分开来，训练了 5 轮。准确率和负似然损失值如表 1 所示。

从训练结果可以看出，微调后的 fakeRoBERTa 表现并不比随机猜测更好，因此生成的评论与真实评论非常相似。与所提出的方法相比，这也表明用户人口统计属性确实有助于区分真实文本评论和生成的文本评论。

所提方法的优势。实验结果表明，在虚假评论检测的准确率指标方面，所提方法显著优于最先进水平的 fakeRoBERTa。此外，所研究的算法在两项分类任务中均表现良好，因此能够有效过滤人类和语言模型生成的虚假内容。尽管如此，仍需开展更深入的语言学和心理学调查，以确定所提算法的指标表现究竟是源于神经网络架构不合适，还是源于用户属性与书面文本之间关系的复杂性。尽管如此，当前方法为虚假文本评论检测开辟了新途径，并可作为未来算法的良好基线，用于检测人类撰写的偏颇评论和生成式语言文本。

## 5 结论

本研究提出了一种新颖的方法，用于检测基于协同过滤的推荐系统中的虚假文本评论。所提出的方法将用户元数据与文本评论结合起来，作为识别虚假评论的额外依据。基于对比学习设置，该方法使推荐平台能够抵御两类攻击策略：（1）人类为获取经济利益而撰写偏颇评论；（2）由语言模型生成的评论。实验结果表明，所提出的算法在这两项分类任务中表现良好，优于仅处理文本的最先进水平方法 fakeRoBERTa [22]。此外，我们还通过构建两个数据集作出了重要贡献，分别模拟了上述两类攻击。第一个数据集用于诚实用户与人工垃圾评论者分类任务，基于 Google 本地商家评论与元数据数据集构建，并通过设计用户人口统计特征和采用所提出的抽样流程进行了进一步修改和预处理。第二个数据集由 GPT-J-6B [25] 模型生成的虚假评论组成，并配以随机用户元数据，可用于诚实人类与生成式语言模型分类任务。总体而言，本研究提出了一种更通用的虚假评论检测方法，为未来研究奠定了良好的基线。

当前项目面临若干研究局限。首先，问题陈述的性质意味着需要使用同时包含评论文本和用户人口统计特征的数据集。现有的开源推荐系统数据集要么包含文本评论信息，要么包含用户元数据，但不会同时包含两者。唯一满足这一条件的开源数据集是 Google 本地商家评论与元数据，我们已在该数据集上进行了数值实验。然而，我们获得的实验结果仅适用于该数据集，未必能推广到其他推荐系统平台。因此，需要获取同时包含评论文本和用户属性的私有工业数据集，以便验证我们的方法。

此外，有研究指出，这一问题陈述对于推荐系统场景而言是全新的，尚未得到研究。因此，目前没有可用于比较我们所提出方法性能的有效算法。今后有必要进一步探究这一问题，以开发更强大的算法，并将本文提出的方法作为基准。

另一项研究局限在于评论文本与用户人口统计数据之间的联系。我们假设真实用户的元数据与评论文本存在潜在关联，并且深度神经网络能够捕捉这种关联。所使用的深度神经网络架构可能并不十分适合执行此类任务，因而无法捕捉这种潜在关联；也可能是关联本身较弱，任何其他模型都无法捕捉。应通过开展心理学和语言学实验，更具体地研究这一局限，并检验用户的人口统计特征与其所写文本评论之间联系的强度。此外，还可以在上文所述的其他类似数据集上检验这些方法的性能，以了解模型在不同数据集上捕捉模式的能力。

**致谢**

作者感谢 Anton Voronov 就该问题进行了富有成效的讨论，并指出在当前情境下使用对比学习的可行性。

**参考文献**

[1] Mehmet Aktukmak、Yasin Yilmaz 和 Ismail Uysal。2019。基于用户属性的推荐系统中快速准确的攻击检测。计算机协会 (2019)，348–352。[2] Jasem M. Alostad。2019。使用支持向量机高斯混合模型改进推荐系统中的刷单攻击检测。《信息与知识管理杂志》18 (2019)。第 8 期。[3] Alper Bilge、Zeynep Özdemir 和 Huseyin Polat。2014。一种新型刷单攻击检测方法。《计算机科学程序》31 (2014)，165–174。[4] Jane Bromley、Isabelle Guyon、Yann LeCun、Eduard Säckinger 和 Roopak Shah。1993。使用“孪生”时延神经网络进行签名验证。载于第六届神经信息处理系统国际会议论文集(科罗拉多州丹佛) *(NIPS'93)*。Morgan Kaufmann 出版社，737–744。[5] Tom Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared D Kaplan、Prafulla Dhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda Askell 等。2020。语言模型是少样本学习器。《神经信息处理系统进展》33 (2020)，1877–1901。[6] Paweł Budzianowski 和 Ivan Vulić。2019。你好，这是 GPT-2——我能为你做什么？探索预训练语言模型在面向任务的对话系统中的应用。*arXiv* 预印本 *arXiv:1907.05774* (2019)。[7] Sumit Chopra、Raia Hadsell 和 Yann LeCun。2005。以判别方式学习相似度度量，并将其应用于人脸验证。载于*2005*年 *IEEE* 计算机学会计算机视觉与模式识别会议论文集，*CVPR 2005*（*2005*年 *IEEE* 计算机学会计算机视觉与模式识别会议论文集，*CVPR 2005*）。IEEE 计算机学会，539–546。https://doi.org/10.1109/CVPR.2005.202 2005年 IEEE 计算机学会计算机视觉与模式识别会议，CVPR 2005；会议日期：2005年6月20日 至 2005年6月25日。[8] Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。2018。BERT：用于语言理解的深度双向变换器预训练。*CoRR* abs/1810.04805 (2018)。arXiv:1810.04805 http://arxiv.org/abs/1810.04805[9] Yingtong Dou、Guixiang Ma、Philip S. Yu 和 Sihong Xie。2020。通过纳什强化学习实现鲁棒的垃圾信息发送者检测*(KDD '20)*。计算机协会，924–933。[10] Luciano Floridi 和 Massimo Chitiatti。2020。GPT-3：其本质、范围、局限与影响。《心智与机器》1-14 (2020)。[11] Leo Gao、Stella Biderman、Sid Black、Laurence Golding、Travis Hoppe、Charles Foster、Jason Phang、Horace He、Anish Thite、Noa Nabeshima、Shawn Presser 和 Connor Leahy。2021。The Pile：一个用于语言建模的 800GB 多样化文本数据集。*CoRR* abs/2101.00027 (2021)。arXiv:2101.00027 https://arxiv.org/abs/2101.00027[12] R. Hadsell、S. Chopra 和 Y. LeCun。2006。通过学习不变映射进行降维。载于*2006*年 *IEEE* 计算机学会计算机视觉与模式识别会议（*CVPR' 06*），第 2 卷。1735–1742。https://doi.org/10.1109/CVPR.2006.100[13] Ruining He、Wang-Cheng Kang 和 Julian McAuley。2017。基于翻译的推荐。*RecSys'17* (2017)。[14] Naima Iltaf、Abdul Ghafoor 和 Uzman Zia。2013。一种在间接信任计算中检测不诚实推荐的机制。《*EURASIP* 无线通信与网络期刊》2013 (2013)。[15] Jong-Seok Lee 和 Dan Zhu。2012。刷单攻击检测——实现可信推荐系统的新方法。《*INFORMS* 计算期刊》24 (2012)。[16] Cong Li 和 Zhigang Luo。2011。协同过滤推荐系统中的刷单攻击检测。*2011*年国际软计算与模式识别会议（*SoCPaR*）(2011)，190–193。[17] Chen Lin、Si Chen、Hui Li、Yanghua Xiao、Lianyun Li 和 Qian Yang。2020。利用增强型用户画像攻击推荐系统。*CoRR*abs/2005.08164 (2020)。[18] Yinhan Liu、Myle Ott、Naman Goyal、Jingfei Du、Mandar Joshi、Danqi Chen、Omer Levy、Mike Lewis、Luke Zettlemoyer 和 Veselin Stoyanov。2019。RoBERTa：稳健优化的 BERT 预训练方法。*CoRR* abs/1907.11692 (2019)。arXiv:1907.11692 http://arxiv.org/abs/1907.11692 稿件已提交至 ACM

[19] Michael Luca。2016。评论、声誉与收入：Yelp.com案例。哈佛商学院 *NOM* 单位工作论文 12-016 (2016)。[20] Rajiv Pasricha 和 Julian McAuley。2018。用于序列推荐的基于翻译的分解机。*RecSys'18* (2018)。[21] Alec Radford、Jeffrey Wu、Rewon Child、David Luan、Dario Amodei、Ilya Sutskever 等。2019。语言模型是无监督多任务学习器。*OpenAI* 博客 1, 8 (2019)，9。[22] Joni Salminen、Chandrashekhar Kandpal、Ahmed Mohamed Kamel、Soon-gyo Jung 和 Bernard J. Jansen。2022。在线产品虚假评论的创建与检测。《零售与消费者服务期刊》64 (2022)。[23] Noopur Samaiya、Sandeep Raghuwanshi 和 Raghuwanshi R. K. Pateriya。2019。使用主成分分析和支持向量机检测推荐系统中的刷单攻击。(2019)。[24] Jiaxi Tang、Hongyi Wen 和 Ke Wang。2020。重新审视针对推荐系统的对抗式学习注入攻击。*CoRR* abs/2008.04876 (2020)。[25] Ben Wang 和 Aran Komatsuzaki。2021。GPT-J-6B：一个拥有 60 亿参数的自回归语言模型。[26] Chenwang Wu、Defu Lian、Yong Ge、Zhihao Zhu、Enhong Chen 和 Senchao Yuan。2021。以火攻火：通过对抗投毒训练实现鲁棒推荐系统。计算机协会，1074–1083。[27] Qin Yang。2018。一种基于攻击检测的鲁棒推荐系统。《并发与计算：实践与经验》32 (2018)。[28] Wei Zhou、Junhao Wen、Qiang Qu、Jun Zeng 和 Tian Cheng。2018。基于群体用户可信度和评分时间序列的推荐系统刷单攻击检测。《公共科学图书馆·综合》32 (2018)。