基于信息熵和 GBDT 算法的 AI 生成与人类撰写检测研究

戎 蓉[1] 杨 行[1] 韩 叙[1] 胡 仕[1]
RONG Rong YANG Hang HAN Xu HU Shi

摘 要 针对人工智能生成文本智能识别与检测的问题，研究基于词频计算的信息熵和机器学习相结合的方法，建立了基于信息熵和 GBDT 算法（梯度提升决策树）预测模型。通过对大量文本样本进行特征提取和模型训练，构建分类检测模型。为了检验梯度提升决策树模型用来判断是否能由 AI 生成文本的精确度，进行相似度的计算，基于 ROC 曲线方法论对其进行验证。所提出模型的预测准确率较高，为 AI 生成文本的质量评估提供了新的视角，也为未来 AI 文本生成技术的发展提供了重要参考。此外，有助于保护知识产权，防止 AI 生成文本被误用或滥用。

关键词 人工智能；信息熵；机器学习；决策树；GBDT 算法

doi：10.3969/j.issn.1672-9528.2024.07.040

0 引言

近年来，随着信息技术的不断发展，各类生成式人工智能得到突破性的进展，使得人工智能生成的内容（AI generated content，AIGC）得到发展。随着 ChatGPT 的发布，生成式人工智能技术得到广泛关注，由 ChatGPT 领导的大型语言模型在世界范围内得到了普及，并得到了广泛的推广和使用。自然语言大模型呈现出强大的自然语言处理、整合、荟萃和生成能力，利用本身强大的自然语言处理能力[1-2]生成自然流畅的文本文体。同时，以生成式对抗网络（generative adversarial networks，GAN）模型、Diffusion 扩散化模型为代表的生成式人工智能在图像、音频、视频领域表现出优异的内容合成、修复、预测和生成性能等[3]。人们充分认识到这些模型给人们带来的丰富、智能和方便的体验。同样重要的是，也意识到人工智能文本生成等工具相关的许多风险。对于许多要求原创性的工作，越来越多的人采用人工智能生成内容，人工检测方法很难判断内容是否为人类撰写，这给原创性工作带来很多伦理风险和安全隐患。如何识别 AI 生成文本与人类撰写成为当前人工智能研究的热点，也是防范人工智能文本生成风险的重要手段。

1 相关研究

针对 AI 生成文本的特点，国内外研究人员对生产文本与人类撰写识别进行了深入研究[4-5]。现有的生成文本检测方法大多采用融合多种特征的学习算法来检测生成文本，包括文本内容、文本结构、生成技术等方面来发现生成文本。目前主流的文本检测方法有基于机器学习的检测方法、基于人机交互的检测方法和基于概率规则的检测方法[6-7]。

（1）基于机器学习的检测方法。基于机器学习的生成文本检测方法通常涉及训练一个模型，使其能够根据文本的统计特性、语言结构或语义特征来区分机器生成的文本和人类生成的文本。常见的方法有深度学习、语义分析、统计模型等。Solaiman 等人[8]提出了一种基于 TF-IDF 特征的逻辑回归分类器，在较小 TGM 基于 Top-K 采样策略的生成文本上具有较好的识别能力。Bakhtin 等人[9]提出基于 EBMs 模型进行机器学习训练，区别人类生成与机器生成文本。这些方法的准确性和有效性取决于多种因素，包括训练数据的数量和质量、模型的复杂性和泛化能力以及检测任务的特定要求等。

（2）基于人机交互的检测方法。通过利用人类的视觉解读技能、常识知识和计算机的统计速度，可以建立一个系统来识别机器生成文本。常见的方法有语义分析、专家评估等方法。Ippolito 等人[10]研究了人类和自动检测器对 TGM 生成的文本识别能力的差异。曹娟等人[11]提出基于事实信息的取证研究，能够在提升检测性能的同时，提供更好的可解释性、可展示性。基于人际交互的检测方法虽然不如机器学习方法自动化，但它可以作为其他方法的补充。

（3）基于概率规则的检测方法。基于概率的文本生成检测方法通常依赖于统计模型来预测下一个单词或字符的可能性。Gehrmann 等人[12]设计了一个统计工具 GLTR，可以突出显示生成文本和人工文本的分布差异。基于概率规则的检测方法，还可以考虑结合其他技术如自然语言处理（NLP）和深度学习来提高检测效果。

1. 昭通学院 云南昭通 657000

[基金项目] 云南省教育厅科学研究基金项目（2024J1064）；云南省高等学校计算机教学研究会项目（云高计教 202305）

2 方法提出

借鉴国内外研究成果，本文提出基于信息熵和 GBDT 算法的 AI 生成与人类撰写检测方法。结合信息熵和 GBDT 算法，计算待检测文本的信息熵，并将其作为文本的一个特征，将其他文本特征（如词汇频率、词语丰富度等）与信息熵特征一起输入 GBDT 分类器中。通过训练和优化模型，用于区分 AI 生成和人类撰写的文本。

信息熵是信息论中的一个概念，用于衡量一个随机变量的不确定性。在文本分析中，信息熵可以用来衡量文本内容的复杂度或不确定性。对于 AI 生成和人类撰写的文本，它们的词汇分布和句子结构可能存在差异，从而导致它们的信息熵有所不同。通过计算文本的信息熵，可以获得一个量化指标，用于区分这两种文本。

GBDT 即梯度提升决策树，是一种集成学习方法，它通过构建多个决策树并将它们的预测结果进行集成来提高预测精度，通过收集大量 AI 生成的文本和人类撰写的文本作为训练数据，将训练好的 GBDT 模型部署到实际应用中，用于检测新生成的文本是否由 AI 生成。

3 研究假设

对于此次研究问题的探索与分析，做出以下基本假设。（1）假设 AI 生成的文本文体词汇多样，情感与句式分析上与人类存在差异。（2）假设 AI 生成的文本文体与人类的写作在统计分析上有明显的区分。（3）假设这些差异的明显区分可以通过人工自然语言处理进行特征量化标记。（4）假设可以使用与访问到足够相关的信息资料进行原始数据的验证。（5）假设提取的部分文本有着个人的主观认识，AI 的情感相对局限。

4 模型建立

4.1 文本特征指标

根据互联网上寻找到的 20 篇文章信息，提取其中的文段，并使用 AI 进行重新改写，最后将得到的 64 篇文本段落实现的分类作为训练集。为了反映 AI 写作文本与人类写作文本之间的差异性，构建并选取了 12 个指标。通过分词、统计词频、计算句子长度等方式对文本数据进行处理和分析，从而提取出句子平均长度、句子中的平均词数、句子最长长度、信息熵、词变异指数、词密度等有用的信息和指标。详细介绍 7 个指标，其余指标为词类别指标。

（1）词频

AI 智能写作是一种利用机器学习算法和自然语言处理技术，通过分析大量数据和语言模型来自动生成文章的技术。然而，其只能按照预设的规则进行操作，难以表达丰富的情感和人类的创造力。因此，不同词性的词频可以一定程度上反映文本是否由 AI 生成。在代称方面中，AI 或许比人类用得更多，它会频繁出现人称来保证它的句式结构等，而我们则会更注重其他来减少句子的冗杂性。基于此，人称词频越大，AI 写作的可能性越大。相反，对于连词词频、形容词词频、副词词频和动词词频，AI 创作的情感不足，因此认为这四种词频越小，AI 写作的可能性越大。词频计算公式为：

$$\text{词频}=\frac{\text{文本中出现的次数}}{\text{文本中总词数}}\times 100\% \tag{1}$$

（2）词类别密度

词类别密度是指在一个文本中，不同的词类别（如名词、动词、形容词等）的数量占总词数的比例。人类写作时，情感是细化且复杂的，而相对于人类来说，AI 的情感线就比较单一，在与人类表达同一个文段文本的时候，文词的混乱就比较小。具体的计算公式为：

$$\text{词类别密度}=\frac{\text{不同词类别的数量}}{\text{文本中的总词数}}\times 100\% \tag{2}$$

词类别密度越高表示第 i 个文段的词语利用率越高，很大程度可能是 AI 创作。词类别密度是一种衡量文本中词类别多样性的指标。它可以在文本中实现不同词类别的分布，反映语言的多样性和表达方式。

（3）信息熵

信息熵用以衡量一个随机变量的不确定性或信息量的大小。当所有可能的结果都是相等的概率时，信息熵达到了最大值，表示不确定性此时最高；而当某些结果的概率高于其他结果时，信息熵较低，表示不确定性较小。然而对于人类的写作来说，AI 涉及的词汇、结构、情感等都是比较单调的，所以结果的概率高于其他结果时，信息熵较低，表示不确定性较小。计算公式为：

$$H(X)=-\sum_{i=1}^{n}P_{ij}\times\log_2\left(P_{ij}\right) \tag{3}$$

$H(X)$ 表示的是随机变量 X 的信息熵。n 表示随机变量 X 可能的取值总数。P 表示随机变量 X 中第 i 段第 j 个词语的时使用频率。n 表示词语的种类数。$H(X)$ 越大表示第 i 个文段的信息熵越大，词频的混乱程度越高，越可能为人类创作。

（4）词汇丰富度

用来衡量文本文段中特征词的数量与总词汇数间的占比。常用的指标是基于特征量化词与总的词汇的类型标记比例。词汇丰富度的计算式为：

$$\text{词汇丰富度}=\frac{\text{不同词汇的数量(类型数)}}{\text{总词汇量(标记数)}} \tag{4}$$

（5）句子平均词数

人与 AI 写作时，存在一定的差异。文本的写作人与 AI

的想象丰富性则不同，人具有一定的联想能力，在进行某一方面的写作时，会引入其他方面的内容，而 AI 是固定的数据库，只会根据提供的指令生成文本，文本句子平均词数较少。对每个句子进行词汇计数分析，对每个句子包含的词汇数量进行计算。句子平均次数计算公式为：

$$句子平均次数=\frac{句子中的词汇总数}{句子的总数} \tag{5}$$

若是句子平均次数越大，表示想象的文本丰富性越强，即可认为文本是人类所作，不是 AI 写作。

（6）最大句子长度

最大句长度是指在一个文本或语料库中，最长的句子所包含的词汇数量。由于人在写作文段的时候，文本文段的结构复杂度相对较高，句长就比较长。计算最大句长度的公式可以表示为：

$$最大句子长度=Max(每个句子的词汇数量) \tag{6}$$

最大句长度越大表示第 i 个文段中最长句子的字数越多，越可能为人创作。

（7）变异指数

变异指数通常用于衡量一组数据的变异程度或离散程度，使用标准差来计算。AI 写作相对于人类风格更加单一，因此在句子字数相近时，人在表达意思时会比 AI 引入更多的词语。由于词语的种类有很多，所以人类写的文段中词频数的均值就较低，标准差会更高，进而变异指数更大。具体的计算公式为：

$$变异指数=\frac{|标准差|}{|平均值|}\times 100\% \tag{7}$$

变异指数越大，表示数据集的变异程度越高，数据点之间的差异性越大，而变异指数较小，则表示数据点之间的差异性较小。变异指数越大，表示第 i 个文段的离散程度越大，越可能为人创作。

4.2 GBDT 分类模型

GBDT 分类模型是一种强大的机器学习算法，通过串行训练多个决策树模型，并利用梯度提升的方法不断优化模型性能。在 GBDT 分类模型中，每棵决策树都是基于之前模型的残差进行训练，通过不断迭代来达到降低损失函数和提升模型预测的准确性的目标。GBDT 分类算法公式为：

$$F(x)=\sum_{m=1}^{M} v h_m(x) \tag{8}$$

式中：$F(x)$ 表示模型对输入特征 x 的预测值，是所有基本分类树的预测值；v 为学习率，$h_m(x)$ 表示第 m 个分类树模型。总损失函数公式为：

$$L=\sum_{i=1}^{N} L(y_i, F(x_i)) \tag{9}$$

式中：y_i 表示真实标签，$F(x_i)$ 表示模拟的预测值。

5 实验和结果

5.1 实验数据分析

相关系数的绝对值越接近 1，则表示两个变量之间的线性关系越强，反之，越接近 0，则表示两个变量之间的线性关系越弱。Pearson 相关性分析的计算公式为：

$$r=\frac{\mathrm{cov}(X,Y)}{\sigma X\times \sigma Y} \tag{10}$$

利用 SPSS 软件进行数据的相关性分析，结果如图 1 所示。

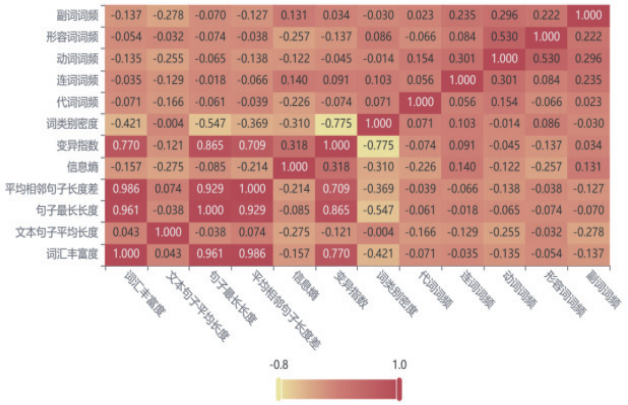

图 1 特征词之间的相关系数热力图

由图 1 可知，词汇的丰富度与平均相邻句子长度差、句子最长长度拥有强正相关性，变异指数与词类别密度有强负相关性。句子最长长度与文本句子平均长度有弱负相关性，副词词频和变异指数有弱正相关性。经过对数据的处理，划分出来的特征词所拥有的重要程度如表 2 所示。

表 2 特征词重要程度数值表 /%

特征词	代词词频	动词词频	变异指数	连词词频	副词词频	句子最长长度	信息熵	相邻句子长度差	形容词词频	此类别密度	文本句子平均长度	词汇丰富度
重要程度	41.2	16.5	12.1	10.4	6.4	4.3	3.4	2.0	1.4	1.2	0.8	0.3

根据表 2 数据可看出代词词频的重要性值为 41.2，说明代词词频是最重要的特征之一，动词词频和变异指数分别排在第二和第三位，重要性值分别为 16.5% 和 12.1%，其他特征的重要性值相对较小，对目标变量的贡献程度较低。词汇丰富度的贡献值为 0.3%，贡献率几乎没有影响，但由于文本的组成离不开所选特征的构成，所以保留特征。

5.2 预测结果

在收集好数据指标后，将预处理好的数据按照 7:3 的比

例划分训练集和测试集，依据中文和英文的不同，对其分别进行模型训练，中文与英文的训练数据如表 3 所示。

表 3 中文和英文的指标结果

	准确率	召回率	AUC	F_1
中文	0.943 182	0.904 762	0.979 389	0.883 721
英文	0.909 091	0.913 043	0.982 609	0.933 33

从表 3 的数据中可以看出，中文与英文数据集的准确率、召回率、AUC 和 F_1 值都较好，表明该模型在分类任务中的表现较为稳定和可靠，具有较高的预测准确性和分类能力。两者的 ROC 曲线如图 2 和图 3 所示。

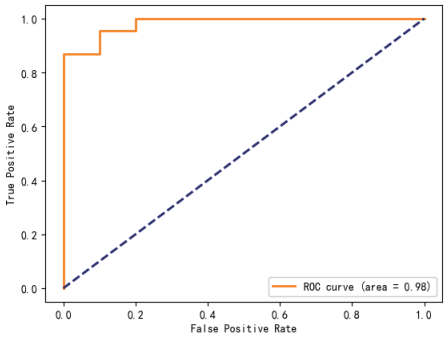

图 2 英文分类 ROC 曲线图

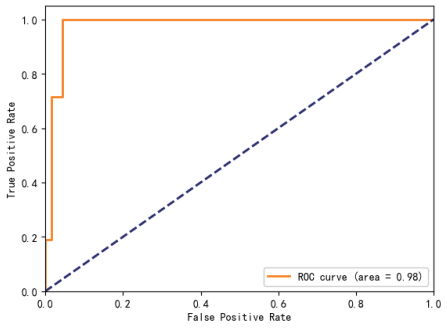

图 3 中文分类 ROC 曲线图

5.3 模型优缺点

（1）优点：高准确性。GBDT 在训练集和测试集上都能表现出良好的性能，能够处理高维度、稀疏特征以及非线性关系等复杂问题。它具有强大的泛化能力，通过组合多个弱分类器形成一个强分类器，减少过拟合风险。其适用于混合类型数据，可以同时处理二元特征与连续特征。

（2）缺点：训练时间较长。由于其串行算法特性，每棵决策树的构建都需要按顺序进行，导致训练时间较长。需要手动调节参数，如树的数量、学习率等，需要通过交叉验证等方法进行调优。

6 结语

本文基于信息熵和 GBDT 算法对 AI 生成文本与人类撰写文本进行了深入检测研究，取得了较好的检测效果。通过信息熵分析文本的信息复杂度，结合 GBDT 算法的高精度分类能力，构建了一个高效的检测模型，有效区分了 AI 生成文本与人类撰写文本。未来将通过不断优化算法和模型，进一步提高检测精度和效率，为文本内容的真实性和可信度提供有力保障。

参考文献：

[1] 朱禹，陈关泽，陆泳溶，等．生成式人工智能治理行动框架：基于 AIGC 事故报道文本的内容分析 [J]. 图书情报知识，2023, 40(4):41-51.

[2] 张钹，朱军，苏航．迈向第三代人工智能 [J]. 中国科学（信息科学），2020,50(9):1281-1302.

[3]BORJI A.Pros and cons of GAN evaluation measures[J]. Computer vision and image understanding,2019,179:41-65.

[4] 王一博，郭鑫，刘智锋，等 .AI 生成与学者撰写中文论文摘要的检测与差异性比较研究 [J]. 情报杂志 ,2023,42(9):127-134.

[5] 张华平，李林翰，李春锦 .ChatGPT 中文性能测评与风险应对 [J]. 数据分析与知识发现 ,2023,7(3):16-25.

[6] 董腾飞，杨频，徐宇，等．基于事实和语义一致性的生成文本检测 [J]. 四川大学学报（自然科学版），2023,60(4):59-66.

[7] 郭美城．基于逐层相关性传播的机器生成文本检测研究 [D]. 内蒙古：内蒙古工业大学 ,2021.

[8]SOLAIMAN I, BRUNDAGE M, CLARK J, et al. Release strategies and the social impacts of language models[EB/OL]. (2019-08-24)[2024-03-02].https://arxiv.org/abs/1908.09203.

[9]BAKHTIN A, GROSS S, OTT M, et al. Real or fake? learning to discriminate machine from human generated text[EB/OL]. (2019-06-07)[2024-03-02].https://arxiv.org/abs/1906.03351.

[10]IPPOLITO D, DUCKWORTH D, CALLISON-BURCH C, et al. Automatic detection of generated text is easiest when humans are fooled[EB/OL].(2019-11-02)[2024-03-02]. https://arxiv.org/abs/1911.00650.

[11] 曹娟，朱勇椿，亓鹏，等．数字内容生成、检测与取证技术综述 [J]. 大数据 ,2023,9(5):150-173.

[12]GEHRMANN S, STROBELT H, RUSH A M. GLTR: statistical detection and visualization of generated text[EB/OL].(2019-06-10)[2024-03-04].https://arxiv.org/abs/1906.04043.

【作者简介】

戎蓉（1989—），女，云南大理人，硕士，讲师，研究方向：信息管理与应用。

（收稿日期：2024-05-15）