基于 Shapley 加性解释的 ChatGPT 生成文本检测模型研究

刘 冬[1] 陈一民[2*]

[1](上海公安学院 上海 200137)

[2](上海建桥学院 上海 201306)

摘 要 针对如何快速识别文本内容是否为 ChatGPT 所生成的问题,提出一种基于 BERT-BiGRU 的 AI 生成文本检测模型。该模型使用预训练的 BERT(Bidirectional Encoder Representations from Transformers)抽取文本的语义特征,并使用 BiGRU(Bidirectional Gated Recurrent Unit)进行特征综合提炼;将 BERT-BiGRU 分类模型在 AI 生成检测数据集 HC3(Human ChatGPT Comparison Corpus)上的分类性能进行相关模型评估;引入 Shapley 加性解释工具(SHAP)从全局和局部两个维度对不同模型所识别出的关键特征和基准值进行比较分析。实验结果表明,虽然深度学习和预训练 BERT 分类模型均取得了较好的分类性能,但在未学习过语种数据集上性能下降严重,然而 BERT-BiGRU 模型表现优秀;不同模型使用可解释工具在同一数据集上计算得到的关键词差异较大,且关键词多为数字、生僻字和标点符号,模型并未真正理解人类撰写文本与 AI 生成文本的内在特征区别,基于已有封闭数据集训练得到的模型无法真正应对开放式的实际应用场景。

关键词 ChatGPT SHAP BERT BiGRU HC3 AI 生成文本检测

中图分类号 TP3 **文献标志码** A DOI:10.3969/j.issn.1000-386x.2024.10.032

CHATGPT GENERATED TEXT DETECTION MODEL BASED ON SHAPLEY ADDITIVE EXPLANATIONS

Liu Dong[1] Chen Yimin[2*]

[1](*Shanghai Police College*, *Shanghai 200137*, *China*)

[2](*Shanghai Jian Qiao University*, *Shanghai 201306*, *China*)

Abstract In order to quickly identify whether text content is generated by ChatGPT, this paper proposes an AI generated text detection model based on BERT-BiGRU. We used pre-trained BERT (bidirectional encoder representations from transformers) to extract semantic features of the text, and used BiGRU (bidirectional gated recurrent unit) for comprehensively extracted feature. The classification performance of the BERT-BiGRU classification model on the AI generated detection dataset HC3 (human ChatGPT comparison corpus) was evaluated. The shapley additive exPlanations (SHAP) was introduced to compare and analyze the keywords and benchmark values identified by different models from both global and local dimensions. Experimental results show that although both deep learning and pre-trained BERT classification models have achieved good classification accuracy, their performance has seriously declined on unlearned datasets, BERT-BiGRU model still has high accuracy. These models' keywords which are calculated on same dataset are quite different, and most of the keywords are numbers, rare characters, and punctuation. These models don't truly understand the inherent characteristics of real human-written text and AI generated texts, models trained on existing closed datasets cannot truly cope with open practical application scenarios.

Keywords ChatGPT SHAP BERT BiGRU HC3 AI generated text detection

---

收稿日期:2024-03-29。上海公安学院科研项目(23xkx53)。**刘冬**,讲师,主研领域:警务数据分析与建模,自然语言处理。**陈一民**,教授。

0　引　言

2022 年 11 月 30 日 OpenAI 发布了基于 GPT(Generative Pre-trained Transformer)3.5 大语言模型的 AI 聊天机器人 ChatGPT。相比基于传统深度神经网络的聊天机器人,ChatGPT 一经发布便以其更出色的上下文理解能力、强大的文本生成能力和丰富的知识储备风靡全球。GPT 系列大语言模型基于 Transformer 架构的解码器(Decoder),GPT3.5 除了使用大量语料数据进行预训练之外,同时还使用了人类反馈强化学习(Reinforcement Learning from Human Feedback,RLHF)技术让模型的输出更接近人类表达[1]。紧随其后,国内外发布了一系列基于 Transformer 架构[2]的大语言模型(Large Language Model,LLM),如百度“文心一言”、清华大学的 ChatGLM、复旦大学的 MOSS、谷歌的 PaLM 和 MetaAI 的 LLaMA 等[3]。目前基于大语言模型的各类聊天机器人和辅助工具已经应用到了智能客服、科研、医疗、交通、软件开发等各个领域[4-8]。大语言模型在大幅度提升社会工作效率和知识查询便捷性的同时带来了诸如 AI 幻觉(AI Hallucinations)[9-10]、AI 偏见(AI Bias)[11]、被违法犯罪人员滥用等新挑战[12],甚至有科研人员使用大语言模型生成学术论文、伪造研究数据从而支持一个“未经证实”的科学主张,这极大地增加了学术界对科研诚信的担忧[13]。与传统深度学习模型相比,大语言模型生成的文本具有更强的可阅读性,与人类正常的文本表达更相接近,普通用户和专家难以通过简单阅读的方法进行分辨,因此如何快速有效地检测文本内容是否为 AI 所生成已经成为当下研究的一个热点[14]。

AI 文本生成内容检测的方法可以归纳总结为两大类:(1) 基于文本特征统计的机器学习方法;(2) 基于神经网络的深度学习分类方法[15]。

基于文本特征统计的机器学习方法首先通过人工归纳总结需要统计的文本特征,然后使用这些特征数据训练机器学习模型进行分类。叶露晨等[16]通过构建“不重复单词的比例”“句子平均长度”“单词平均字母数”“空格频率”“数字频率”等 8 个指标对某社交平台上 107 个问题的 8 475 条人工回答和 535 条 ChatGPT 回答进行分类检测,实验结果表明使用上述特征的机器学习模型取得了 99.1% 的最高分类准确率。但该研究使用的数据集严重类别不均衡,ChatGPT 的回答只占总数 5.94%,模型只需将所有结果预测为人类回答即可获得 94.06% 的分类准确率。Leon Fröhling 等[17]从前人的工作中总结出了“句法和词汇多样性”“重复性”“连贯性”“目的性”和“基本特征”(字符、音节、单词、句子统计)五大类特征,实验结果表明多层前馈神经网络模型使用上述特征可以在“GPT2”生成数据集上取得了 92.7% 的分类准确率,但在“GPT-3”生成的数据集上仅取得了 77.9% 的分类准确率。Kristina Schaaff 等[18]总结了“复杂度(Perplexity)”“语义(Semantic)”“可读性(Readability)”“文本向量(TextVector)”等 8 大类共 37 个统计指标进行 AI 生成文本检测研究,实验结果表明随机森林模型(Random Forest)使用上述 37 个特征在 970 条英语文本(其中 555 条由 ChatGPT 生成)上取了 98% 的分类准确率。但如果直接使用 BERT[19]模型提取的文本向量这单一特征进行分类就可以取得 95% 的分类准确率,剩下的 36 个特征只提升了 3% 的分类准确率。由于特征的好坏直接决定了机器学习算法的分类性能,因此特征的选取就成了基于文本特征统计分类方法的关键,然而这些文本特征的归纳总结过度依赖研究人员的语言学知识和选取技巧,不利于模型的推广使用[20]。

深度学习的检测方法首先对输入的文本进行预处理(分词、去除停用词等),然后将文本转换成词向量,通过深度神经网络提取文本特征后进行分类。范志武[21]研究了 TextCNN(Text Convolutional Neural Network)、RNN(Recurrent Neural Network)、GRU(Gated Recurrent Unit)、DPCNN(Deep Pyramid CNN)等深度学习算法在中文人类文本与 ChatGPT 生成的中文文本中分类性能,实验结果表明基于深度金字塔卷积神经网络(DPCNN)取得了 96.58% 的最高分类准确率,但该结果不如目前更主流的基于预训练语言模型(如 BERT、RoBERTa 等)的分类模型性能。Daphne Ippolito 等[22]使用 250 000 条 GPT-2 生成的英语文本和相同数量的人类专家写的英语文本对 BERT 分类模型进行微调,微调后的模型在测试集上取得了 88% 的分类准确率,同时发现分类模型的准确率与生成模型的自回归算法高度相关:与“Top-k”方法相比,使用“核方法”和“随机采样法”生成的文本分类时准确率分别下降了 7% 和 10%。Guo 等[23]收集了 24 322 个英文问题和 12 853 个中文问题,分别让人类专家和 ChatGPT 对上述问题进行回答,再将上述问答对组成了用于 AI 文本生成检测的 HC3 数据集,实验结果表明使用基于 RoBERTa 的分类模型在英语和中文数据集上分别取得了 98.78% 和 96.40% 的最高分 F1 值,比逻辑回归算法高了 8.72% 和 20.14%。

通过分析上述文献给出的研究结果,我们可以发现,深度学习的检测方法无须开展特征工程,因此其建

模过程更简单,而且这类方法相比基于文本特征统计的机器学习方法分类准确率更高,在 AI 文本检测领域有着更为广阔的应用价值。但所有的深度学习模型都是“端到端”(End to End)的复杂“黑箱”模型,我们无法知道模型如何进行分类决策。分类模型是通过学习训练集的数据进行语义理解来区分人类文本和 AI 生成文本?还是仅仅通过统计两者之间的外在文本差异来分类?通过特定数据集训练的分类模型是否可以用于其他数据集上的检测?这些问题直接决定了 AI 文本检测模型预测结果的可靠性和可信度。为此我们提出了一种基于 BERT-BiGRU 的 ChatGPT 生成文本检测模型,首先测试了该模型在 AI 生成检测数据集 HC3 上的分类性能,并进行模型评估;最后通过 Shapley 加性解释工具从全局和局部两个维度对不同模型所识别出的关键词和基准值进行比较分析。结果表明我们提出的 BERT-BiGRU 模型在中文和英文测试集中分别获得了 98.99% 和 98.62% 的最高分类准确率,但模型所识别出的特征关键词多为数字、生僻词和标点符号,没有明显的规律模型并未真正理解人类撰写文本与 AI 生成文本的内在特征区别。

1 本文研究方法设计

本文研究框架如图 1 所示,分为数据获取与准备、模型训练、模型评估和模型解释四个部分。

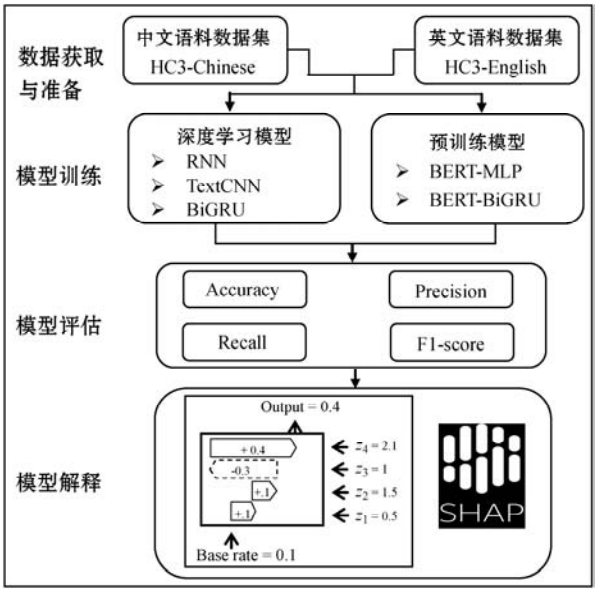


图 1 本文研究框架

1.1 数据获取与准备

本研究使用 HC3 数据集,该数据集由中文和英文两种语料组成,数据集的基本信息如表 1 所示。其中 AI 生成文本使用 ChatGPT(GPT3.5)根据指定问题生成回答,人类文本为人类专家针对上述问题撰写的文本。

表 1 HC3 数据集信息

文本语言	问题数量	人类回答	AI 回答
中文	12 841	22 231	17 328
英语	7 208	7 210	10 225

为了研究问答对和文本语言对模型结果的影响,我们将 HC3 数据集拆解成了 Chinese-A、Chinese-QA、English-A、English-QA、Mix-A 和 Mix-QA 六个数据集,其中 A 表示模型输入文本只是用人类专家和 AI 生成的回答,QA 表示将问题和回答数据拼接组成问答对作为输入,Mix 表示将中文文本和英语文本随机混合组成数据集。我们将上述数据集均按照 8∶1∶1的常用比例随机拆分成训练集、评估集和测试集。为了减少因不同分词工具对模型分类性能的影响,我们选用预训练 BERT 模型的 tokenizer 进行分词并序列化转换。经过上述预处理后得到的数据集具有统一的格式和标准化的分词,更加有利于后续生成检测模型的训练和评估。

1.2 分类模型

本研究综合运用了 RNN、TextCNN、BiGRU 三种深度学习算法以及预训练 BERT 模型,提出了一种基于 BERT-BiGRU 的 ChatGPT 生成文本检测模型。图 2 为 BERT-BiGRU 分类模型示意图,先使用预训练 BERT 提取文本数据的特征,接着使用 BiGRU 将上述特征进一步融合后做分类,然后采用 MLP + Softmax 进行进一步的细分,通过交叉熵损失函数计算模型的损失,最后通过误差逆传播方法更新网络的权重参数。BERT 模型采用了 Transformer 架构中的 Encoder 作为自己的核心模块,它使用了多头自注意力机制(Multi-Head Attention)直接获取输入文本中所有词之间的语义信息,计算过程中并不涉及任何循环或者卷积。因此相比 RNN、TextCNN 等深度神经网络,多头注意力机制可以获得更长效的记忆效果,提取的语义特性也更具代表性。

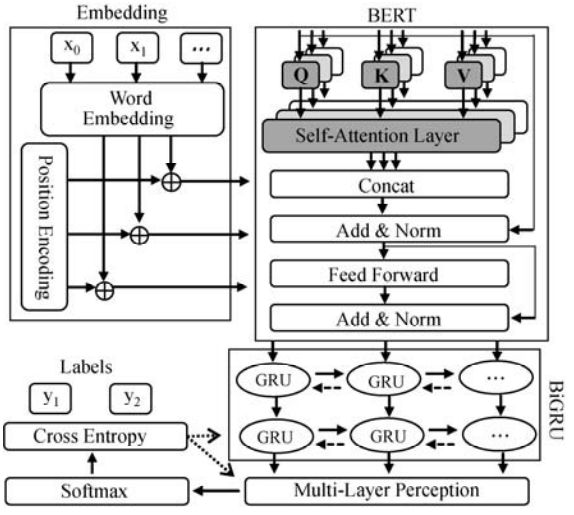


图 2 BERT-BiGRU 模型示意图

本文中的预训练 BERT 模型是经我们改写和适配后的开源“bert-base-chinese”模型,该模型使用的字典大小为 21 128 个字,Transformer Encoder layer 层数为 12,词嵌入维度(Word Embedding dim)为 768,参数总量为 1.03 亿。通过高维度以及巨量数据的预训练,我们可以更好地从文本中提取出语义特征,提高模型分类性能。

BiGRU 既双向 GRU 网络[24],是一种在 LSTM(Long Short-Term Memory)基础之上改进而来的门控循环神经网络,它在保持了 LSTM 出色的文本语义提取能力的同时结构更加简单。GRU 由更新门 z_t 和重置门 r_t 控制着上一步的信息输入、当前步骤的信息状态和信息状态的更新量。GRU 的单元状态计算过程如式(1)-式(4)所示。

$$\boldsymbol{z}_t = \sigma(\boldsymbol{W}_z[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_z) \tag{1}$$

$$\boldsymbol{r}_t = \sigma(\boldsymbol{W}_r[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_r) \tag{2}$$

$$\boldsymbol{h}'_t = \tanh(\boldsymbol{W}_h[\boldsymbol{r}_t h_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_h) \tag{3}$$

$$\boldsymbol{h}_t = \boldsymbol{z}_t \boldsymbol{h}_{t-1} + (1 - \boldsymbol{z}_t)\boldsymbol{h}'_t \tag{4}$$

式中:$\boldsymbol{x}_t$ 为当前步骤的输入,$\boldsymbol{b}_z$、$\boldsymbol{b}_r$ 和 $\boldsymbol{b}_h$ 为偏置项,$\boldsymbol{W}_z$、$\boldsymbol{W}_r$ 和 $\boldsymbol{W}_h$ 为权重矩阵,$\boldsymbol{h}'_t$ 为当前步骤的信息更新量,$\boldsymbol{h}_t$ 为当前步骤的输出。GRU 除了可以正向输入计算,也可以进行反向输入计算,将正反两个 GRU 计算结果拼接在一起后即为 BiGRU。

本文实验使用的操作系统为 Windows 10,编程语言为 Python 3.9,深度学习框架为 PyTorch 2.0.1,CPU 为 Intel i9,主频 2.80 GHz,GPU 为 NVIDIA 3080Ti,内存 32 GB。深度学习模型和预训练 BERT 分类模型使用的训练参数保持一致,训练轮次为 20 个 epoch,学习率为 0.001,batch_size 设置为 48,输入文本最大长度 max_len 为 500。

1.3　模型评估

本文分类实验中模型评估指标有 4 个,分别为准确率 A(Accuracy)、精准率 P(Precision)、召回率 R(Recall)和 F1 值。TP(True Positive)真正例,预测为正例且标签为正例的样本数;FP(False Positive)假正例,预测为正例但标签为负例的样本数;TN(True Negative)真负例,预测为负例且标签为负例的样本数;FN(False Negative)假负例,预测为负例但标签为正例的样本数。

准确率为模型分类正确的样本数占测试样本数的百分比。

$$A = \frac{T_P + T_N}{T_P + F_P + T_N + F_N} \tag{5}$$

精准率 Precision 表示预测为正例的样本中真实为正例的比例。

$$P = \frac{T_P}{T_P + F_P} \tag{6}$$

召回率 Recall 表示正例样本中有多少比例的样本被正确预测为正例。

$$R = \frac{T_P}{T_P + F_N} \tag{7}$$

在 F_1 值是对精准率和召回率的综合评价。

$$F_1 = \frac{2 \times P \times R}{P + R} \tag{8}$$

对于二分类任务,模型总的 Precision、Recall、F1 值是正例和反例 Precision、Recall、F1 值的算术平均。

1.4　SHAP 模型可解释性研究

Shapely 加性解释(Shapley Additive ExPlanations,SHAP)[25]是一种通用模型可解释框架,既可以解释传统机器学习模型,也可以解释深度学习模型,它采用加性特征归因方法(Additive Feature Attribution Methods)将待解释模型的最终预测结果分解成了输入特征重要性值的线性组合,如式(9)所示。

$$f(x) = g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i \tag{9}$$

式中:f 为原始待解释模型,x 为单个输入样本,g 为解释模型,$z' \in \{0,1\}^M$,M 为简化后输入数据的特征个数,ϕ_i 为第 i 个输入特征对应的重要性值。

SHAP 解释框架中 SHAP 值的计算方法引入了合作博弈论中 Shapley 值算法,Shapley 值在计算单个特征贡献度时采用边际贡献率的方法,即单个特征的贡献度(重要性值)等于该特征为模型输出值的边际贡献平均值。Shapley 值计算方法如式(10)所示。

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S)] \tag{10}$$

式中:F 为所有特征的集合,S 为特征子集,$f_{S \cup \{i\}}$ 为使用包含特征子集 S 和第 i 个特征训练的模型,f_S 为特征子集 S 训练的模型,$f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S)$ 为在当前特征子集 S 上第 i 个特征的边际贡献值,遍历所有可能特征子集的边际贡献值加权平均得到的结果即为第 i 个特征的 Shapley 值。由于经典的 Shapley 值计算方法需要遍历所有可能的特征子集,当输入特征较多时上述计算方法将会非常缓慢,因此在计算 Shapley 值我们引入了简化计算方法进行近似估计。

SHAP 值为 Shapley 值的一种简化版估算方法,计算公式如式(11)-式(15)所示。

$$\phi_0 = f(\phi) = E[f(z)] \tag{11}$$

$$\phi_1 = E[f(z) \mid z_1] - E[f(z)] \tag{12}$$

$$\phi_2 = E[f(z) \mid z_{1,2}] - E[f(z) \mid z_1] \tag{13}$$

$$\phi_3 = E[f(z) \mid z_{1,2,3}] - E[f(z) \mid z_{1,2}] \tag{14}$$

$$\vdots$$

$$\phi_i = E[f(z) \mid z_{1,2,\cdots,i}] - E[f(z) \mid z_{1,2,\cdots,i-1}] \tag{15}$$

式中:ϕ 为空集;z_i 为第 i 个输入特征;$E[f(z)]$ 为模型预测值的期望,可用训练样本的模型预测平均值近似;$E[f(z) \mid z_{1,2,\cdots,i}]$ 为当输入特征为 $\{z_1, z_2, \cdots, z_i\}$ 时模型预测的期望。

SHAP 既可以对单条输入数据进行局部解释,也可以在计算得到每条数据的局部解释之后开展全局解释。图 3 为 SHAP 模型解释示意图。

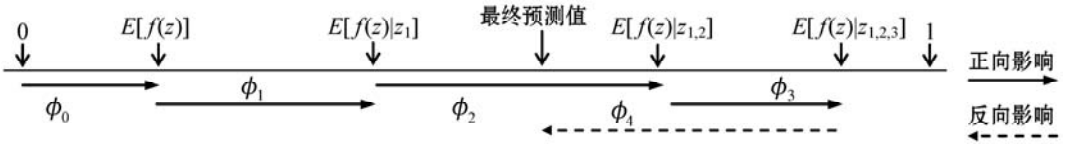


图 3 SHAP 模型解释示意图

SHAP 为输入数据的每个特征分别计算其对所有类别的特征重要性值,如当前特征对某一类别产生正向影响时其 SHAP 值为正,反之其 SHAP 值为负,将输入数据的所有 SHAP 值线性相加得到的结果即为解释模型预测的当前类别概率值。给定一条由 n 个词组成的文本序列样本 $\{x_1, x_2, \cdots, x_j, \cdots, x_n\}$,该样本对类别 c 的预测概率如式(16)所示。

$$p_i^c = \phi_0^c + \phi_1^c + \cdots + \phi_j^c + \cdots + \phi_n^c \tag{16}$$

式中:c 为数据类别标签,二分类任务中 $c \in \{0,1\}$;n 为输入数据序列长度,$j \in \{1,2,\cdots,n\}$;p_i^c 为解释模型将第 i 个文本数据预测为类别 c 的概率;ϕ_j^c 为输入数据中第 j 个词对类别 c 的重要性值,在二分类任务中 ϕ_j^0 和 ϕ_j^1 互为相反数;ϕ_0^c 为解释模型计算得到的类别 c 的基线值,基线值越接近于 1,模型越倾向于将所有数据预测为该类别。循环遍历所有输入数据计算每个词的特征值并按从大到小的顺序排列,即可得到输入特征的全局性解释。

SHAP 使用的加性特征归因方法具有三个理想属性:(1) 局部保真性(Local Accuracy),即针对特定输入 x 当解释模型的输入 z' 逼近原模型简化输入 x' 时,解释模型 g 的输出逼近于待解释模型 f 的输出;(2) 缺失性(Missingness),即原始输入数据中缺失的特征属性对解释模型的输出没有影响;(3) 连续性(Consistency),即待解释模型如果发生改变(某些简化输入特征的贡献增加或者不变),那么该输入特征的贡献值不应减少。上述三个理想属性保证了 SHAP 给出的解释具有一致性和唯一性:即对同一个模型和同一输入计算得到的 SHAP 值是唯一的。上述特性使得 SHAP 能够给出稳定的模型解释,从而帮助我们更好地理解“黑箱”模型的决策过程。

2 实验及结果分析

2.1 文本特征统计

表 2 为 HC3 数据集的文本特征统计。从词汇量上分析,ChatGPT 生成的中文本和英语文本比人类的回答分别少了 22.52% 和 7.75%。从平均长度上看,ChatGPT 生成的文本比人类的回答长了 22.72% 和 42.66%。为了进一步综合比较不同文本的词汇量和平均长度,我们定义了词密度,其计算方法如式(17)所示。

$$D = \frac{V}{L \times 10} \tag{17}$$

式中:D 为词密度;V 为词汇量;L 为平均长度。词密度更加直观地表示了单位句子长度内所使用的不同词汇量:词密度越大,相同长度句子运用的词汇量就更丰富。在四个文本类别中,中文和英文人类回答的词密度分别比 ChatGPT 生成回答高了 58.49% 和 54.65%。综合上述分析,ChatGPT 相比人类倾向于使用更少的词汇量生成更长的句子,两者在文本特征统计上具有较为明显的差异。

表 2 生成文本检测模型分类性能评估结果

文本语种	模型	问答对形式							
		A				QA			
		Accuracy/%	Precision	Recall	F1	Accuracy/%	Precision	Recall	F1
中文	RNN	56.26	0.281 3	0.500 0	0.360 0	56.26	0.281 3	0.500 0	0.360 0
	TextCNN	94.85	0.948 9	0.946 4	0.947 7	93.99	0.939 7	0.9381	0.938 9
	BiGRU	97.32	0.973 8	0.971 9	0.972 8	97.48	0.975 6	0.973 2	0.974 4
	BERT-MLP	98.03	0.980 2	0.979 8	0.980 0	97.58	0.9758	0.974 9	0.975 4
	BERT-BiGRU	**98.86**	**0.9883**	**0.988 7**	**0.988 5**	**98.99**	**0.989 6**	**0.989 9**	**0.989 7**

续表 2

文本语种	模型	问答对形式							
		A				QA			
		Accuracy/%	Precision	Recall	F1	Accuracy/%	Precision	Recall	F1
英语	RNN	61.87	0.690 8	0.542 9	0.608 0	72.76	0.7210	0.7256	0.723 3
	TextCNN	95.47	0.953 5	0.953 0	0.953 3	95.07	0.9480	0.9508	0.949 4
	BiGRU	96.67	0.967 5	0.963 8	0.9657	96.73	0.9668	0.9658	0.966 3
	BERT-MLP	96.62	0.965 3	0.977 5	0.967 5	96.79	0.978 3	0.966 8	0.953 6
	BERT-BiGRU	**98.62**	**0.985 6**	**0.986 0**	**0.985 8**	**98.51**	**0.989 2**	**0.985 4**	**0.979 3**
混合	RNN	51.62	0.558 1	0.504 7	0.530 0	53.06	0.661 6	0.519 3	0.581 9
	TextCNN	94.78	0.948 3	0.947 4	0.947 9	94.02	0.940 7	0.939 9	0.940 3
	BiGRU	96.91	0.969 1	0.969 2	0.969 2	97.13	0.971 2	0.971 2	0.971 2
	BERT-MLP	97.42	0.974 2	0.974 4	0.974 3	97.23	0.972 2	0.972 4	0.972 3
	BERT-BiGRU	**98.63**	**0.986 3**	**0.986 4**	**0.986 3**	**98.44**	**0.984 3**	**0.984 5**	**0.984 4**

2.2　AI 生成检测模型评估

本文实验在 HC3 数据集上分别使用传统深度学习和基于预训练 BERT 的分类模型来实现 AI 生成文本检测,上述模型在测试集上的实验结果如表 3 所示。实验结果表明除了 RNN 模型之外,剩下的模型都取得了 94% 以上的分类准确率,能够正确区分出绝大部分人类文本和 AI 生成文本。但 BiGRU 算法在所有测试集上均取得了 96% 以上的分类准确率,明显优于 RNN 和 TextCNN 算法。

表 3　HC3 文本特征统计

语种	类别标签	词汇量	平均长度	词密度
中文	人类	8 626	146.67	5.88
	ChatGPT	6 683	179.99	3.71
英语	人类	3 430	221.75	1.55
	ChatGPT	3 164	316.34	1.00

借助于在预训练时所学习的海量文本知识,基于预训练 BERT 的分类模型相比传统的深度神经网络拥有更强的语义特征提取能力。因此我们提出使用 BiGRU 网络综合提炼预训练 BERT 提取的文本语义特征,BERT-BiGRU 模型在所有数据集上均取得了最高的分类准确率、精准率、召回率和 F1 值。使用问答对语料(QA)训练的模型相比回答语料(A)对 RNN、BiGRU 和 BERT-BiGRU 模型的分类性能有小幅提升,但在 TextCNN 和 BERT-MLP 算法上略微下降,总体对分类效果提升不显著。为了降低人类撰写的问题对分类模型解释的负面影响,我们在研究中仅采用回答语料训练的模型。

虽然使用深度神经网络算法和基于预训练 BERT 的分类模型在同语种测试集上表现出了较好的分类性能,但为了进一步研究 AI 生成检测模型的泛化性能,我们使用与模型训练集语种不同的测试集进行泛化性能测试,实验结果如表 4 所示。

表 4　跨语种测试结果

训练集	测试集	模型	Accuracy/%	F1
中文	英语	RNN	41.28	0.292 2
		TextCNN	66.69	0.669 1
		BiGRU	59.92	0.599 8
		BERT-MLP	58.83	0.614 6
		BERT-BiGRU	**62.21**	**0.628 4**
英语	中文	RNN	44.45	0.594 6
		TextCNN	52.09	0.574 1
		BiGRU	59.49	0.591 5
		BERT-MLP	61.51	0.590 9
		BERT-BiGRU	**75.16**	**0.747 6**

可以看出,所有模型在与训练集不同语种测试集上分类性能都大幅度下降,但 BERT-BiGRU 模型的综合分类性能仍优于其他模型。结合表 2 中混合语料训练模型优异的分类性能表现,我们可以发现无论是基于预训练 BERT 的分类模型还是深度神经网络模型,虽然它们能在已学习过的封闭数据集进行正确分类,但在未学习过的语种语料上分类性能均会大幅度下降,其预测结果将不再可靠。

2.3　模型全局性解释

BiGRU 模型和 BERT-BiGRU 模型分别在深度神经网络模型和预训练 BERT 模型中取得了最佳分类性能,因此我们分别选择中文 BiGRU、英语 BiGRU、中文 BERT-BiGRU 和英语 BERT-BiGRU 四个模型进行解

释。由于 SHAP 解释框架需要的运算量极大,为了将计算开销控制在合理的时间范围内,我们在测试集中随机选取其中的 1 500 条样本进行解释。

表 5 为按照特征重要性值由大到小排序的中文数据集关键词 Top15 和其对应的重要性值,others 表示解释样本中剩余词的特征值总和。在这些特征关键词中大多是数字、生僻字、英语单词和标点符号,没有明显的规律。虽然 BiGRU 和 BERT-BiGRU 模型的分类准确率都很高,但两者只在 AI 生成文本中发现了“error”“network”和“was”三个相同关键词。通过搜索发现中文语料库中共出现了 34 次“network error”,标签都为 AI 生成文本,这是使用 ChatGPT 的 API 生成回答时因网络连接错误生成的信息,这些信息被模型识别成了 AI 生成文本的重要特征。

表 5 中文数据集关键词 Top 15 和重要性值

BiGRU				BERT-BiGRU			
人类	重要性值	ChatGPT	重要性值	人类	重要性值	ChatGPT	重要性值
455	0.09	**network**	0.22	1 948	0.02	**error**	0.14
の	0.08	**error**	0.20	1 893	0.02	請	0.09
401	0.07	谴	0.13	蔚	0.01	179	0.06
1 880	0.07	ana	0.12	cookies	0.01	宫	0.06
杠	0.06	吾	0.10	sound	0.01	**was**	0.04
ner	0.06	旻	0.09	12	0.01	**network**	0.04
application	0.05	束	0.08	1945	0.01	畿	0.04
<	0.05	≈	0.08	bl	0.01	県	0.04
info	0.05	蛛	0.08	吾	0.01	university	0.04
/	0.04	三	0.07	library	0.01	ner	0.03
亩	0.04	lio	0.07	1 982	0.01	please	0.03
1 938	0.04	**was**	0.07	51	0.01	chicken	0.03
躬	0.04	它	0.06	res	0.01	sorry	0.03
1 942	0.04	900	0.06	970	0.01	纬	0.03
512	0.04	express	0.06	181	0.00	溥	0.03
others	-7.29	others	4.97	others	-10.97	others	10.07

数字在人类文本关键词中分别出现了 6 次和 8 次,但在 AI 生成文本关键词中只出现了 1 次,这表明分类模型发现人类在撰写文本时比 AI 更喜欢用数字。从重要性值上看,同一分类模型计算得到的人类文本特征词重要性值普遍小于 AI 生成文本的重要性值,并且所有模型的人类文本 others 都为负数,AI 生成文本的 others 都为正数。这表明模型在训练时更容易捕捉和发现 AI 生成文本的用词规律,这与表 2 中 AI 生成文本句子长度更长但词汇量更少的规律相符。

基线值是解释模型的重要指标之一,它可以有效解释模型在分类决策时的策略。解释模型基线值的算术平均如表 6 所示,所有模型的人类文本基准值都约等于 1,而 AI 生成文本的基准值都约等于 0。结合表 2 和表 5 的数据,我们认为分类模型在做分类决策时将输入的文本数据都默认为是人类撰写文本:通过捕捉和统计计算 AI 生成特征词汇的重要性值来判断该文本是否为 AI 所生成。人类文本因词汇量更大,但特征重要性值较小且有正有负,因此其最后的重要性值求和接近于 0,解释模型的输出结果仍为基线值 1;AI 生成文本的基线值虽然为 0,但模型更容易捕捉到 AI 生成文本中的特征词,结合其特征重要性值较大且句子更长的特点,因此这些重要性值求和后变为了 1。

表 6 解释模型的基线平均值

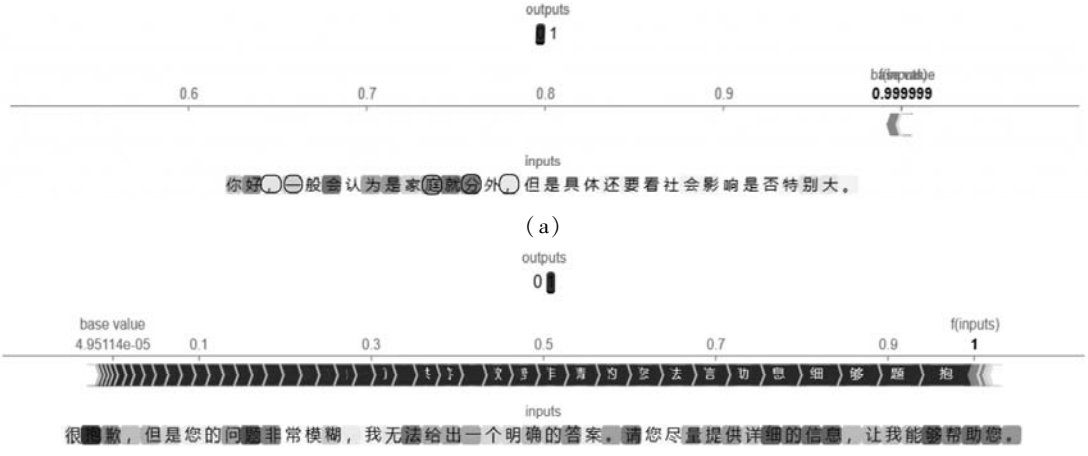
文本	Model	人类	ChatGPT
中文	BiGRU	1.00E+00	9.08E-09
	BERT-BiGRU	9.99E-01	9.20E-04
英语	BiGRU	9.19E-01	8.12E-02
	BERT-BiGRU	1.00E+00	2.81E-04
混合	BiGRU	1.00E+00	1.24E-04
	BERT-BiGRU	1.00E+00	3.98E-06

2.4 模型局部性解释

为了进一步研究分类模型的分类决策过程,我们使用 SHAP 对具体个案样本进行解释,在中文和英语数据集上我们各挑选一条人类回答文本和 ChatGPT 生成回答文本进行解释,被解释的模型选用分类性能最

优的 BERT-BiGRU 模型。

图 4 是基于 SHAP 解释工具的 BERT-BiGRU 模型中文个案分析,图 4(a)为人类撰写的文本,图 4(b)为 ChatGPT 生成的文本。图中“output”为解释模型的输出结果,“0”为“人类文本”,“1”为“AI 生成”。“inputs”为输入的待解释文本,圆角矩形“▭”圈出的字对当前分类为负贡献,其余为正贡献,字体背景颜色越深代表其特征重要性的绝对数值越大,反之则越小。图(a)中 SHAP 计算得到“你”“好”“会”“为”“是”“就”“别”“大”8 个词可以增加该样本是人类文本概率,而“,”“—”“庭”“分”这 4 个词有反向贡献,剩下的 16 个词对分类几乎没有贡献。

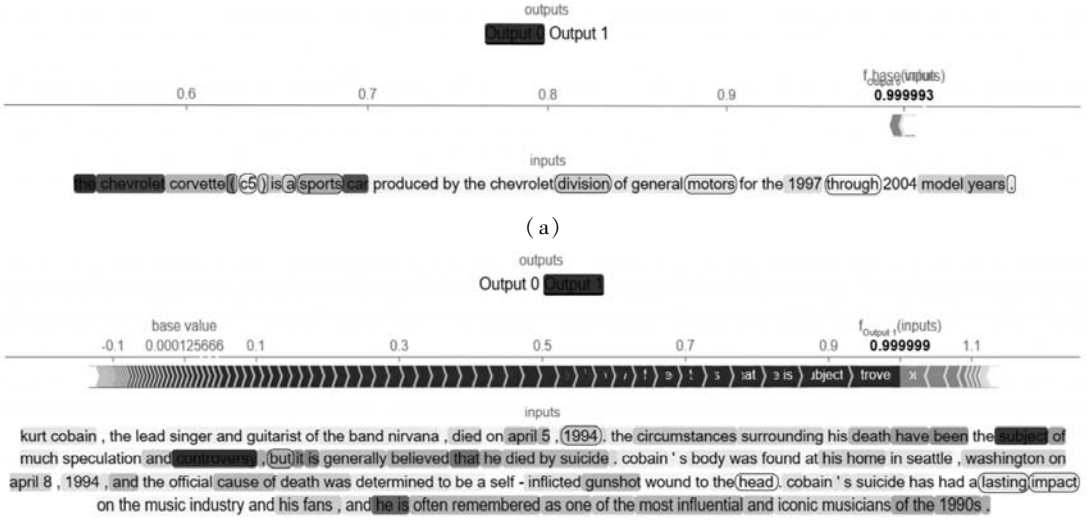


(a)　　(b)

图 4　BERT-BiGRU 模型的中文个案解释

通过上下文分析我们可以发现上述输入文本中的“就分”是错别字,此处应该是“纠纷”两字更合理。相比 AI,人类撰写的文本更有可能包含较多的错别字,但模型只给予了“就”字较高的正向贡献,而“分”字却较大的反向贡献,这与人的直觉并不相符。从图(b)中可以发现模型将大部分词认定为对“AI 生成”这一标签有正向贡献。上述这些特征词都是一些常用词,即没有特殊的含义也没有好坏之分,并不能单独确定该文本是否为 AI 生成。

图 5 是基于 SHAP 解释工具的 BERT-BiGRU 模型英语个案分析,(a)为人类撰写的文本,(b)为 ChatGPT 生成的文本。我们发现英语个案遵循着与中文个案同样的规律:模型将待检测的文本全部默认为人类文本,再计算重要性值后判断该文本是否为 AI 所生成。模型检测发现的关键词也多为“he”“subject”“controversy”“(”“the”“chevrolet”“car”等常规字符,并没有特殊含义,无法简单判定该文本是否为 AI 生成。



(a)　　(b)

图 5　BERT-BiGRU 模型的英语个案解释

综上所述,基于语义的 AI 生成检测模型虽然能够在封闭数据集上取得优异的分类效果,但这些模型捕捉到的特征更多地依赖于训练数据集特有的外在特征,模型并没有真正理解 ChatGPT 生成文本与人类文

本之间的内在区别。

3 结 语

针对如何快速识别文本内容是否为 ChatGPT 所生成文本难以检测的问题,本文提出了一种 BERT-BiGRU AI 生成文本分类检测模型,首先使用预训练 BERT 模型提取输入文本的语义特征,然后使用 BiGRU 对提取的语义特征进行综合提炼从而提升模型的分类性能。实验结果表明基于预训练 BERT 的分类模型性能优于传统深度学习模型,BERT-BiGRU 模型在中文和英语测试集上分别取得了 98.99% 和 98.62% 的最高分类准确率。为了进一步研究分类模型是否真正理解人类撰写文本和 ChatGPT 生成文本之间的内在区别,本文首先使用与训练数据不同语种的数据进行分类测试,然后使用 SHAP 框架从全局和局部两个维度对分类模型进行解释。研究结果表明虽然基于深度学习和预训练 BERT 的分类模型在封闭测试集上取得了优异表现,但模型并没有真正理解人类撰写文本与 AI 生成文本的内在特征区别,在未学习过的语料上分类性能下降严重;模型将所有的待检测文本都默认为了人类撰写文本,模型所识别出的关键特征词多为数字、生僻词和标点符号,没有明显的规律,模型只是利用 AI 生成文本所用词汇量更少、文本长度更长这一点来统计发现 AI 生成文本中的关键词,最终判断该文本是否为 AI 所生成。

参 考 文 献

[1] Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback[EB]. arXiv:2203.02155,2022.

[2] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[EB]. arXiv:1706.03762,2017.

[3] Cerf V G. Large language models[J]. Communications of the ACM,2023,66(8):7.

[4] Biswas S. Potential use of chat GPT in global warming[J]. Annals of Biomedical Engineering,2023,51(6):1126 - 1127.

[5] Remedios D, Remedios A. Transformers, codes and labels: Large language modelling for natural language processing in clinical radiology[J]. European Radiology,2023,33(6):4226 - 4227.

[6] Jablonka K M, Schwaller P, Ortega-Guerrero A, et al. Leveraging large language models for predictive chemistry[J]. Nature Machine Intelligence,2024,6(2):161 - 169.

[7] Lee G H, Lee K J, Jeong B, et al. Developing personalized marketing service using generative AI[J]. IEEE Access,2024,12:22394 - 22402.

[8] Biswas S. Role of chat GPT in public health[J]. Annals of Biomedical Engineering,2023,51(5):868 - 869.

[9] Siontis K C, Attia Z I, Asirvatham S J, et al. ChatGPT hallucinating: Can it get any more humanlike? [J]. European Heart Journal,2024,45(5):321 - 323.

[10] Kreps S, McCain M R, Brundage M. All the news that's fit to fabricate: AI-generated text as a tool of media misinformation[J]. Journal of Experimental Political Science,2022,9(1):104 - 117.

[11] 汪怀君,汝绪华. 人工智能算法歧视及其治理[J]. 科学技术哲学研究,2020,37(2):101 - 106.

[12] 谢梅,王世龙. ChatGPT 出圈后人工智能生成内容的风险类型及其治理[J]. 新闻界,2023(8):51 - 60.

[13] Naddaf M. ChatGPT generates fake data set to support scientific hypothesis[J]. Nature,2023,623(7989):895 - 896.

[14] Sadasivan V S, Kumar A, Balasubramanian S, et al. Can AI-generated text be reliably detected? [EB]. arXiv:2303.11156,2023.

[15] Crothers E N, Japkowicz N, Viktor H L. Machine-generated text: A comprehensive survey of threat models and detection methods[J]. IEEE Access,2023,11:70977 - 71002.

[16] 叶露晨,范渊,王欣,等. 大型语言模型内容检测算法和绕过机制研究[J]. 信息安全研究,2023,9(6):524 - 532.

[17] Fröhling L, Zubiaga A. Feature-based detection of automated language models: Tackling GPT-2, GPT-3 and Grover [J]. PeerJ Computer Science,2021,7:e443.

[18] Schaaff K, Schlippe T, Mindner L. Classification of Human and AI-generated texts for English, French, German, and Spanish[C]//International Conference on Natural Language and Speech Processing,2023:1 - 10.

[19] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding [EB]. arXiv:1810.04805,2018.

[20] Sandler M, Choung H, Ross A, et al. A linguistic comparison between human and ChatGPT-generated conversations [EB]. arXiv:2401.16587,2024.

[21] 范志武,姚金良. 基于深度金字塔卷积神经网络的 ChatGPT 生成文本检测方法[J]. 数据分析与知识发现,2024,8(7):14 - 22.

[22] Ippolito D, Duckworth D, Callison-Burch C, et al. Automatic detection of generated text is easiest when humans are fooled[C]//58th Annual Meeting of the Association for Computational Linguistics,2019:1808 - 1822.

[23] Guo B, Zhang X, Wang Z Y, et al. How close is ChatGPT to human experts? Comparison corpus, evaluation, and detection[EB]. arXiv:2301.07597,2023.

[24] Cho K, Merrienboer B, Gulcehre C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[EB]. arXiv:1406.1078,2014.

[25] Lundberg S M, Lee S. A unified approach to interpreting model predictions [C]//31st International Conference on Neural Information Processing Syste,2017:4768 - 4777.