学校代码：10246
学　　号：22210690089

復旦大學

硕 士 学 位 论 文

（专业学位）

基于注意力进行数据增强的机器生成文本检测

Detecting Machine-Generated Text via Attention-Based Data Augmentation

院　　系：　管理学院
专业学位类别(领域)：　应用统计
姓　　名：　冯超
指 导 教 师：　张成洪　教授
完 成 日 期：2024 年 5 月 10 日

目 录

插图目录

表格目录

摘　要

大语言模型的发展为文本生成和信息检索带来了诸多便利，但误用或滥用机器生成的文本可能会带来重大风险，因此准确检测机器生成文本具有重要意义。国内外文献主要基于英文语料数据集，且对混合来源文本的检测问题关注较少。本文构建了一个多领域、多模型生成的人类和机器生成中文文本数据集，包含 8 万余条文本，并基于此数据集对单一来源文本和混合来源文本检测问题进行了实证分析。

针对单一来源文本的检测问题，本文首先构造了文本长度、用词丰富度、文本困惑度等 6 个文本统计特征，发现人类和机器生成的文本在一些统计特征上存在显著差异。基于这些特征的 XGBoost 分类模型可以取得 92.76% 的 F1 值，其中文本困惑度对分类效果贡献最大，占比达 62.63%。根据文本所属领域和来源是否相同，设计了 4 种不同的实验组合，使用 fastText 进行文本分类，可以取得约 90% 的 F1 值，并发现跨领域、跨模型的分类难度最大，模型效果最差。基于 Transformer 架构的 BERT 模型展现了最先进的分类效果，F1 值达到 98.6%。而直接使用 ChatGPT 进行零样本学习识别机器生成文本，在人类生成文本召回率上接近于随机猜测。

本文使用机器续写的方式模拟构造混合来源文本，并使用基于注意力和分类效果筛选的数据增强算法改进 BERT 模型的分类效果。本文可视化分析了该数据增强算法在分类错误样本上的注意力转移情况，说明了经数据增强改进后的模型可以正确关注混合来源文本中的由机器生成的部分。与随机拼接和随机筛选的数据增强方法对比，本文改进的数据增强算法在混合来源文本检测任务上取得了最好的效果，F1 值提升了约 5%。最后，本文提供了一个机器生成文本检测应用，可供用户在线识别机器生成文本并进行分句可解释性分析。

关键词：机器生成文本；文本分类；数据增强；注意力机制
中图分类号：TP391

Abstract

The advancement of large language models significantly aids text generation and information retrieval but also raises concerns over the potential misuse of machine-generated text. Accurate detection of such text is crucial. Previous research has largely concentrated on English datasets, with a limited focus on texts from mixed sources. This paper constructs a comprehensive dataset with over 80000 samples of both human and machine-generated Chinese texts across multiple domains and generative models. It provides an empirical analysis of detecting texts from single and mixed sources.

In detecting texts from a single source, we construct 6 statistical features, such as text length, lexical diversity, and text perplexity, which can be used to distinguish between human and machine-generated texts effectively. An XGBoost model using these features achieves a 92.76% F1 score, with perplexity being the most significant contributor, accounting for 62.63%. According to whether the texts' domain and source are the same, 4 different testbeds are designed. Using fastText for text classification achieves an F1 score of 90%. The cross-domain and cross-model testbed has the highest task difficulty and worst classification performance. The transformer-based BERT model demonstrates state-of-the-art detection capability, with an F1 score reaching 98.6%. Directly using ChatGPT itself for zero-shot learning approaches random guessing in terms of recall rate for human-written text.

We use continuation methods to create mixed-source texts and apply a data augmentation algorithm, leveraging attention and classification logit value, to enhance the BERT model's performance. We visualize the attention shift on a misclassified sample, demonstrating that the model, after being enhanced by data augmentation, can correctly focus on the machine-generated part within mixed-source texts. Compared with random concatenation and random selection, our method achieves the best results in detecting mixed-source texts, with an F1 score improvement of about 5%. We also develop an application offering online detection and sentence-level interpretability analysis.

Keywords: Machine-Generated Text; Text Classification; Data Augmentation; Attention Mechanism

CLC number: TP391

第 1 章　绪论

1.1　研究背景及意义

近年来，随着人工智能技术的快速发展，以 ChatGPT[1] 为代表的大语言模型在文本生成领域展现了卓越的能力。这些大语言模型十分擅长生成流畅、符合语法规范且看似令人信服的文本。随着大语言模型能力的不断增强，由机器生成的文本质量也逐渐接近人类，甚至在某些任务上体现出超越人类的水平，这使得大语言模型在许多领域都有着丰富的应用场景。例如，使用机器生成的文本作为营销文案、智能辅助客服问答等[2]，极大地提高了人类的生产效率。

尽管大语言模型的发展为文本生成和信息检索等带来了诸多便利，但误用或滥用机器生成的文本可能会带来重大风险，尤其是在教育、新闻、法律、医疗等对文本真实性、可靠性和完整性要求较高的领域[3–4]。一份对 1000 名大学生的调查数据显示，有超过 89% 的学生使用 ChatGPT 来帮助完成家庭作业，超过一半的学生曾用 ChatGPT 写过一篇论文[5]。这些行为不仅可能违反学术诚信的原则，还可能抑制学生的批判性思维和创造力。2023 年初，一些个人和媒体利用 ChatGPT 生成并发布了关于“杭州取消限行”、“甘肃火车事故”等虚假新闻，并迅速被广泛传播，对广大民众产生了误导并引起社会混乱[6]。这些案例表明，若不恰当地使用机器生成的文本，可能会对社会产生严重的负面影响。

为了对机器生成文本的使用进行规范和治理，从而能够更负责任地推动人工智能技术的进步，有效地检测人类与机器生成文本的方法就显得尤为重要。对于使用机器生成文本的创作者而言，了解文本中的哪些部分最像由机器生成，不仅有助于保持作品的原创性和独特性，而且对于提升写作品质、减少机器生成内容的不当使用都有重要意义。对于接收信息的读者而言，检测机器生成的文本能增强他们对信息来源和准确性的认识，帮助评估文本的可信度，从而在面对可能含有欺骗、误导或不准确信息的内容时作出更明智的判断。此外，准确检测文本来源还能促进社会对信息质量的关注，鼓励更负责任的内容生产和传播行为，有效减少盲目传播不实信息的事件发生。

综上所述，随着以 ChatGPT 为代表的大语言模型能力持续提升，其应用场景也在不断扩展。然而，误用或滥用机器生成文本可能造成的负面影响也不容忽视。因此，准确地区分人类与机器生成的文本，对于提高文本创作的质量、保障信息接收者免受误导以及维护社会信息秩序的稳定具有至关重要的意义。

1.2　国内外研究现状

1.2.1　文本生成模型

本文的研究对象是机器生成的文本，它是指由机器制造、修改或拓展的文本[7]。具体地，本文关注的是机器生成的自然语言中文文本，它有别于编程语言等非自然语言。本小节介绍现有机器生成自然文本的方法，它们可以按照发展历程分为四个阶段：统计模型、神经网络模型、预训练模型和大语言模型。

早期的语言模型主要是基于统计模型的方法，例如 Jelinek[8]、Rosenfeld[9]和 Szymanski et al.[10]。这些模型的基本思想是基于马尔科夫假设，使用最近的若干上文来预测下一个词。由于不同词之间存在的转换可能性随词数呈指数级增长，高维度降低了统计模型对马尔科夫转移概率矩阵的估计准确度。Katz[11] 提出的平滑方法可以缓解维数灾难的问题，使模型可以对更多样的上文进行预测，即使这些上文在训练集中没有出现过。

随着深度学习技术的发展，基于多层感知机、循环神经网络（RNN）等神经网络的文本生成模型逐渐成为主流。Bengio et al.[12] 提出了一种使用神经网络学习每个单词的分布式表征的方法，并用这些分布式表征来计算语言模型的概率函数。之后，Mikolov et al.[13–14] 提出的 word2vec 词向量模型的出现，使神经网络模型在学习文本表示上展现出了更好的性能，对自然语言处理领域中的许多任务都产生了重要影响。

2018 年，作为较早被提出的预训练模型，ELMo[15] 使用了一个双向长短期记忆网络（biLSTM）来学习基于上下文的词向量表示，而不是使用固定的词向量。这一预训练模型可以根据下游任务的不同而进行微调。2019 年，Devlin et al.[16] 在大量无标注语料上训练了 BERT 模型，它使用了基于多头自注意力机制的 Transformer[17] 架构。这些预训练模型极大地提升了许多自然语言处理任务的表现，也吸引了大量研究者关注预训练模型这一个研究方向，例如使用单向 Transformer 架构的 GPT-2[18]。

当预训练模型的训练预料和参数的规模逐渐增大后，许多研究发现，模型的能力也随之增强，甚至出现“涌现”现象[19]。例如，拥有 1750 亿个参数的 GPT-3[20] 能够通过上下文实现小样本学习，而仅有 15 亿个参数的 GPT-2 则无法做得很好。研究者们将这些大规模的预训练模型称为大语言模型（LLM）[21–22]。目前影响力最大的大语言模型之一是 OpenAI 于 2022 年发布的 ChatGPT[1]，它使用 GPT 系列模型实现对话效果，展现了强大的文本生成能力。

1.2.2 机器生成文本的威胁

大语言模型出现之后，机器生成文本已经展现出了令人印象深刻的能力，其质量足以与人类专家作者相媲美，因此迅速在社交媒体内容创作、客户服务、教育辅助和语言翻译等多个领域中得到了广泛应用。

然而，和任何强大的技术一样，机器生成文本也带来了一系列潜在的威胁。Baki et al.[23] 使用机器生成大规模的垃圾邮件，并发现人类无法准确识别这些由机器生成的有害文本。在社交媒体内容创作领域，Crothers et al.[7] 指出大语言模型可以基于用户信息、历史对话和其他上下文生成高度个性化的内容，在社交媒体上制造虚假信息和误导性内容，成为社交蠕虫、网络诈骗的工具。Stiff 和 Johansson[24] 发现，机器可以在短时间内生成大量文本，其生成的虚假新闻、误导信息可以影响社会舆论甚至政治选举。在学术领域，使用机器生成的文本完成作业、论文变得越来越容易实现[25–26]，某些由机器生成的科研论文甚至能够通过同行评审[27]。

当人们无法准确识别文本是否为机器生成时，容易引发信任危机和对信息真实性的质疑。例如，鉴于求职者可以轻松使用机器生成一封优秀的求职信，某些雇主可能会对求职信丧失信心，从而扰乱正常的招聘流程[7]。作为与大语言模型密切相关的学术组织，计算语言学协会（ACL）在 2023 年发布了一份声明[28]，要求研究者规范使用机器生成文本，例如不鼓励使用机器生成的新想法和新文本、在某些使用场景下需要承认存在机器生成的文本，以提高研究对公众的透明度，以避免机器生成文本带来的负面影响。

1.2.3 机器生成文本检测方法

机器生成文本检测问题可以视为一个二分类问题，即判断文本是由人类还是机器生成的。目前，国内外学者们提出了许多机器生成文本检测的方法，这些方法可以分为基于特征的方法和基于神经网络的方法两大类。

基于特征的文本检测方法

许多研究发现，人类和机器生成的文本在一些统计特征上通常表现不一致，这可以帮助区分两者。基于特征的方法通常需要使用自然语言处理技术构造文本的特征向量，再基于这些特征向量使用逻辑回归、支持向量机、随机森林等机器学习算法进行分类。

按照构造特征的角度，研究者们构造的特征主要包括基本文本特征、基于词频的特征和借助其他自然语言处理模型的语义特征。基本文本特征包括文本长度、标点符号个数等，主要是对一段文本进行基础统计指标的计算，例如 Fröhling et al.[29]。基于词频的特征则是统计文本中特定词汇出现的频率。齐夫定律指出，

自然语言中的词汇出现频率呈幂律分布，即少数词汇出现频率很高，而大部分词汇出现频率很低[30]。Nguyen-Son et al.[31]发现，人类生成的文本通常比较符合齐夫定律，而机器生成的文本则有所背离。Gehrmann et al.[32] 观察到机器生成的文本更容易出现重复的词汇或句子，而人类生成的文本则更加多样化。一些研究借助其他自然语言处理模型，例如命名实体识别、词性标注、情感分析和文本困惑度等，以此构造更加复杂的语义特征，例如 Guo et al.[33] 和 Li et al.[34]，他们发现机器生成文本通常具有低困惑度，且文本情感更偏中性。

基于神经网络的文本检测方法

和许多自然语言处理任务一样，基于神经网络模型的算法通常能取得最先进的表现，尤其是基于 Transformer 架构的神经网络模型。基于神经网络的文本检测方法通常有两种研究方法，分别是基于深度学习模型在特定的检测任务上进行微调，以及直接使用语言模型自身进行零样本学习。

基于深度学习模型进行微调的方法主要使用 BERT[16] 和 RoBERTa[35] 等预训练模型，例如 Liao et al.[36]和 Liu et al.[37]等。这些研究使用预训练模型得到文本表示向量，再使用全连接分类层即可输出模型对于文本来源的分类，在特定数据集上的准确率、F1 值等评价指标能达到 95% 以上的检测效果。更大规模的预训练模型通常能带来更好的模型表现，但其训练和微调说耗费的计算资源和时间也更多。

零样本学习不需要额外的标注数据，在结合一定的提示词后，就可以直接使用它生成的内容作为分类结果。例如，在将提示词设定为“该文本是否由机器生成？”后，可用模型回答的“是”或“不是”作为分类结果。例如，Zellers et al.[38] 发现抵御由 Grover 模型生成的虚假新闻的最佳方法是 Grover 本身，其检测准确率高达 92%。Mitchell et al.[39] 使用语言模型本身计算各个词出现的对数概率，发现在略微打乱文本顺序后，机器生成的文本的对数概率值会明显下降。这种方法属于白盒模型，即它需要检测者能够获得语言模型内部的输出值。当生成文本的模型与用于检测的模型不同时，检测效果通常会下降。

1.3　研究内容与框架

本文共包含六章，文章结构和各章的具体内容安排如下：

第一章为绪论。本章首先介绍了机器生成文本检测的研究背景及意义。其次，介绍了国内外现有对机器生成文本检测的相关研究，从文本生成模型、潜在威胁以及机器生成文本检测算法三个方面进行了梳理和总结。最后介绍了本文的研究内容与框架，并阐述了创新点和潜在的局限性。

第二章为相关理论及技术。本章介绍了本研究用到的若干理论和技术，包括XGBoost、fastText、BERT 这些适用于文本检测问题的经典分类算法，以及文本分类任务的分类效果评价指标。

第三章为数据构造。本章首先介绍了本研究使用的多领域、多模型生成的数据集的构造和清洗过程，然后对数据集的基本特征进行了统计和可视化分析，说明了人类和机器生成的文本的异同点。最后根据机器生成文本是否来自相同领域、是否为相同模型生成这两个标准，设计了 4 种不同的实验组合，为后续的实证分析划分了不同的数据集。

第四章为单一来源文本检测的实证分析。本章针对第四章构建的不同实验组合下的数据集，使用 fastText 进行文本分类。在不同领域、不同模型生成这一最为复杂的实验组合中，进一步使用 XGBoost、BERT 和 GPT-3.5 模型进行文本分类，并比较了各个分类方法的检测效果和优劣。

第五章为混合文本检测的实证分析。本章通过机器续写的方式构建由人类和机器混合生成的文本，并使用第四章构建的 BERT 模型进行文本分类，指出了混合文本检测的难点和挑战。通过使用基于注意力机制的文本混合方法，以及基于分类效果提升的筛选标准，提出了一种适用于人类和机器混合文本检测的数据增强方法。同时，本章可视化案例分析了数据增强前后模型对混合文本的 SHAP 分句可解释性，多种数据增强方法的实验结果也说明了数据增强方法的有效性。

第六章为总结与展望。本章对本文的研究内容和成果进行了总结。针对研究中存在的局限性与不足之处，对未来的研究方向进行了展望。

1.4 创新点

本文的创新点主要体现在中文语料数据集的构建、针对混合文本检测的数据增强方法和实验结论这三个方面。

1. 本文构建了一个多领域、多模型生成的中文语料数据集。本文基于金融、法律、医疗等 7 个领域的人类问答中文文本数据，调用了 OpenAI、讯飞星火、百度等 5 个知名的国内外大语言模型，构建了 8 万余条人类和机器生成的文本和 8 千余条由机器续写的混合文本。

2. 本文改进了一种适用于人类和机器混合文本检测的数据增强方法。基于注意力机制对文本进行分割，将错误分类与正确分类的文本进行混合，并基于对分类效果提升的程度对数据增强的文本进行筛选，从而扩充训练数据。该方法能够有效提升混合文本检测的分类效果，为混合文本检测提供了一种新的改进思路。

3. 本文对中文语料下的人类和机器生成文本的检测效果进行了较为全面的

实证分析。相对英文语料而言，目前中文语料下的人类和机器生成文本检测研究较少。本文使用多个文本分类算法，在多个实验设定下实证分析了单一来源文本检测和混合文本检测的效果，为中文语料下的人类和机器生成文本检测研究提供了一定的补充和参考。

第 2 章　相关理论及技术

2.1　文本分类算法概述

本节将介绍本研究中所使用的文本分类算法，包括 XGBoost、fastText、BERT 预训练模型这三种经典的机器学习和深度学习算法。

2.1.1　XGBoost

eXtreme Gradient Boosting[40] (XGBoost) 是一种基于决策树的集成学习算法，其对样本标签的预测值是由多棵决策树的预测值累加得到的，即加法模型：

$$\hat{y}_i = \phi\left(\mathbf{x}_i\right) = \sum_{k=1}^{K} f_k\left(\mathbf{x}_i\right), \quad f_k \in \mathscr{F} \tag{2.1}$$

式 2.1 中的 $\hat{y}_i$ 是模型对样本 i 标签的预测值，$\mathbf{x}_i$ 是样本 i 的特征，K 是决策树的个数，参数空间 $\mathscr{F}$ 是所有 CART 决策树的集合：

$$\mathscr{F} = \left\{f(\mathbf{x}) = w_{q(\mathbf{x})}\right\} \quad q : \mathbb{R}^m \to T, w \in \mathbb{R}^T \tag{2.2}$$

式 2.2 中的 T 是决策树的叶子节点的个数；$q(x) \in \{1, 2, \ldots, T\}$ 决定了样本 $\mathbf{x}$ 最终归属的决策树叶子节点的索引，即决策树路径划分的抽象表达；$w = (w_1, w_2, \ldots, w_T)^T$ 表示每个叶子节点的输出值。

与梯度提升决策树 (Gradient Boosting Decision Tree，即 GBDT) 类似，XGBoost 的每棵子树都是拟合当前预测值的残差。此外，XGBoost 还考虑了对决策树的叶子节点个数和每个叶子节点的输出值同时进行惩罚，以防止过拟合。式 2.3 中的两项分别对叶子节点个数较大、叶子节点输出值的二范数较大的情形进行了惩罚：

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \|w\|^2 \tag{2.3}$$

其中，T 是叶子节点的个数，w 是叶子节点的输出值。γ 和 λ 是超参数，分别控制对叶子节点个数和叶子节点输出值的惩罚程度。

因此，XGBoost 考虑损失函数和正则化项后的目标函数为：

$$\begin{aligned}\mathscr{L}(\phi) &= \sum_{i=1}^{N} \ell\left(y_i, \hat{y}_i\right) + \sum_{k=1}^{K} \Omega\left(f_k\right) \\ &= \sum_{i=1}^{N} \ell\left(y_i, \sum_{k=1}^{K} f_k\left(\mathbf{x}_i\right)\right) + \sum_{k=1}^{K} \Omega\left(f_k\right)\end{aligned} \tag{2.4}$$

式 2.4 中的 N 是样本数量。给定第 $t-1$ 轮迭代的模型 ϕ_{t-1}，XGBoost 的目标是找到第 t 棵决策树 f_t 使得目标函数 $\mathscr{L}(\phi)$ 最小。对于任意的目标函数，XGBoost 使用泰勒展开式对损失函数进行二阶近似：

$$\begin{aligned}\mathscr{L}^{(t)} &= \sum_{i=1}^{N} \ell\left(y_i, \hat{y}_i^{(t-1)} + f_t\left(\mathbf{x}_i\right)\right) + \Omega(f_t) \\ &\approx \sum_{i=1}^{N} \left(\ell(y_i, \hat{y}_i{}^{(t-1)}) + g_i f_t(\mathbf{x}_i) + \frac{1}{2} h_i f_t^2(\mathbf{x}_i)\right) + \Omega(f_t)\end{aligned} \tag{2.5}$$

式 2.5 中的 g_i 和 h_i 分别是损失函数 ℓ 对预测值 $\hat{y}^{(t-1)}$ 的一阶偏导数和二阶偏导数：

$$g_i = \frac{\partial \ell(y_i, \theta)}{\partial \theta}\Big|_{\theta=\hat{y}_i^{(t-1)}}, \quad h_i = \frac{\partial^2 \ell(y_i, \theta)}{\partial \theta^2}\Big|_{\theta=\hat{y}_i^{(t-1)}} \tag{2.6}$$

在特定的划分路径下，叶子节点的个数是固定的，因此目标函数简化为：

$$\begin{aligned}\tilde{\mathscr{L}}^{(t)} &= \sum_{i=1}^{N} \left[g_i f_t\left(\mathbf{x}_i\right) + \frac{1}{2} h_i f_t^2\left(\mathbf{x}_i\right)\right] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \\ &= \sum_{j=1}^{T} \left[\left(\sum_{i \in I_j} g_i\right) w_j + \frac{1}{2}\left(\sum_{i \in I_j} h_i + \lambda\right) w_j^2\right] + \gamma T\end{aligned} \tag{2.7}$$

式 2.7 中的 I_j 是第 j 个叶子节点的样本索引集合。对式 2.7 求导，得到最优的叶子节点输出值：

$$w_j^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda} \tag{2.8}$$

将式 2.8 代入目标函数式 2.7，得到给定划分 q 时的最优损失（结构分数）为：

$$\tilde{\mathscr{L}}^{(t)}(q) = -\frac{1}{2} \sum_{j=1}^{T} \frac{\left(\sum_{i \in I_j} g_i\right)^2}{\sum_{i \in I_j} h_i + \lambda} + \gamma T \tag{2.9}$$

XGBoost 通过贪心算法搜索最优的划分 q，即最小化结构分数 $\tilde{\mathscr{L}}^{(t)}(q)$。在搜索最优划分的过程中，可构建直方图对特征值进行离散化，以提高搜索效率。

2.1.2 fastText

fastText 是由 Facebook 研究团队在 2016 年提出的一种轻量级的文本分类工具[41]。如图 2-1 所示，fastText 的模型架构包括输入层、隐藏层和输出层。输入层将文本序列通过词袋向量化和使用哈希技术的 N-gram 特征提取方法映射到固定长度的特征向量。在将特征向量进行标准化后，隐藏层采用线性激活函数，计算这些特征向量的平均值，作为文本表示向量。输出层采用 Softmax 函数作为激活函数，将文本表示向量映射到类别概率分布。

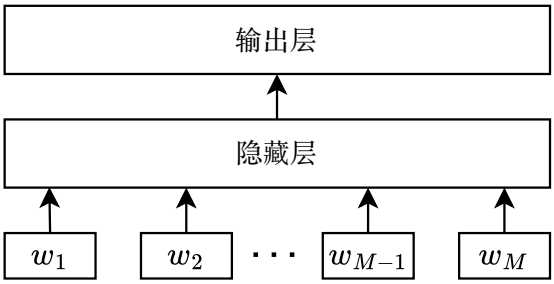


图 2-1 fastText 模型架构

对于分类任务，fastText 的损失函数采用负对数似然损失函数，即：

$$\mathscr{L} = -\frac{1}{N}\sum_{n=1}^{N} y_n \log\left(f\left(\boldsymbol{B}\boldsymbol{A}x_n\right)\right) \tag{2.10}$$

式 2.10 中，N 是样本数量；y_n 是样本 n 的真实标签；x_n 是样本 n 经过标准化后的特征向量均值；$\boldsymbol{A}$ 和 $\boldsymbol{B}$ 是隐藏层和输出层的权重矩阵；f 是 Softmax 函数。

得益于使用哈希技术的 N-gram 特征提取方法和分层 Softmax，fastText 可以在取得和其他深度学习模型相近的分类效果的同时，具有更快的训练速度和更小的模型体积[41]。本文关注的机器生成文本检测问题是一个二分类问题，故分层 Softmax 的优势并不显著。若后续拓展研究检测具体来源（例如预测值为生成该文本的机器名称等）、具体文本构造方式（例如预测值为单一来源、混合文本、机器润色等）等多分类问题，fastText 的分层 Softmax 技术更能为训练速度带来优势。

2.1.3 BERT 预训练模型

BERT (Bidirectional Encoder Representations from Transformers) 是由 Google 研究团队在 2018 年提出的一种预训练模型，其发布时对包括文本分类、命名实体识别等在内的 11 个自然语言处理任务取得了最佳结果[16]。

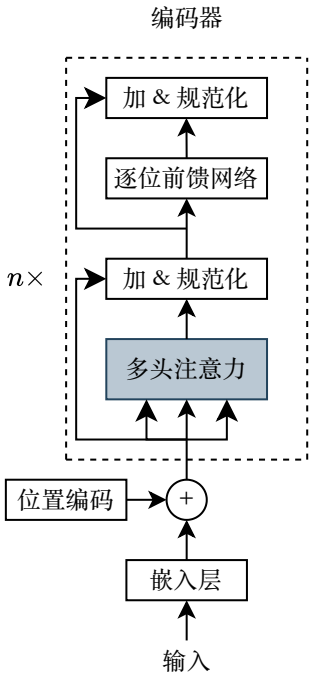


图 2-2　Transformer 编码器

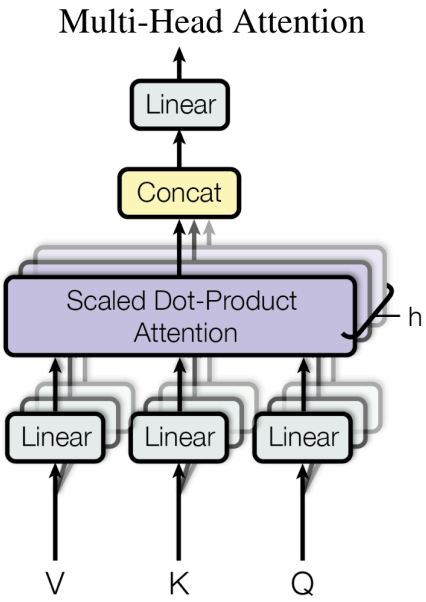


图 2-3　多头注意力

BERT 模型的核心部分是基于 Transformer[17] 的编码器。Transformer 是一个基于自注意力机制的深度学习模型，其编码器的结构如图 2-2 所示。嵌入层将输入序列的每个词转换为词嵌入向量 X。位置编码将词的位置信息加入词嵌入向量，使得模型可以推断词的相对或绝对位置，帮助模型更好地理解语序和语境。多头注意力模块使用了自注意力机制，将词嵌入向量 X 分别与三个可学习的权重参数矩阵 W^Q、W^K 和 W^V 相乘，得到查询、键和值的向量表示 Q、K 和 V，再计算缩放点积注意力：

$$\begin{aligned} Q &= XW^Q \\ K &= XW^K \\ V &= XW^V \\ \text{Attention}(Q,K,V) &= \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \end{aligned} \tag{2.11}$$

本文使用的 `bert-base-chinese` 模型的多头注意力机制中包含 12 个注意力头，每个头的输出向量长度为 $d_k = 64$。单头注意力的输出向量经过拼接后，通过一个全连接层，得到多头注意力模块的 $12 \times 64 = 768$ 维的输出向量。通过使用多头注意力，模型可以在不同的表示子空间中并行地学习信息，这使得多头注意力比仅使用单头能够捕捉到更丰富的语义信息，从而提高了模型处理复杂任务的能力。

经过多头注意力模块后，残差连接和层规范技术有助于减缓梯度消失和爆

炸问题，提高模型的训练稳定性和效率。逐位置前馈网络使用了两个线性全连接层和非线性的 GELU 激活函数，以增强模型的拟合能力。整个编码器的输出向量经过 n（在 `bert-base-chinese` 模型中，$n = 12$）层堆叠后，得到文本表示向量，作为下游任务的输入。

构建 BERT 模型的过程分为预训练阶段和微调阶段。本文使用的 bert-base-chinese 模型在约 2500 万条中文维基百科语句上进行了预训练。在微调阶段，由于本文研究的问题是一个二分类任务，故编码器后的输出层采用了 768×2 的全连接层，最终输出两个类别的逻辑值。

2.2 分类效果评价指标

本节将介绍本研究使用的文本分类效果的评价指标，包括准确率、精确率、召回率、F1 值和 AUC 值。这些评价指标可以从不同角度反映模型的有效性和泛化能力。

对于一个二分类问题，其混淆矩阵如表 2-1 所示。其中，TP (True Positive) 表示真正例的数量；FP (False Positive) 表示假正例的数量；FN (False Negative) 表示假负例的数量；TN (True Negative) 表示真负例的数量。本研究中的正例为机器生成的文本，负例为人类生成的文本。

表 2-1 混淆矩阵

		预测值	
		正例	负例
真实值	正例	TP	FN
	负例	FP	TN

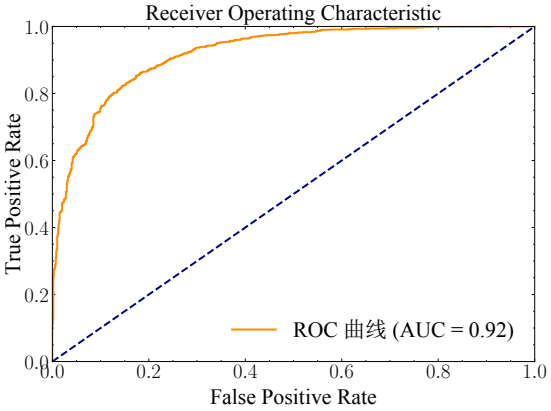


图 2-4 ROC 曲线与 AUC 值

表 2-2 展示了本文使用的评价指标及其计算方法：准确率（Accuracy）是分类正确的样本数量占总样本数量的比例；精确率（Precision）是分类为正例的样本中真正例的比例；召回率（Recall）是真正例中被分类为正例的比例，本研究细分为机器生成文本召回率（将机器生成的文本视为正例）和人类生成文本召回率（将人类生成的文本视为正例），并计算两者的均值作为平均召回率；F1 值是精确率和机器生成文本召回率的调和平均数。

在将样本的预测概率转换为最终的分类标签时，不同的阈值标准会导致不同的分类结果。上述评价指标的计算中，都是将阈值标准设为 0.5，即选择预测概率较大的类别作为最终的分类标签。若阈值标准越高，则模型更容易将样本分类为负例（人类生成的文本）。受试者工作特征（Receiver Operating Characteristic，ROC）曲线是如图 2-4 所示的以假正例率（False Positive Rate）为横轴、真正例率（True Positive Rate）为纵轴的曲线，AUC 值即是 ROC 曲线下的面积，它是一个不依赖于特定阈值标准的评价指标。AUC 的计算方法如式 2.12 所示：

$$AUC = \frac{1}{m^+ m^-} \sum_{x^+ \in D^+} \sum_{x^- \in D^-} \left(I\left(f\left(x^+\right) > f\left(x^-\right) \right) \right) \tag{2.12}$$

式 2.12 中，m^+ 和 m^- 分别是正例和负例的数量；D^+ 和 D^- 分别是正例和负例的样本集合；$f(x)$ 是样本 x 的预测概率；$I(\cdot)$ 是指示函数，若括号中的条件成立则返回 1，否则返回 0。因此，AUC 可以理解为分类模型将真正例样本预测为正例的概率大于将真负例样本预测为正例的概率的可能性。AUC 值越大，说明模型的分类效果越好。

表 2-2　分类效果评价指标

评价指标	含义	计算方法
Accuracy	准确率	$(TP+TN)/(TP+FP+FN+TN)$
Precision	精确率	$TP/(TP+FP)$
Recall-M	机器生成文本召回率	$TP/(TP+FN)$
Recall-H	人类生成文本召回率	$TN/(TN+FP)$
Recall-Avg	平均召回率	(Recall-M + Recall-H)/2
F1 Score	F1 值	$2\times TP/(2\times TP+FP+FN)$
AUC	ROC 曲线下面积	见式 2.12

第 3 章　数据构造

3.1　数据集构造方法

本文首先基于 Guo et al. (2023)[33] 整理的中文人类-ChatGPT 问答对比语料集（HC3-Chinese），提取其中的人类生成文本。这些人类生成文本主要有两个来源：一是公开可用的问答数据集，这些数据集中的答案由特定领域的专家给出，或是网络用户投票选出的高质量答案；二是从维基百科和百度百科等资料中构造的“概念 - 解释”问答语句对。

人类生成文本数据集中包含百科、金融、法律、医疗、检索问答、通用问答、心理咨询共 7 个领域的问答数据，各领域的人类生成文本构造方式如表 3-1 所示。

表 3-1　人类生成文本构造方式

领域	人类生成文本构造方式
baike	使用“我有一个计算机相关/信息科学相关的问题，请用中文回答，什么是 < 词条 >？”作为问题，从百度百科采集相关词条数据作为人类生成文本。
finance	采样自 ChineseNlpCorpus 整理的“金融知道”数据集。链接：`https://github.com/SophonPlus/ChineseNlpCorpus`
law	采样自 LegalQA 整理的法律问答数据集。链接：`https://github.com/siatnlp/LegalQA`
medicine	采样自 He et al.[42] 整理的 MedDialog-CN 医疗对话数据集。
nlpcc_dbqa	采样自 Duan[43] 整理的自然语言处理数据集。
open_qa	采样自 Xu[44] 整理的 WebTextQA 和 BaikeQA 数据集。
psychology	采样自飞桨 AI Studio 发布的中文心理问答数据集。链接：`https://aistudio.baidu.com/datasetdetail/38489`

对于机器生成的文本，本文没有直接使用 HC3-Chinese 中由 ChatGPT 生成的文本，而是将人类生成文本对中的问题作为提示词（Prompt），通过调用大语言模型的 API 接口得到机器生成的文本。这些大语言模型包括 GPT-3.5、Spark、ChatGLM、Ernie 和 QWen 共 5 个大语言模型，具体信息如表 3-2 所示。除了

GPT-3.5 外，其余 4 个大语言模型均由中国企业开发。根据 Xu et al. (2023)[45] 对各个大语言模型的基准测试，这 5 个大语言模型在中文领域均有较好的表现。

表 3-2　大语言模型及其具体信息

大语言模型	开发者	发布日期	参数规模
GPT-3.5	OpenAI	2022 年 11 月	1750 亿
Spark	科大讯飞	2023 年 8 月	未知
ChatGLM	智谱 AI	2023 年 3 月	60 亿
Ernie	百度	2023 年 3 月	2600 亿
QWen	阿里巴巴	2023 年 8 月	70 亿

HC3-Chinese 数据集中包含百科、金融、法律、医疗、检索问答、通用问答、心理咨询共 7 个领域的问答数据。将各个领域中的问题作为提示词，调用上述 5 个大语言模型的 API 接口得到机器生成文本。在调用 API 接口生成数据的过程中，某些提示词中可能包含不道德、不健康或较敏感的信息，导致大语言模型认为不适合回答而返回错误。此外，观察到大部分问题都能在 1 分钟内完成回答，为了避免生成数据的程序耗时过长，本文将调用时间到达 1 分钟但还未得到问题答案的视为超时。这两种情况均被视为机器在生成文本的过程出现错误。对第一轮机器生成文本过程出现错误的问题，再进行第二轮生成，若第二轮仍然出现错误，则最终缺失该条机器生成文本。

表 3-3　各领域下的文本数量和生成文本错误率

领域	问题	人类生成文本	机器生成文本	生成文本错误率
baike	4617	4616	22934	0.65%
finance	689	1560	3411	0.99%
law	372	690	1821	2.10%
medicine	1074	1074	5198	3.20%
nlpcc_dbqa	1709	1709	8340	2.40%
open_qa	3293	7362	16059	2.47%
psychology	1099	5220	5332	2.97%
总计	12853	22231	63095	1.82%

本文调用大语言模型 API 构建数据集的过程共耗时 1 周。表 3-3 展示了最终构建的数据集中各个领域下的文本数量和生成文本错误率。可以看出，在医疗和

心理咨询这两个敏感问题较多领域下的生成文本错误率相对较高，分别为 3.20% 和 2.97%，而绝大部分问题都能由机器顺利生成文本。从整体上看，该数据集包含较丰富的人类和机器生成文本，总计共 8 万余条，且涵盖了多个领域的文本，适合用于文本分类模型的训练和测试。

3.2 文本统计特征

人类和机器生成的文本在一些统计特征上可能存在差异，这些差异也许可以帮助区分人类和机器生成的文本。根据 Gehrmann et al.[32] 和 Guo et al.[33] 等前人研究，本文计算了文本长度、句子数量、单句长度均值、单句长度标准差、用词丰富度和文本困惑度这 6 个统计特征，这些特征在较多文献中展现了较好的分类能力。

首先，统计文本中包含的字符数量，作为文本长度。然后，将文本以逗号、句号、感叹号和问号等用于停顿的标点符号（不区分全角和半角）进行划分，得到若干个短句，将这些短句的数量作为句子数量。单句长度均值和标准差分别是所有短句长度的均值和标准差。

用词丰富度是不同字符的数量占总字符数量的比例，即 $D = 100 \times \frac{V}{L}$，其中 V 为不同字符的数量，L 为总字符数量。该值越大，说明文本中不同的字符越多，用词丰富度越高。

文本困惑度的计算如式 3.1 所示。基于一个语言模型，根据上下文推断文本中每个词的负对数似然值，再对每个位置的负对数似然值取平均值，最后取自然指数函数得到文本困惑度。文本困惑度越大，说明文本中的词语越不符合语言模型的预期，也意味着文本越难以理解。本文选用的语言模型是 chinese-bert-wwm-ext 预训练模型[46]，它在 54 亿词数的中文语料上使用全词掩码技术（Whole Word Masking）进行预训练，因此十分适合中文掩码预测任务。

$$\begin{aligned} \text{Loss}_i &= -\log(p(w_i)) \\ \text{PPL} &= \exp\left(\frac{\sum_{i=1}^{N} \text{Loss}_i}{N}\right) \end{aligned} \tag{3.1}$$

图 3-1 和图 3-2 展示了人类和机器生成文本在这 6 个统计特征上的分布情况。从图中可以看出，机器生成文本普遍更长，这主要是由于其包含的句子数比人类更多，其次是单句长度更长。从单句长度标准差来看，机器生成文本的波动性普遍更大，这意味着机器生成文本中的句子长度差异更大，即长短句交替出现的情况更多。从用词丰富度来看，人类生成文本的用词更丰富，而机器生成的文本更像是相同词语的反复使用。文本困惑度是人类生成文本和机器生成文本之间差异最大的统计特征。在所有的领域上，人类生成文本均有显著高于机器

生成文本的困惑度，反映了机器更倾向于生成通顺、连贯的文本。

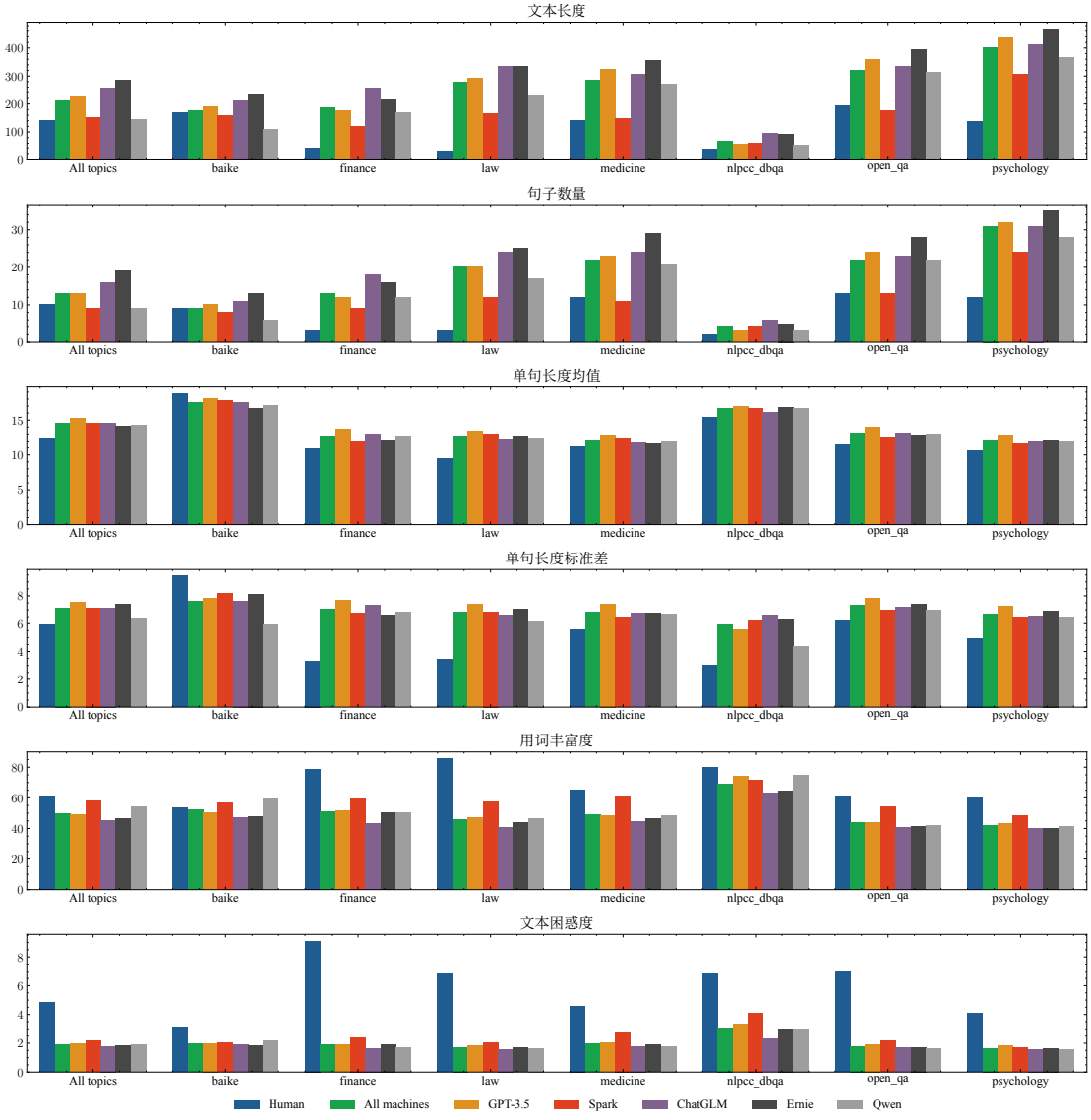


图 3-1　不同领域和机器的文本统计特征

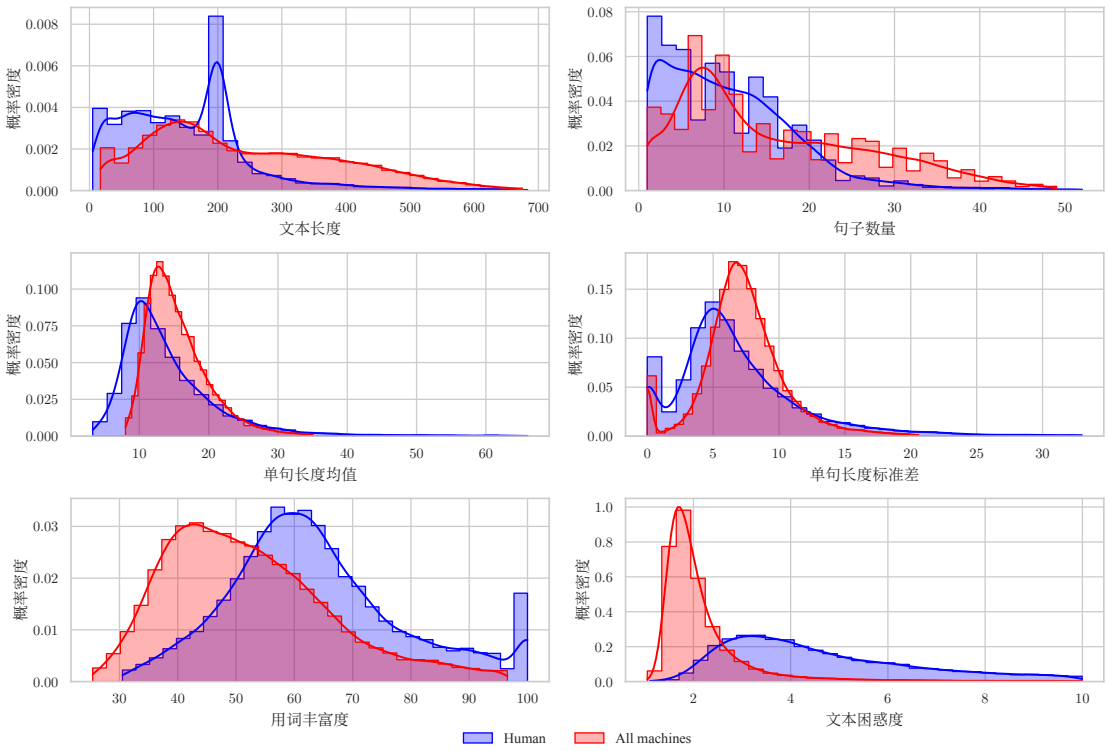


图 3-2 文本统计特征分布

人类和机器生成的文本在统计特征上存在一定的差异，并且这些差异主要体现在人类生成文本和机器生成文本之间，而不同机器生成的文本之间的差异相对较小。这为区分人类和机器生成的文本提供了初步的依据，后续的实证分析也将考虑这些统计特征构建机器学习分类模型。

3.3 数据集划分

为了检验文本分类模型在不同领域、不同大语言模型生成的文本上的性能，本文按照是否来自相同领域、是否为相同模型生成这两个标准，设计了 4 种不同的实验组合，如表 3-4 所示。

表 3-4 根据领域和模型划分实验组合

是否来自相同领域	是否为相同模型生成	实验个数
是	是	$5 \times 7 = 35$
是	否	7
否	是	5
否	否	1

随着领域和模型变得不同，实验场景对分类模型的要求也变得更高。从图

3-1 也可看出：对于来自相同领域、由相同模型生成的文本，数据内部的文本特征分布应该更加相似，因此理论上分类模型的性能也应该更好。而不同领域、不同模型生成的文本可能具有更大的差异，这为分类模型带来了更大的挑战。

在将数据集细分为上述的实验组合后，本文将每个实验组合的数据集按照 8:1:1 的比例划分为训练集、验证集和测试集。训练集用于模型拟合参数，验证集用于调整模型的超参数和判断是否需要早停，测试集用于评估模型的性能。

第 4 章　单一来源文本检测

单一来源文本检测本质上是一个文本二分类问题，即识别一段文本是由人类生成或是机器生成的。文本分类任务自然语言处理的一个被广泛研究的领域，现有文献提出了多种的文本分类方法，在机器生成文本检测方面亦有大量应用。

目前中文领域的机器生成文本检测相对较少，何种模型检测效果较好、跨领域和跨模型的语料会对模型检测效果带来何种影响等，这些问题尚未被充分讨论。因此，本章首先考察机器生成中文文本检测在跨领域和跨模型的场景下的检测难度是否也更大，再在难度最大的检测问题上对多种模型的检测效果和优缺点进行对比，为机器生成中文文本检测的研究作实证补充。

4.1　fastText 模型检验不同实验组合的难易度

我们首先使用 fastText 这一基于词袋模型和 N-gram 的分类器，它具有效果较好且训练高效的特点，常在自然语言处理研究中作为基线模型。对于表 3-4 中的 48 个实验，以 F1 值为目标，基于 fastText 的 autotune 方法自动搜索最优学习率和 N-gram 的 N 值等超参数，最长搜索时间为 60 秒。将每个实验组合下的所有实验结果合并后计算分类效果评价指标，以衡量该实验组合下的模型表现。

表 4-1　不同实验组合下的模型表现

领域	模型	Accu.	Prec.	Recall-M	Recall-H	Recall-Avg	F1	AUC
相同	相同	95.26	96.12	90.46	97.95	94.21	93.20	0.97
相同	不同	95.63	95.61	**98.60**	87.29	92.95	97.08	0.99
不同	相同	**96.17**	96.94	92.33	**98.34**	**95.34**	94.58	0.98
不同	不同	96.13	**97.24**	97.55	92.09	94.82	**97.39**	**0.99**

表 4-1 显示了不同实验组合下的模型表现。在不同领域和不同模型这一实验组合下，模型在精确率、F1 值和 AUC 指标上的表现最好，不同领域、相同模型这一实验组合下，也取得了最优的准确率和平均召回率。这与我们在 3.3 节中“越复杂的实验组合下，分类任务越困难”的预期不一致。

考虑到随着领域和模型变得不同，实验组合中的数据量也会自然增大，因此数据量这一因素可能对模型的表现产生了影响。因此，接下来我们通过下采样

的方法，控制每个实验组合中训练集样本量最多为 4000，接近于相同领域、相同模型这一实验组合下的平均数据量。

从表 4-2 中可以看出，当数据量被控制在相同的水平时，不同领域、不同模型这一实验组合下的模型表现最差。这一观察符合我们的预期，即当某些特征分布差异较大时，从这些差异较大的特征中学习规律的难度更大，分类任务变得更具有挑战。

表 4-2　不同实验组合下的模型表现 - 控制数据量

领域	模型	Accu.	Prec.	Recall-M	Recall-H	Recall-Avg	F1	AUC
相同	相同	94.03	91.52	91.91	95.22	**93.57**	91.72	0.97
相同	不同	**95.38**	**96.28**	**97.51**	89.43	93.47	**96.89**	**0.98**
不同	相同	94.03	93.34	89.93	**96.36**	93.14	91.60	0.97
不同	不同	87.87	88.84	95.63	65.69	80.66	92.11	0.93

同时，我们也从各实验组合下的模型表现中注意到两个现象。一是相同领域、不同模型这一实验组合在大部分评价指标上的表现最好，说明将不同模型生成的文本混合在一起作为训练数据时，分类模型能够从 GPT-3.5、Spark 等多个模型生成的文本中学习到机器生成文本的特征，这能够帮助增强分类模型的泛化能力。

二是机器生成文本召回率和人类生成文本召回率较不平衡，具体体现在：若训练数据均来自相同模型生成的文本，则不论是否将不同领域的数据进行混合，都观察到机器生成文本召回率相对偏低；若训练数据为不同模型生成的文本混合而成，则人类生成文本召回率明显更低。结合表 3-3 中的人类生成文本和机器生成文本数量，分析各实验组合下的数据构成可知：当机器生成的文本仅来自一个模型时，由于某些问题下收集到的人类生成文本不止一条，因此人类生成文本数量会略多；当机器生成文本来自不同模型时，由于数据集中包含 5 个模型生成的文本，因此机器生成文本数量会是人类生成文本数量的数倍。在训练数据中，若某种标签的数据量更大，则分类模型更倾向于将样本预测为该标签，这可能是造成召回率不平衡的原因。

为消除样本不平衡的影响，我们接下来在控制训练集样本量最多为 4000 的基础上，随机抽取相同数量的人类和机器生成文本。若某一实验场景下的样本量不足 4000，则所有样本全部抽取。表 4-3 显示了在样本平衡的情况下，不同实验组合下的模型表现。

从表 4-3 中可以看出，当训练集样本量被控制在相同的水平且样本平衡时，在人类和机器生成文本上的召回率较为接近。从表 4-3 、图 4-1 图 4-2 中对比四

表 4-3 不同实验组合下的模型表现 - 控制数据量和平衡样本

领域	模型	Accu.	Prec.	Recall-M	Recall-H	Recall-Avg	F1	AUC
相同	相同	**94.04**	93.45	89.71	96.47	93.09	91.54	**0.98**
相同	不同	93.59	**98.91**	**92.33**	**97.15**	**94.74**	**95.50**	0.97
不同	相同	92.10	87.40	91.37	92.51	91.94	89.34	0.96
不同	不同	88.28	94.95	88.91	86.48	87.70	91.83	0.93

种不同的实验组合，前文讨论的结论仍然成立：不同领域、不同模型这一实验组合下，由于特征分布差异较大，分类任务变得更加困难，分类效果最差。

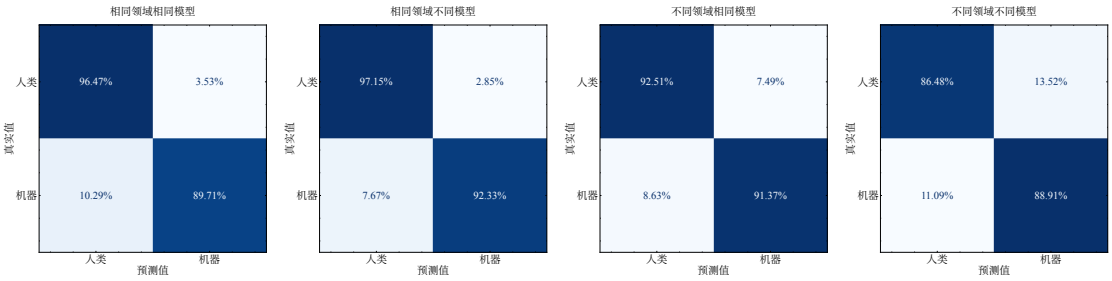


图 4-1 不同实验组合下的混淆矩阵

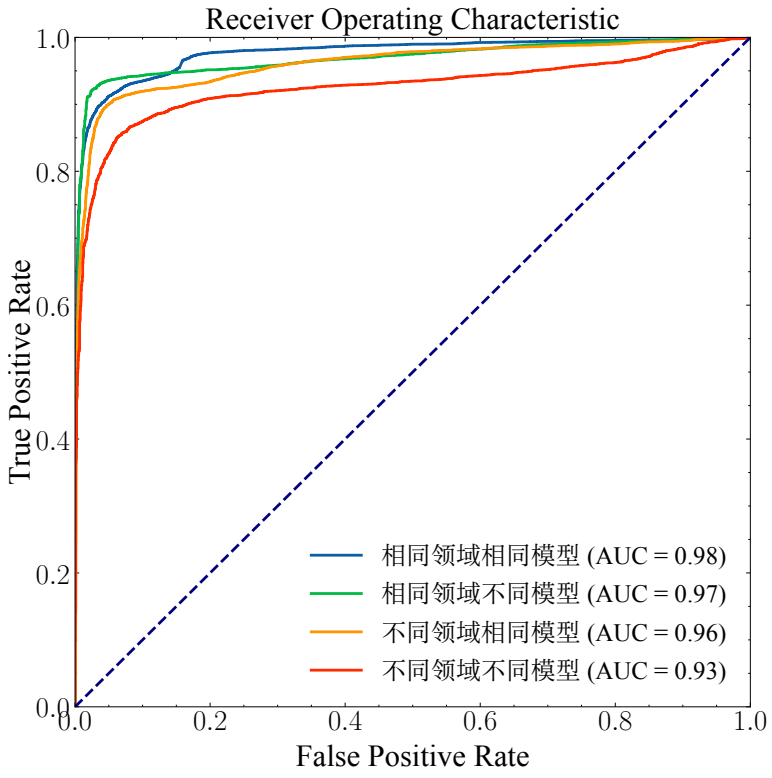


图 4-2 不同实验组合下的 ROC 曲线与 AUC 值

在后续的实验中，我们针对不同领域、不同模型这一最复杂的组合进行实验，考察各分类模型在该实验组合下的表现及可解释性分析等。

4.2　基于文本统计特征的 XGBoost 分类模型

在 3.2 节中，计算了文本长度、句子数量、单句长度均值、单句长度标准差、用词丰富度和文本困惑度这 6 个统计特征，且从这些特征的分布中观察到人类和机器生成的文本存在差异，尤其是在困惑度指标上最为明显。因此，这启发我们可以直接使用这些统计特征进行分类建模，并探究各特征的预测能力和重要性等。下面我们使用 XGBoost 分类模型，将损失函数设为如式 4.1 所示的二元交叉熵损失，以文本统计特征作为输入，对不同领域、不同模型这一实验组合下的数据进行分类。

$$\mathscr{L} = -\sum_{i=1}^{N} \left[y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i) \right] \tag{4.1}$$

为选择较合适的模型超参数，我们使用网格搜索的方法，设置超参数的候选值如表 4-4 所示。

表 4-4　XGBoost 超参数候选值

超参数	候选值
最大树深度	[2, 4, 6, 8, 10]
基学习器个数	[100, 200, 300, 400, 500]
学习率	[0.01, 0.05, 0.1, 0.2, 0.3]

在 3 折交叉验证的超参数搜索后，得到最优的 XGBoost 分类模型。最终 XGBoost 的分类效果如表 4-5 所示。第一行为前文使用的 fastText 在同样数据集下训练的分类效果。第二行展示了使用 6 个统计特征的 XGBoost 分类效果，在 F1 值、AUC 等指标上比 fastText 略高。第三行为去除文本困惑度这一特征后的分类效果，可以看出，当去除文本困惑度这一特征后，所有评价指标均有较为明显的下降。这说明文本困惑度这一特征对分类模型的预测能力有较大的贡献。

为定量刻画各特征对分类模型的贡献，我们使用 XGBoost 的特征重要性和 SHAP（SHapley Additive exPlanations）[47] 方法对 XGBoost 分类模型进行解释。

XGBoost 基于各特征的如式 2.9 所示的结构分数带来的增益，计算节点分裂前后结构分数的差值，并考虑节点分裂带来的的惩罚项，如式 4.2 所示。其中 $G_I = \sum_{i \in I} g_i$ 和 $H_I = \sum_{i \in I} h_i$ 分别为节点 I 中所有样本的一阶导数和二阶导数

表 4-5 XGBoost 分类效果

	Accu.	Prec.	Recall-M	Recall-H	Recall-Avg	F1	AUC
fastText	88.28	94.95	88.91	86.48	87.70	91.83	0.93
XGboost	89.69	92.29	93.22	81.12	87.17	92.76	0.95
-Perplexity	82.30	83.96	92.72	57.03	74.88	88.12	0.86

之和，L 和 R 分别代表左右子节点，λ 和 γ 均为正则化项的系数。结构分数的增益越大，说明该特征对模型的预测贡献越大。

$$Gain = \frac{1}{2}\left[\frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\right] - \gamma \tag{4.2}$$

图 4-3 展示了 6 个统计特征的特征重要性。绝大部分对预测效果的贡献都来自文本困惑度这一特征，占比达 62.63%，其余特征的贡献均接近或不到 10%。这与我们在 3.2 节中的观察也较为一致，即文本困惑度这一特征在人类和机器生成的文本中的分布差异最为明显，因此也是分类模型预测能力最大的来源。

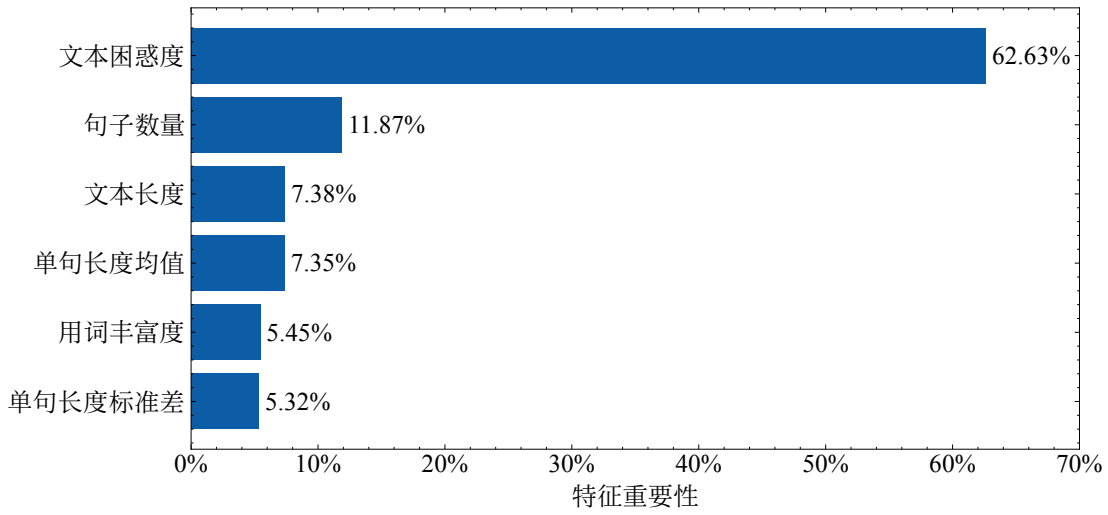


图 4-3 XGBoost 特征重要性

SHAP 是一种基于博弈论的机器学习解释性分析方法，它计算每个特征对模型预测的边际贡献，平均在所有可能的特征组合中，并引入了多种优化和近似算法来高效地计算近似。与 XGBoost 计算的特征重要性不同的是，SHAP 值不仅反映了该特征对模型预测结果贡献的绝对大小，还反映了贡献的方向。

图 4-4 展示了 6 个统计特征的 SHAP 值。图中每个圆点代表一个样本；颜色代表特征值的大小，红色代表特征值越大，蓝色代表特征值越小；横坐标代表 SHAP 值，横坐标越大代表让模型越有可能预测为机器生成的文本，横坐标越小

代表让模型越有可能预测为人类生成的文本。从图中可以看出，文本困惑度这一特征对模型的预测有最大的贡献，且展现出明显的单调性，即文本困惑度极低的样本绝大部分都是机器生成的文本，文本困惑度高的样本更可能是人类生成的文本。

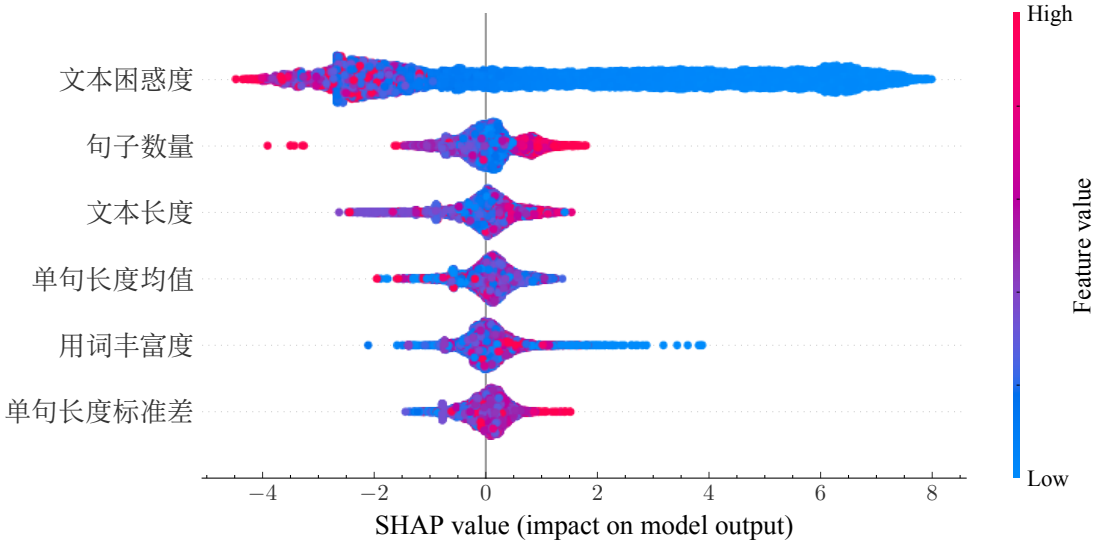


图 4-4　各统计特征的 SHAP 值

4.3　BERT 分类模型

第 4.1 节和第 4.2 节均是通过构建特征后使用机器学习分类器进行分类。在本节中，我们使用 BERT 预训练模型进行端到端的文本分类，它不需要人工构建特征，而是直接使用原始文本作为输入，通过分词、编码器等网络层，再通过全连接层进行分类。

本文使用的 BERT 预训练模型为中文版的 `bert-base-chinese` 模型。相关硬件和软件环境如表 4-6 所示。

表 4-6　实验环境

环境	配置
CPU	Intel Xeon Gold 6330 @2.10GHz
GPU	NVIDIA GeForce RTX 3090 24GB
操作系统	Ubuntu 20.04
Python	3.9.12
PyTorch	2.1.2

由表 3-3 的数据描述和表 4-2 的实验结果可知，训练数据中的人类和机器生成文本数量不平衡，大致比例为 1:3，这可能会导致人类和机器生成文本的召回

率差异较大。在 BERT 模型中，我们使用带有权重的交叉熵损失，即对于人类生成文本和机器生成文本赋予 3:1 的权重，以缓解样本不平衡的问题。训练器为 AdamW，学习率为 5×10^{-5}，训练轮数为 3 轮，批大小为 32。

BERT 模型的分类效果如表 4-7 所示。BERT 模型在所有评价指标上的绝对数值都超过 96%，相比 fastText 和 XGBoost 都有大幅的提升。这与 Li et al. (2023)[34] 使用多种模型进行分类的实验结果相一致。

表 4-7 BERT 分类效果

Accu.	Prec.	Recall-M	Recall-H	Recall-Avg	F1	AUC
97.94	98.82	98.39	96.65	97.52	98.6	0.9978

4.4 使用 ChatGPT 识别人类和机器生成文本

最后，作为检测大模型能否分辨人类和其自身生成文本的一种尝试，我们随机选取 1000 条测试文本，直接使用 GPT-3.5 模型进行零样本学习，输出每条文本是否为机器生成。

具体地，本文使用的提示词为“Is this text generated by a large language model?”借助 `marvin` 这一辅助校准机器回答格式的工具，我们可以得到 GPT-3.5 对某一条文本是否由机器生成的 `True` 或 `False` 的回答。

表 4-8 ChatGPT 分类效果

Accu.	Prec.	Recall-M	Recall-H	Recall-Avg	F1	AUC
67.8	63.82	82.2	53.4	67.8	71.85	0.68

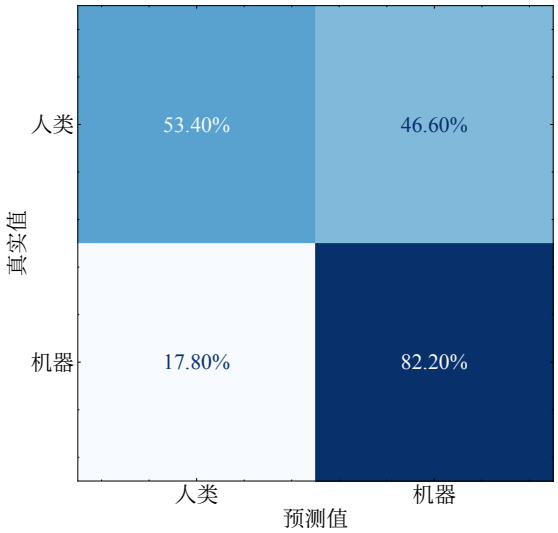


图 4-5　ChatGPT 分类结果混淆矩阵

最终，GPT-3.5 对 1000 条测试文本的分类效果如表 4-8 和图 4-5 所示。ChatGPT 作为大语言模型，分辨人类和自身生成的文本效果比随机猜测要好，但其劣势体现在人类生成文本召回率较低，即容易将人类生成的文本预测为机器生成的文本。此外，由于大语言模型生成结果存在一定的随机性，同样的文本也可能得到不同的分类结果。因此，使用 ChatGPT 作为机器生成文本检测器并不如基于训练数据得到的机器学习模型可靠。

4.5　机器生成文本检测模型总结

本章首先使用 fastText 在根据是否跨领域、是否跨模型生成划分的数据集上进行了多个实验，在统一数据量大小和平衡正负样本比例后，发现跨领域和跨模型的文本检测任务难度最大，模型表现效果最差。接着，我们使用 XGBoost 模型和 BERT 模型以及 ChatGPT 这一本用于生成机器文本的大语言模型进行了机器生成文本检测。这些模型的使用方式、分类效果等各有不同，其优缺点可总结如表 4-9 所示。

本章的实验结果丰富了中文语料下的机器生成文本检测研究的实证支持，主要回答了“跨领域和跨模型生成的文本对检测效果有何影响”和“何种检测模型表现较好”这两个问题。在后文的研究中，我们将使用 BERT 这一分类效果最好的模型进行更深入的实验。

表 4-9 机器生成文本检测模型总结

检测模型	优点	缺点
XGBoost	训练速度较快；可进行特征可解释性分析，方便理解模型决策依据。	需要事先人为构造量化的文本特征，特征质量对模型表现影响较大。
fastText	训练速度较快；无需人为构造特征。	不便于进行可解释性分析。
BERT	分类效果最好；可基于词或句进行可解释性分析。	模型训练时间长，推理消耗内存大。
ChatGPT	无需训练数据，可直接进行零样本学习。	检测效果较差；检测结果不稳定可靠。

第 5 章　混合来源文本检测

第 4 章使用多种分类模型对单一来源的人类和机器生成文本进行检测，并在测试集上取得了较好的效果，多个评价指标都达到 95% 以上。单一来源文本的生成者为纯人类或纯机器，即一整段文本由同一来源生成。然而，在现实使用场景中，人们通常不会完全使用整段机器生成文本，而是会采用如补全、续写等方式，将人类编写的文本与机器生成文本混合，即一段文本中既包含人类生成的部分，又包含机器生成的部分。因此，本章将研究混合来源文本检测问题。

5.1　构造混合来源文本

为了模拟现实使用场景中机器续写文本的方式，本文利用纯人类生成的文本构造混合来源文本。考虑到长度过短的文本表达的语义可能不够完整，在要求机器续写时的效果不佳，因此首先根据逗号、句号、感叹号、问号等表示停顿和断句的标点符号，筛选出至少包含 4 句话的纯人类生成文本，共 8741 条。然后，将后一半的句子用 `[MASK]` 代替，使用提示词“Continue writing the given text. Only return the new text.”要求大语言模型进行续写。最后，将原始文本与续写文本拼接在一起，构造混合来源文本。

表 5-1　GPT-4-32K 续写样本示例

原始文本	GPT-4-32K 续写文本
时间是否够用？取决于你想做的事情所需要的时间以及时间的总量，可能跟事情的难易，做事情的效率等很多因素都有关系。其实大家的时间总量都是一样的，当我们想做的事情太多，不足以在我们的时间内完成的时候，就会觉得不够用。	因此，时间管理变得至关重要。好的时间管理能帮助我们更有效地利用时间，完成更多的工作。首先，我们需要明确优先级，了解哪些事情需要首先完成，哪些可以稍后处理。其次，提升效率，减少因为低效率而浪费的时间。使用一些专门的时间管理工具，比如番茄时钟法、四象限管理法等也可以帮助提升我们的时间使用效率。智能的使用时间，让我们不再感到时间不够用。

在使用 GPT-3.5 和其他国产大模型进行续写时，得到的文本有时会包含原始文本，或仅仅是对原始文本的重复，因此续写的效果可能较差。本文最终用于续写的模型为 GPT-4-32K，它是目前最强大的大语言模型之一，其续写的样本示例如表 5-1 所示。

本文亦曾尝试过将人类生成文本的任意句子替换为 `[MASK]`，再要求大语言模型进行补全，但观察到这种方式生成的文本通常会改写其他位置的文本，这并不严格符合补全的要求，因此最终没有使用补全这一方式。

图 5-1 展示了续写文本与原文本的长度差异。在没有将长度要求考虑进提示词的情况下，得到的机器续写文本比原文本的字符个数之差均值约为 90.4，标准差约为 103.5，计算配对 t 检验统计量约为 81.7，说明机器续写文本比原文本显著更长。我们统计了原文本的长度均值约为 247，这也意味着混合来源文本中大致有 58% 的文本是机器续写的。

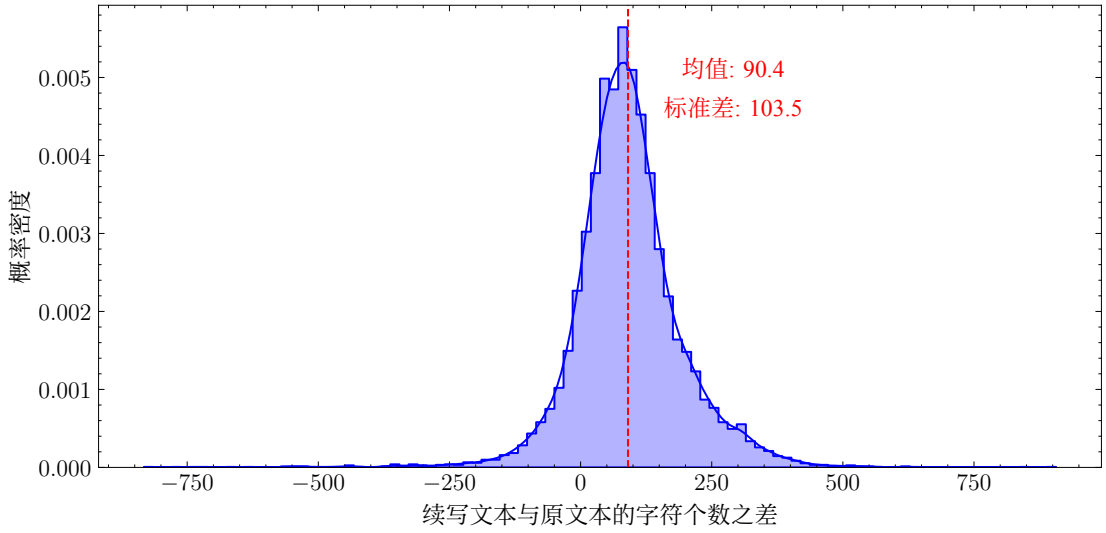


图 5-1　续写文本与原文本的长度差异

5.2　人类检测机器生成文本

在得到混合来源文本后，本文通过问卷调查的形式，邀请约 100 名志愿者对机器生成文本进行检测。问卷池包含 50 条人类生成文本、25 条纯机器生成文本和 25 条混合来源文本。每位志愿者需对问卷池中的随机 5 条样本进行分类，即判断某段文字是否全部为或部分包含机器生成的文本。

经统计，人类检测机器生成文本的分类效果如表 5-2 所示。可以发现，人类对机器生成文本的检测效果并不理想，各项分类指标均接近于随机分类，远不如前文所构建的各分类模型。在对混合文本的检测上，人类检测机器生成文本的分类效果更差，说明从混合文本中识别出机器生成的部分是更加困难的。

表 5-2 人类检测机器生成文本的分类效果

测试范围	Accu.	Prec.	Recall-M	Recall-H	Recall-Avg	F1	AUC
单一来源文本	48.63	34.66	53.04	46.26	49.65	41.92	0.5
混合来源文本	46.5	31.95	46.96	46.26	46.61	38.03	0.47
差异	-2.13	-2.71	-6.08	0	-3.04	-3.89	-0.03

5.3 BERT 模型检测混合来源文本

首先，我们直接使用第 4.3 节中使用单一来源文本训练得到的 BERT 模型对混合来源文本进行检测。表 5-3 中的第二行为混合来源文本下的实验结果，其中人类生成的文本仍为单一来源，但机器生成的文本为混合来源。可以看出，人类生成文本召回率仍为 96.65%，但机器生成文本召回率从 98.39% 下降到 61.83%，降幅达 36.56%，导致除人类生成文本召回率外，所有分类效果评价指标都有下降，这说明 BERT 模型在混合来源文本上的检测效果较差。

表 5-3 BERT 模型在混合来源文本上的分类效果

测试范围	Accu.	Prec.	Recall-M	Recall-H	Recall-Avg	F1	AUC
单一来源文本	97.94	98.82	98.39	96.65	97.52	98.6	0.9978
混合来源文本	87.24	87.26	61.83	96.65	79.24	72.38	0.9061
差异	-10.7	-11.56	-36.56	0	-18.28	-26.22	-0.0917

仅在单一来源文本上训练的 BERT 模型容易将混合来源的文本误判为人类生成文本，这可能是因为模型还没有学习过“混合来源的文本应被视为机器生成的”这一规则。因此，接下来我们将混合来源的文本加入训练集，重新训练 BERT 模型，训练和测试所使用的数据构成如表 5-4 所示。

表 5-4 BERT 模型检测混合来源文本的数据构成

	纯人类生成文本	纯机器生成文本	机器续写混合文本
训练集	17815	50445	7031
验证集	2204	6329	890
测试集	2212	6321	820

基于表 5-4 中包含了机器续写混合文本的训练数据，我们得到了新的 BERT 模型，将其在测试集上进行检测，结果如表 5-5 所示。第一行是模型仅在单一来

源文本上测试的表现；第二行是在人类生成文本和机器续写混合文本上的表现，即排除了纯机器生成的文本。

表 5-5　基于混合文本训练的 BERT 模型分类效果

测试范围	Accu.	Prec.	Recall-M	Recall-H	Recall-Avg	F1	AUC
单一来源文本	96.51	96.89	98.45	90.96	94.7	97.66	0.9618
混合来源文本	91.29	79.08	92.2	90.96	91.58	85.14	0.9545
差异	-5.22	-17.81	-6.25	0	-3.12	-12.52	-0.0073

将表 5-5 中的第二行与表 5-3 中的第二行进行对比，可以看出：在将机器续写混合文本加入训练数据之后，模型在人类生成文本召回率这一指标上有所下降，但在机器生成文本召回率这一指标上有所上升。这说明训练集中的机器续写混合文本能够引导模型将混合文本识别为机器生成的文本。

然而，即使加入了机器续写混合文本作为训练数据，对比表 5-5 中的第一行与表 5-3 中的第一行后可以看出，模型在混合来源文本上的检测效果仍然不如在单一来源文本上的检测效果。这说明混合来源文本的检测问题相对于单一来源文本的检测问题更为困难，仅通过基于混合文本的训练并不能很好地解决这一问题。

5.4　基于注意力和分类效果筛选的数据增强算法

5.4.1　算法介绍

为提高模型对于混合文本检测的效果，本文考虑使用数据增强方法扩充训练数据。一些经典的数据增强算法并不十分适合于本研究，例如反向翻译方法，它需要将中文文本翻译成其他语言，再由机器翻译回中文。这样得到的文本必定全部为机器生成的，而不是人类和机器的混合文本。因此，本文将采用对训练集内部的文本进行拆分和重组的方法，得到新的训练数据。

参考 Jiang et al. (2023)[48] 针对文本分类问题提出的基于注意力的数据增强方法，将原始文本拆分为多个子句后，通过 BERT 词嵌入和编码器模块得到完整文本和各子句的文本表示向量，分别记为 K 和 Q。再计算完整语句与各子句之间的缩放点积注意力，如图 5-2 所示。与 BERT 模型训练时的自注意力机制不同的是，此处的注意力是完成文本与各子句之间的注意力，它在一定程度上反映了完整句子与各子句的相似度和对其关注的程度。

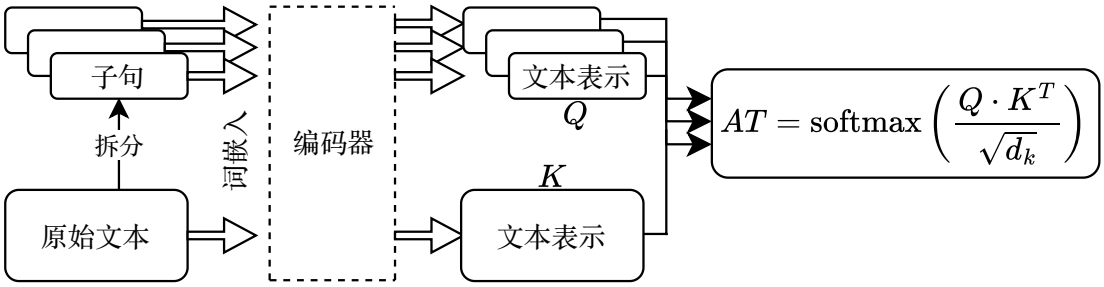


图 5-2 计算各子句注意力

对于一条被模型错误分类的文本，它对不同的子句有着不同的注意力，其中注意力最高的子句很可能是导致模型给出决策的最重要因素，也就是模型错误分类的原因。若我们希望模型能够正确分类这条文本，一个直觉的做法是将其注意力转移到其他子句上，这样也许能够使模型正确分类。

为了让模型将分类错误的文本的注意力转移到其他子句上，可以直接将注意力较高的子句删除，这样剩余子句的注意力自然能够得到提高。但是，只删除语句会让文本长度、句子数量等统计特征也发生变化，且删除后的文本与原始文本相似度较高，无法为训练数据提供多样性。因此，我们还需要为删除的子句找到替代的子句，即经过删除和新添这两个步骤。

在理想的情况下，新添的子句应具有两个特点：一是与原始文本的标签相同，否则两个不同标签的文本混合在一起后，无法明确定义其标签；二是新添子句在其所属的完整文本中的注意力应较低，否则数据增强后的文本很可能将大部分注意力放在新添子句上，而没有达到我们需要的效果：将注意力转移到原始文本中注意力较低的子句上。

记 $\text{sub}(X)$ 为文本 X 的子句集合，$R(X)$ 为 X 的文本表示向量。基于一条分类错误的文本 E 和一条分类正确的文本 C 得到的数据增强文本 $\hat{x}$ 的计算方法如式 5.1 所示：

$$
\begin{aligned}
\hat{x} &= \text{Mix}(E, C) = (E \setminus A) \cup B, \\
A &= \left\{ s \in \text{sub}(E) \mid \text{AT}(s, E) > \text{median}\left(\left\{\text{AT}\left(s', E\right) \mid s' \in \text{sub}(E)\right\}\right)\right\}, \\
B &= \left\{ s \in \text{sub}(C) \mid \text{AT}(s, C) < \text{median}\left(\left\{\text{AT}\left(s', C\right) \mid s' \in \text{sub}(C)\right\}\right)\right\}.
\end{aligned}
\tag{5.1}
$$

式 5.1 中的 $\text{AT}(s, E)$ 表示文本 E 在子句 s 上的注意力值。对于式 5.1 中的 E 和 C，我们要求它们的标签相同，即 $y_E = y_C$。因此，对于某个分类错误的文本 E，我们可以找到许多条与其标签相同且分类正确的文本，从而可以构造许多条数据增强文本。若将每个分类错误的文本经式 5.1 计算得到的数据增强文本都加入训练集，会使训练集的规模增大数百倍，这需要消耗大量的计算资源和时间，

并且数倍的训练规模也不一定能带来数倍的训练效果提升。因此，我们需要对构造出的候选数据增强文本进行筛选。

不同的数据增强文本加入训练集后，会使模型参数发生不同的变化。我们希望数据增强的样本 $\hat{x}$ 能够帮助模型纠正对文本 E 分类的错误，也就是希望模型对 $\hat{x}$ 应取得准确的分类结果。基于这一逻辑，我们对所有数据增强样本的候选值进行预测，并选择对应标签的逻辑值最大的那一个作为最终的数据增强样本。数据增强下的 BERT 模型训练算法如算法 1 所示。由于生成数据增强的样本耗时较长，本文选取的数据增强候选样本数 n 为 10。

Algorithm 1 基于注意力和分类效果筛选的数据增强算法

Input 训练数据 $D \leftarrow \{(x, y)\}$，BERT 预训练模型 M，训练轮数 $epoch$，数据增强候选样本数 n

Output 微调后的 BERT 分类模型 M

function DATA_AUGMENTATION($D, M, epoch, n$)
  while epoch **do**
    Train M on D
    $C \leftarrow \varnothing$, $E \leftarrow \varnothing$
    for $(x, y) \in D$ **do**
      $y' \leftarrow M(x)$
      if $y' \neq y$ **then**
        $E \leftarrow E \cup \{(x, y)\}$
      else
        $C \leftarrow C \cup \{(x, y)\}$
      end if
    end for
    if $E \neq \varnothing$ **then**
      $D_{aug} \leftarrow \varnothing$
      for $(x_e, y_e) \in E$ **do**
        $\hat{X} \leftarrow \varnothing$, $\hat{L} \leftarrow \varnothing$
        $C_n \leftarrow$ random n samples from C
        for $(x_c, y_c) \in C_n$ **do** ▷ 生成 n 个数据增强候选样本
          $\hat{X} \leftarrow \hat{X} \cup \{\text{Mix}(x_e, x_c)\}$, $\hat{L} \leftarrow \hat{L} \cup \{M(\hat{x})_{y_e}\}$
        end for
        $i \leftarrow \text{argmax}(\hat{L})$ ▷ 筛选分类效果最好的数据增强样本
        $D_{aug} \leftarrow D_{aug} \cup \{\hat{X}_i\}$
      end for
      $D \leftarrow D \cup D_{aug}$
    end if
  end while
  return M
end function

5.4.2　可视化案例分析

第 5.4.1 小节介绍了基于注意力和分类效果筛选的数据增强算法的理论和直觉，本小节通过可视化展示如何对一条错误分类的文本进行数据增强，以及数据增强前后的模型对文本的预测变化。

图 5-3 展示了一条由机器续写的文本，以及 SHAP 对各句的可解释性分析。图中蓝色代表对分类结果的负贡献，即使 BERT 分类模型更倾向于将文本分类为人类生成的文本；红色代表对分类结果的正贡献，即使 BERT 分类模型更倾向于将文本分类为机器生成的文本。颜色越深代表贡献的程度越大。可以看出，模型大致将靠前部分的文本识别为人类生成，而将靠后部分的文本识别为机器生成，但前者的贡献更大。最终模型输出该条文本是由机器生成的概率值为 1.1%，而实际上为机器生成，因此分类错误。

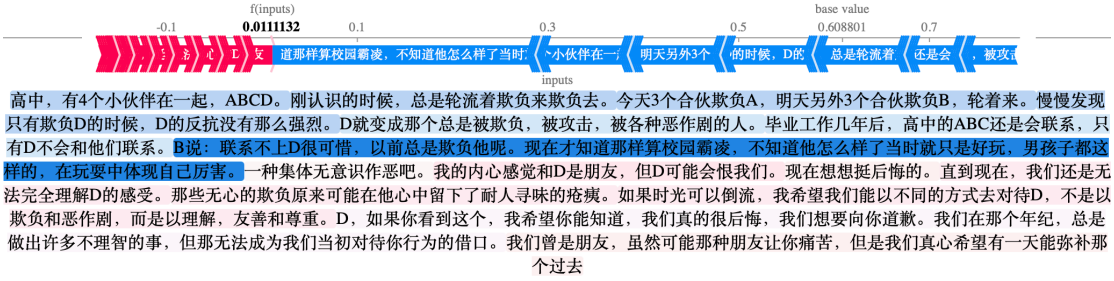


图 5-3　分类错误文本的 SHAP 分句可解释性分析

图 5-4 是该条文本在每个子句上的注意力分布。纵坐标中的红色代表该句是由机器续写的。BERT 模型将大部分注意力放在了靠前的子句上，而对靠后的子句的注意力较低。我们以所有注意力得分的中位数为阈值，将注意力较高的一半子句删除，最终保留的子句为五角星标记的条形图。

将保留的子句重新组合成文本后，再次计算 BERT 模型在各子句上的注意力分布，如图 5-5 所示。可以看出，最后四句由机器续写的子句的注意力得分明显提升。再与某个分类正确的文本进行拼接后，模型也更有可能会关注这四句由机器续写的文本，从而正确分类。

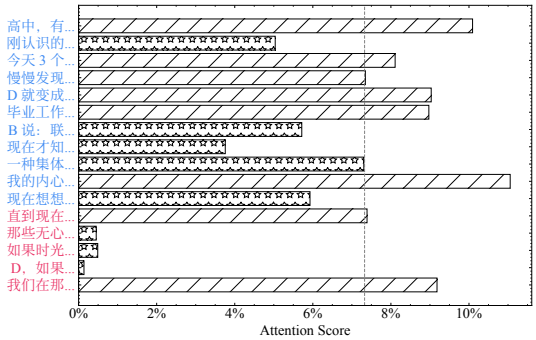


图 5-4 分类错误文本的注意力分布

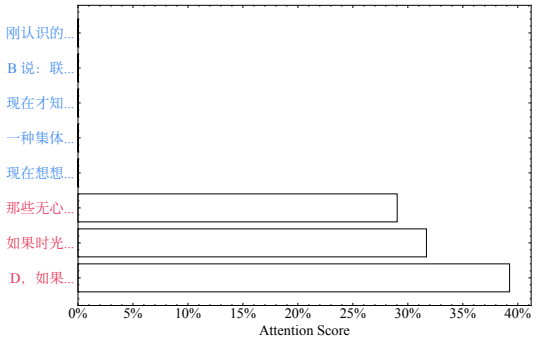
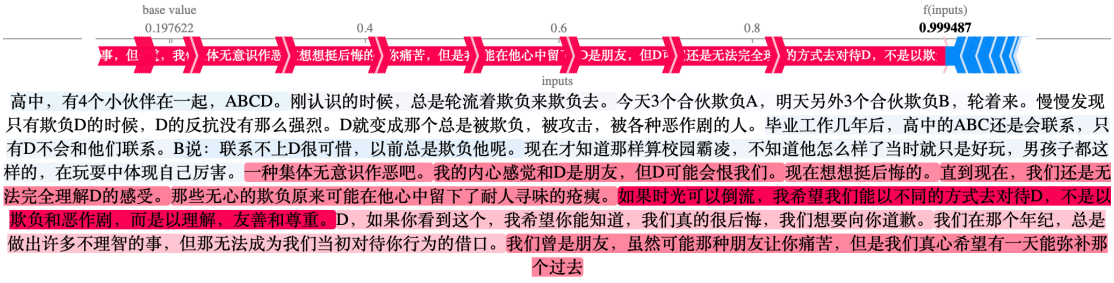


图 5-5 删除高注意力子句后的注意力分布

在应用基于注意力和分类效果筛选的数据增强算法后，我们再次使用 SHAP 对该条原本分类错误的文本进行分句可解释性分析，如图 5-6 所示。此时后一部分由机器续写的文本呈现显著的红色，其对 BERT 模型的预测贡献相比图 5-3 明显更大了。



高中，有4个小伙伴在一起，ABCD。刚认识的时候，总是轮流着欺负来欺负去。今天3个合伙欺负A，明天另外3个合伙欺负B，轮着来。慢慢发现只有欺负D的时候，D的反抗没有那么强烈。D就变成那个总是被欺负，被攻击，被各种恶作剧的人。毕业工作几年后，高中的ABC还是会联系，只有D不会和他们联系。B说：联系不上D很可惜，以前总是欺负他呢。现在才知道那样算校园霸凌，不知道他怎么样了当时就只是好玩，男孩子都这样的，在玩耍中体现自己厉害。一种集体无意识作恶吧。我的内心感觉和D是朋友，但D可能会恨我们。现在想想挺后悔的。直到现在，我们还是无法完全理解D的感受。那些无心的欺负原来可能在他心中留下了耐人寻味的疮痍。如果时光可以倒流，我希望我们能以不同的方式去对待D，不是以欺负和恶作剧，而是以理解，友善和尊重。D，如果你看到这个，我希望你能知道，我们真的很后悔，我们想要向你道歉。我们在那个年纪，总是做出许多不理智的事，但那无法成为我们当初对待你行为的借口。我们曾是朋友，虽然可能那种朋友让你痛苦，但是我们真心希望有一天能弥补那个过去

图 5-6 数据增强后模型的 SHAP 分句可解释性分析

5.5 实验结果

为了探究基于注意力和分类效果筛选的数据增强算法的优势，在第一组无数据增强样本的实验后，我们同时进行了多组对比实验。

第二组是随机拼接的数据增强算法（Random Concat），即随机将分类错误的文本中的一半子句与分类正确的文本中的一半子句进行拼接。这组实验是为了考察基于注意力将文本拆分为不同子句后，分别选择分类错误语句中的低注意力和分类正确语句中的高注意力的子句进行拼接的效果。

第三组是基于注意力的数据增强算法，再从生成的数据增强候选样本中随机选择一条（AT+Random）。这组实验是为了考察基于反映分类效果的逻辑值进行筛选的效果，体现这种特定的筛选方式为模型效果带来的增益。

第四组是基于注意力的数据增强算法，再从生成的数据增强候选样本中选择分类效果最好的一条（AT+Logit）。这组实验是使用本文提出的基于注意力和分类效果筛选的数据增强算法。

加入数据增强算法后训练得到的 BERT 模型在测试集上的分类效果如表 5-6 所示，其中前四行为所有测试集数据上的分类效果，后四行为仅在人类生成文本和机器续写混合文本上的分类效果。

表 5-6　应用数据增强算法的 BERT 模型分类效果

	Accu.	**Prec.**	**Recall-M**	**Recall-H**	**Recall-Avg**	**F1**	**AUC**
Without Aug	96.13	97.21	97.73	90.96	94.34	97.47	0.961
Random Concat	96.94	97.27	**98.77**	91.05	94.91	98.01	0.9953
AT+Random	**97.36**	97.8	**98.77**	92.81	95.79	**98.28**	0.9966
AT+Logit	96.9	**99.27**	96.65	**97.69**	**97.17**	97.94	**0.9967**
Without Aug	91.29	79.08	92.2	90.96	91.58	85.14	0.9545
Random Concat	92.71	80.1	**97.2**	91.05	94.12	87.82	0.9888
AT+Random	93.8	83.26	96.46	92.81	94.64	89.38	0.9902
AT+Logit	**96.83**	**93.83**	94.51	**97.69**	**96.1**	**94.17**	**0.9944**

应用数据增强算法后，模型分类效果均比未应用数据增强算法时（Without Aug）有所提升。其中基于注意力和分类效果筛选的数据增强算法在大部分分类效果评价指标上的提升均最大，尤其是在针对机器续写文本的检测中取得了最高的 F1 值和 AUC 值。随机拼接的训练效果提升最低，说明基于注意力挑选的子句作为数据增强样本对模型效果提升的效率更高。同时，基于预测逻辑值筛选的数据增强样本在大部分评价指标上也优于随机筛选的样本。

5.6　稳健性分析

5.6.1　子句替换比例

基于注意力进行数据增强的一个关键步骤是，将分类错误文本中注意力较高的子句用分类正确文本中注意力较低的子句进行替换。在表 5-6 中，我们将分类错误文本中一半的子句替换为分类正确文本的一半文本。为了探究替换比例更多或更少是否会对模型标签产生较大影响，我们设置了 1/3 和 2/3 两种不同的子句替换比例，得到的 BERT 模型在测试集上的分类效果如图 5-7 所示。

从图 5-7 中可以看出，在大部分评价指标上，替换比例为 1/2 时的模型效果最好。替换比例为 1/3 和 2/3 的模型效果较为接近。整体来看，三种替换比例下的模型分类效果均比未进行数据增强的模型效果有显著提升，说明子句替换比例的稳健性较好。

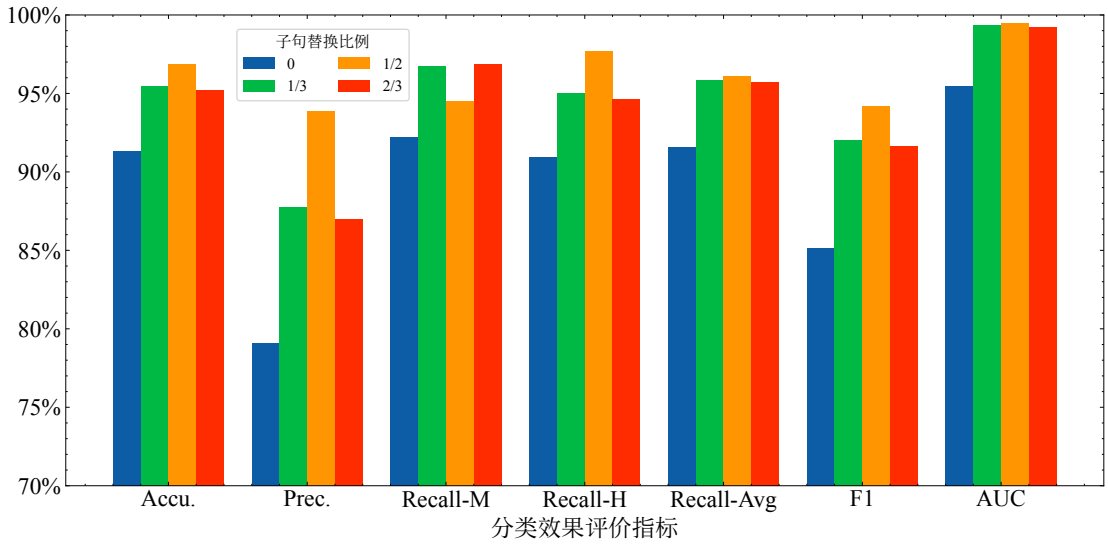


图 5-7 不同子句替换比例下的 BERT 模型分类效果

5.6.2 数据增强样本量

表 5-6 中，我们从 10 个数据增强候选样本中选取对应标签的逻辑值最大的 1 个样本进行数据增强。为了考察数据增强样本数量对模型表现的影响，我们也设置了 1 到 3 个不同的数据增强样本数量，得到的 BERT 模型在测试集上的分类效果如图 5-8 所示。

随着数据增强样本数量的增多，模型的表现并没有提升，在除了机器生成文本召回率之外的其他指标上均有所下降。这说明数据增强的样本并不是越多越好，过多的数据增强样本可能会带来样本不平衡的问题，反而使模型的表现下降。在数据增强样本为 1 或 2 时，模型的标签均好于未应用数据增强算法时的表现，更多的数据增强样本并不会带来更好的效果。

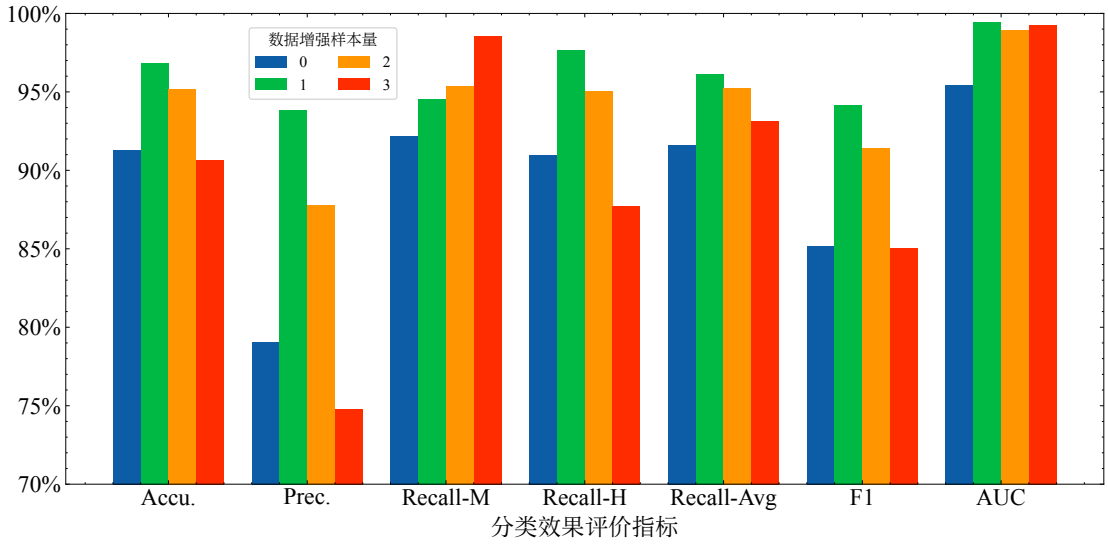


图 5-8 不同数据增强样本数量下的 BERT 模型分类效果

5.7　机器生成文本检测应用演示

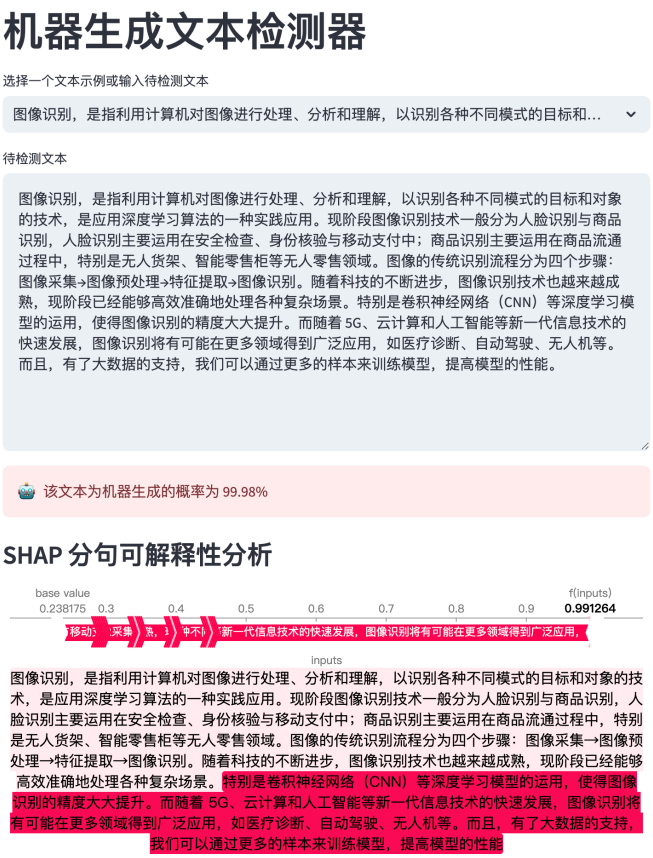


图 5-9　机器生成文本检测应用演示

最后，本文基于前文使用注意力和分类效果筛选的数据增强算法训练的 BERT 模型，开发了一个机器生成文本检测器应用，应用的界面如图 5-9 所示。该应用的在线地址为 `https://machine-generated-text-detection.streamlit.app/`。

用户输入待检测的文本后，即可查看模型对文本是否为机器生成的概率预测值，以及基于 SHAP 的分句可解释性分析。在分句可解释性分析中，我们使用句号、问号和感叹号将长文本划分为单句，并考察各单句对模型预测结果的影响。图中红色代表使模型预测结果为机器生成文本的因素，蓝色代表使模型预测结果为人类生成文本的因素。在图 5-9 中的示例中，最后若干句话呈现明显的红色，说明模型能够识别出这部分文本是机器生成的，即机器续写文本。

第 6 章　总结与展望

6.1　总结

随着以 ChatGPT 为代表的大语言模型能力持续提升，机器生成文本在各领域的应用场景也在不断扩展。虽然大语言模型的强大能力为人们的生产和生活带来了诸多便利，但不恰当地使用机器生成的文本也会带来严重的负面影响。因此，准确识别人类和机器生成的文本变得至关重要。过往研究大部分集中在英文语料的机器生成文本检测上，且主要关注单一来源的文本检测，对混合来源文本检测问题关注较少。本文构建了多领域、多模型生成的人类和机器生成中文文本数据集，在该数据集的基础上进行实证研究。

首先，本文计算了文本长度、句子数量、单句长度均值、单句长度标准差、用词丰富度和文本困惑度这 6 个统计特征，发现人类和机器生成的文本在一些统计特征上存在差异，尤其是机器生成文本的用词丰富度和文本困惑度显著低于人类生成文本。这些差异为区分人类和机器生成的文本提供了初步的依据，基于这 6 个特征的 XGBoost 分类器便可以取得 92.76% 的 F1 值，其中对预测效果贡献最大的特征是文本困惑度，占比达 62.63%。

根据是否来自相同领域和是否由相同模型生成设计了 4 种不同的实验组合，使用 fastText 进行文本分类，可以取得约 90% 的 F1 值。实验结果表明，不同模型这一实验组合下，由于特征分布差异较大，分类任务最困难。BERT 模型在不同领域、不同模型这一实验组合下的分类效果最好，F1 值达到 98.6%。

使用 ChatGPT 自身进行零样本学习识别机器生成文本，虽表现优于随机猜测，但其人类生成召回率仅有 53.4%，F1 值为 71.85%，并且分类结果具有一定的随机性。

利用机器续写的方式构造混合来源文本，发现基于单一来源文本训练的 BERT 模型对混合来源文本的分类效果有明显下降。将混合来源文本加入训练集后，模型表现虽有提升，但在混合来源文本测试数据上的检测效果仍显著更低，F1 值相比在单一来源文本上的表现下降 12.52%。

本文改进了一种适用于人类和机器混合文本检测的数据增强方法，将分类错误的样本中注意力较高的子句用分类正确的样本中注意力较低的子句进行替换，并筛选出分类效果最好的作为数据增强样本。与随机拼接和随机筛选的数据增强方法相比，本文提出的数据增强方法在混合来源文本检测任务上取得了

最好的效果，F1 值提升了约 5%。该算法对子句替换比例和数据增强文本数量两个超参数较稳健，在大多数超参数设置下均能取得比未进行数据增强时更好的分类效果。

最后，将本文训练的机器生成文本检测模型部署到了一个在线服务中，用户可以通过输入文本，获取文本检测结果和 SHAP 分句可解释性分析。

6.2　不足与展望

本文构建了人类和机器生成中文文本数据集，并在此基础上进行了单一来源和混合来源文本检测的实证研究。然而，本文的研究仍存在一些局限性和不足之处：

第一，本文在构建数据集的过程中，未对模型的输出添加额外的指令，这可能无法模拟有意避开检测的机器生成文本。例如，生成者可能会要求机器按照某种风格、语气、文风等规范来生成文本，这些文本可能与本文使用的训练数据差异较大，从而使分类模型效果下降。后续可根据生成者常用的指令迭代训练数据，以提升模型对各种风格文本检测的稳健性。

第二，本文在构建文本统计特征时考虑了文本困惑度等 6 个特征，受限于相关计算模型的准确度和计算效率，未对情感、词性、命名实体等角度进行挖掘。若能得到更丰富的文本特征，也许可以进一步提升 XGBoost 等机器学习分类模型的效果。

第三，针对混合来源文本检测问题，在应用本文提出的基于注意力和分类效果筛选的数据增强算法时，直接不同句子的子句进行替换的策略很可能会导致生成的文本不通顺，这有可能影响模型的分类效果。后续可以考虑结合语义相似度等指标，进一步改进数据增强算法。

参考文献

[1] OPENAI. Introducing ChatGPT[EB/OL]. 2022. https://openai.com/blog/chatgpt.

[2] GEORGE A S, GEORGE A H. A review of ChatGPT AI's impact on several business sectors[J]. Partners Universal International Innovation Journal, 2023, 1 (1): 9-23.

[3] COTTON D R, COTTON P A, SHIPWAY J R. Chatting and cheating: Ensuring academic integrity in the era of ChatGPT[J]. Innovations in Education and Teaching International, 2023: 1-12.

[4] ANDERSON N, BELAVY D L, PERLE S M, et al. AI did not write this manuscript, or did it? Can we trick the AI text detector into generated texts? The potential future of ChatGPT and AI in Sports & Exercise Medicine manuscript generation[J]. BMJ Open Sport—Exercise Medicine, 2023, 9(1).

[5] STUDY.COM. Productive teaching tool or innovative cheating?[EB/OL]. 2023. https://study.com/resources/perceptions-of-chatgpt-in-schools.

[6] 黄楚新, 张迪. ChatGPT 对新闻传播的机遇变革与风险隐忧[J]. 视听界, 2023 (30-35).

[7] CROTHERS E, JAPKOWICZ N, VIKTOR H L. Machine-generated text: A comprehensive survey of threat models and detection methods[J]. IEEE Access, 2023.

[8] JELINEK F. Statistical methods for speech recognition[M]. MIT press, 1998.

[9] ROSENFELD R. Two decades of statistical language modeling: Where do we go from here?[J]. Proceedings of the IEEE, 2000, 88(8): 1270-1278.

[10] SZYMANSKI G, CIOTA Z. Hidden Markov models suitable for text generation [C]//WSEAS International Conference on Signal, Speech and Image Processing (WSEAS ICOSSIP 2002). 2002: 3081-3084.

[11] KATZ S. Estimation of probabilities from sparse data for the language model component of a speech recognizer[J]. IEEE transactions on acoustics, speech, and signal processing, 1987, 35(3): 400-401.

[12] BENGIO Y, DUCHARME R, VINCENT P. A neural probabilistic language model [M]//Advances in Neural Information Processing Systems: Vol. 13. MIT Press, 2000.

[13] 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.

[14] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[A]. 2013.

[15] PETERS M E, NEUMANN M, IYYER M, et al. Deep contextualized word representations[C]//Proceedings of NAACL-HLT. 2018: 2227-2237.

[16] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[A]. 2019. arXiv: 1810.04805.

[17] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Advances in Neural Information Processing Systems: Vol. 30. Curran Associates, Inc., 2017.

[18] RADFORD A, WU J, CHILD R, et al. Language models are unsupervised multitask learners[J]. OpenAI blog, 2019, 1(8): 9.

[19] WEI J, TAY Y, BOMMASANI R, et al. Emergent abilities of large language models[A]. 2022. arXiv: 2206.07682.

[20] BROWN T, MANN B, RYDER N, et al. Language models are few-shot learners [J]. Advances in neural information processing systems, 2020, 33: 1877-1901.

[21] HOFFMANN J, BORGEAUD S, MENSCH A, et al. Training compute-optimal large language models[A]. 2022. arXiv: 2203.15556.

[22] SHANAHAN M. Talking about large language models[J]. Communications of the ACM, 2024, 67(2): 68-79.

[23] BAKI S, VERMA R, MUKHERJEE A, et al. Scaling and effectiveness of email masquerade attacks: Exploiting natural language generation[C]//Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security. 2017: 469-482.

[24] STIFF H, JOHANSSON F. Detecting computer-generated disinformation[J]. International Journal of Data Science and Analytics, 2022, 13(4): 363-383.

[25] FENG X, LIU M, LIU J, et al. Topic-to-essay generation with neural networks [C]//IJCAI. 2018: 4078-4084.

[26] DEHOUCHE N. Plagiarism in the age of massive Generative Pre-trained Transformers (GPT-3)[J]. Ethics in Science and Environmental Politics, 2021, 21: 17-23.

[27] CABANAC G, LABBÉ C. Prevalence of nonsensical algorithmically generated papers in the scientific literature[J]. Journal of the Association for Information Science and Technology, 2021, 72(12): 1461-1476.

[28] BOYD-GRABER A R J, OKAZAKI N, ROGERS A. ACL 2023 policy on AI writing assistance[J]. 2023-02-09]. https://2023. aclweb. org/blog/ACL-2023-policy, 2023.

[29] 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.

[30] ZIPF G K. Human behavior and the principle of least effort[M]. Addison-Wesley Press, 1949.

[31] NGUYEN-SON H Q, TIEU N D T, NGUYEN H H, et al. Identifying computer-generated text using statistical analysis[C]//2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). IEEE, 2017: 1504-1511.

[32] GEHRMANN S, STROBELT H, RUSH A M. GLTR: Statistical detection and visualization of generated text[A]. 2019.

[33] GUO B, ZHANG X, WANG Z, et al. How close is ChatGPT to human experts? Comparison corpus, evaluation, and detection[A]. 2023. arXiv: 2301.07597.

[34] LI Y, LI Q, CUI L, et al. Deepfake text detection in the wild[A]. 2023. arXiv: 2305.13242.

[35] LIU Y, OTT M, GOYAL N, et al. RoBERTa: A robustly optimized BERT pretraining approach[A]. 2019. arXiv: 1907.11692.

[36] LIAO W, LIU Z, DAI H, et al. Differentiating ChatGPT-generated and human-written medical texts: Quantitative study[J/OL]. JMIR Medical Education, 2023, 9: e48904. http://dx.doi.org/10.2196/48904.

[37] LIU Y, ZHANG Z, ZHANG W, et al. ArguGPT: evaluating, understanding and identifying argumentative essays generated by GPT models[A]. 2023. arXiv: 2304.07666.

[38] ZELLERS R, HOLTZMAN A, RASHKIN H, et al. Defending against neural fake news[J]. Advances in neural information processing systems, 2019, 32.

[39] MITCHELL E, LEE Y, KHAZATSKY A, et al. DetectGPT: Zero-shot machine-generated text detection using probability curvature[A]. 2023. arXiv: 2301.11305.

[40] CHEN T, GUESTRIN C. XGboost: A scalable tree boosting system[C]// Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016: 785-794.

[41] JOULIN A, GRAVE E, BOJANOWSKI P, et al. Bag of tricks for efficient text classification[A]. 2016. arXiv: 1607.01759.

[42] HE X, CHEN S, JU Z, et al. MedDialog: Two large-scale medical dialogue datasets[A]. 2020. arXiv: 2004.03329.

[43] DUAN N. Overview of the NLPCC-ICCPOL 2016 shared task: Open domain Chinese question answering[C]//Natural Language Understanding and Intelligent Applications: 5th CCF Conference on Natural Language Processing and Chinese Computing, NLPCC 2016, and 24th International Conference on Computer Processing of Oriental Languages, ICCPOL 2016, Kunming, China, December 2–6, 2016, Proceedings 24. Springer, 2016: 942-948.

[44] XU B. Nlp chinese corpus: Large scale chinese corpus for nlp[J]. Zenodo, 2019.

[45] XU L, LI A, ZHU L, et al. SuperCLUE: A comprehensive Chinese large language model benchmark[A]. 2023. arXiv: 2307.15020.

[46] CUI Y, CHE W, LIU T, et al. Revisiting pre-trained models for Chinese natural language processing[C/OL]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings. Online: Association for Computational Linguistics, 2020: 657-668. https://www.aclweb.org/anthology/2020.findings-emnlp.58.

[47] LUNDBERG S M, LEE S I. A unified approach to interpreting model predictions[M]//GUYON I, LUXBURG U V, BENGIO S, et al. Advances in Neural Information Processing Systems 30. Curran Associates, Inc., 2017: 4765-4774.

[48] JIANG S, CHU Y, WANG Z, et al. Explainable text classifcation via attentive and targeted mixing data augmentation[C]//International Joint Conference on Artificial Intelligence. 2023.

致　谢

三年前的夏天，我收到了来自复旦大学管理学院硕士研究生的录取通知。这是一封期待已久的邮件，是我人生中的一次重要转折，也是让我一次又一次感到幸运和感激的开始。

感谢教授过我课程的窦一凡老师、方冠华老师、冯项楠老师、刚博文老师、耿哲老师、刘彬老师、刘炎老师、孙海老师、王有为老师、夏寅老师、虞嘉怡老师、张妮拉老师、张新生老师、朱祁老师，他们在课堂上传授的专业知识培养了我的数理统计素养和编程技能，让我能够自信地面对学术研究和业界实践的各种课题与挑战。

感谢本硕博教育中心的王耀珍老师、虞盛达老师、郑班举老师，他们为硕士项目的课程安排与优化、班级文化活动等付出了许多。

感谢职业发展中心的李诞新老师和张敏仪老师，他们为我提供了诸多职业发展上的指导和机会，让我能够找寻并从事自己热爱的事业。

感谢张成洪老师、陈刚老师、程坦师兄在本论文的研究方向、实验设计和论文写作过程中的悉心指导和帮助，让我力求科学、创新地研究有趣且有价值的课题。感谢管理学院提供的实验所需算力资源。感谢人工智能领域的前辈学者们，他们不断创新发展的学术理论和经验让我能站在巨人的肩膀上进行学术研究。感谢众多的技术分享者和开源软件贡献者，他们带来的技术进步让我有了实现本文众多复杂计算程序的可能。

特别感谢 2022 级数据科学与商务分析硕士项目的所有同学们，是他们与我共同学习、共同进步、共同成长。在复旦两年的求学时光是幸运且美好的，它将永远珍藏在我的回忆中。

复旦大学
学位论文独创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。论文中除特别标注的内容外，不包含任何其他个人或机构已经发表或撰写过的研究成果。对本研究做出重要贡献的个人和集体，均已在论文中作了明确的声明并表示了谢意。本声明的法律结果由本人承担。

作者签名：冯超　　日期：2024.6.6

复旦大学
学位论文使用授权声明

本人完全了解复旦大学有关收藏和利用博士、硕士学位论文的规定，即：学校有权收藏、使用并向国家有关部门或机构送交论文的印刷本和电子版本；允许论文被查阅和借阅；学校可以公布论文的全部或部分内容，可以采用影印、缩印或其它复制手段保存论文。涉密学位论文在解密后遵守此规定。

作者签名：冯超　　导师签名：张以洗　　日期：2024.6.6