基于改进Transformer的电信重投报告自动生成方法研究

梁伟明[1,3]，肖军[2,3]，马晓亮[1,3]，辛盛[2,3]，徐荣彬[2,3]
（1. 中国电信股份有限公司广州分公司，广东 广州 510620；
2. 中数通信息有限公司，广东 广州 510630；
3. 马晓亮劳模与工匠人才创新工作室，广东 广州 510620）

摘　要：在电信行业中，客户对未解决或处理不满意的投诉进行重复投诉的现象较为常见。手动生成重投报告不仅耗时且主观性较强，难以满足企业对高效性和一致性的要求。针对这一问题，提出了一种基于改进Transformer模型的自动化报告生成方法。该方法通过引入情绪嵌入，有效捕捉客户在对话中的情绪变化，改善了生成报告对客户态度和诉求的理解能力。同时，结合定制化位置编码，提升了模型对投诉时序信息的感知能力，从而增强了生成内容的时间逻辑性和细节完整性。实验结果表明，改进后的模型在BLEU（bilingual evaluation understudy）和ROUGE（recall-oriented understudy for gisting evaluation）指标上分别达到0.352和0.482，显著优于原始Transformer和其他对比模型。此外，与人工对比，工作效率提高了89%。生成的报告内容不仅更加准确贴合实际需求，还在语义细节与时序一致性上表现优异。

关键词：电信客服；Transformer；重投报告；自动报告生成
中图分类号：TP399
文献标志码：A
doi: 10.11959/j.issn.1000−0801.2025110

Research on the automatic generation method of telecom re-complaints report based on improved Transformer model

LIANG Weiming[1,3], XIAO Jun[2,3], MA Xiaoliang[1,3], XIN Sheng[2,3], XU Rongbin[2,3]
1. China Telecom Corporation Guangzhou Branch, Guangzhou 510620, China
2. China DataCom Co., Ltd., Guangzhou 510630, China
3. Ma Xiaoliang Innovation Studio for Model Workers and Creative Talents, Guangzhou 510620, China

Abstract: In the telecommunications industry, repeated customer complaints about unresolved or unsatisfactory issues are a common challenge. Manually generating re-investment reports is not only time-consuming and prone to subjectivity but also fails to meet enterprise demands for efficiency and consistency. To address this issue, an automatic report generation method based on an improved Transformer model was proposed. This method introduced emotion embedding, enabling the model to effectively capture dynamic emotional changes in customer interactions and better understand customer attitudes and demands during the dialogue. Additionally, the incorporation of custom-

收稿日期：2024−12−11；修回日期：2025−04−07
通信作者：肖军，xj@cndatacom.com

ized position encoding enhanced the model's ability to perceive complaint time series information, significantly improving the time logic and detailed completeness of the generated content. Experimental results demonstrate that the proposed model achieves BLEU (bilingual evaluation understudy) and ROUGE (recall-oriented understudy for gisting evaluation) scores of 0.352 and 0.482, respectively, outperforming the original Transformer and other baseline models. Moreover, compared to manual efforts, the proposed model improves work efficiency by 89%. The generated reports not only align more accurately with real-world requirements but also exhibit superior performance in semantic detail and time sequence consistency.

Key words: telecom customer service, Transformer, re-complaints report, automatic report generation

0 引言

在电信行业，客户的不满情绪常常引发多次投诉（即重投），这一现象在客户服务中普遍存在。这类投诉通常涉及服务不满意、问题未解决或投诉响应不及时等原因，导致客户在第一次投诉后仍持续进行后续投诉[1]。重投现象往往会大幅增加客服人员的工作量，影响客户满意度，同时也导致较高的服务成本和潜在的客户流失风险[2]。因此，如何有效处理重投问题、减轻其对客服工作的影响，成为提升电信服务质量的关键。传统上，电信客服人员在重投处理中需要手动记录和分析每一单投诉，以确保每次投诉能够得到妥善处理。重投报告的内容通常涵盖客户信息、投诉内容、解决方案以及客户反馈等。然而，手动生成这些报告不仅耗时耗力，而且主观判断和标准不一致容易导致内容缺乏连贯性和一致性。因此，如何通过自动化手段生成高质量的重投报告，成为电信服务优化中的关键问题之一。

近年来，随着自然语言处理（natural language processing，NLP）和机器学习技术的发展，研究人员尝试将这些技术应用于自动化报告生成，以降低人力成本，提高处理效率。传统的报告生成方法通常依赖于规则和模板，将关键词替换进预定义的报告格式[3]。然而，模板化的报告生成方法难以应对复杂多样的客户需求，并且在应对不同行业的特定场景时表现有限。近年来，基于深度学习的文本生成技术在报告生成领域取得了显著的进展，特别是随着序列到序列（sequence-to-sequence，Seq2Seq）和注意力机制的提出，自动文本生成技术取得了显著突破。2014年，文献[4]通过长短期记忆（long short-term memory，LSTM）网络实现了高质量的句子生成。2017年，文献[5]提出的Transformer模型仅依靠自注意力机制，进一步推动了自动文本生成技术的发展。此外，近年的研究还显示，基于Transformer模型的变体BERT等预训练模型的生成式文本自动摘要系统，已能够自动从文本中提取关键信息并生成摘要[6]，而基于Transformer的医疗诊断报告生成系统也证明了该技术在其他行业中的广泛应用潜力[7]。尽管基于深度学习的文本生成技术在多个行业中取得了成功应用，但针对电信行业重投报告生成的研究仍然较为匮乏。电信客服对话通常涉及大量非结构化文本数据，且高度依赖对客户情绪和诉求的准确把握，传统的模板化方法难以满足生成高质量重投报告的需求。近年来，情感分析技术在多个领域的应用取得了显著成果。例如，文献[8]结合深度学习和情感分析成功预测了价格走势，文献[9]则通过调整客户评价的情感分析，改进了产品销售预测。这些研究表明，情感分析技术能够有效提升深度学习模型在处理实际应用任务时的表现。

针对电信行业重投报告生成的需求，本文提出了一种基于改进Transformer模型的自动化生成

方法。不同于传统的模板化方法，本文结合深度学习技术，通过情绪嵌入和定制化位置编码来增强模型对客户情绪和诉求的理解能力。通过自定义的Transformer模型，本文能够更精确地关注投诉内容的语义关系和时间序列特征，从而生成语义流畅、信息全面且符合实际业务需求的重投报告。具体来说，本文的主要贡献如下。

- 优化的多头自注意力机制与情绪嵌入相结合，使生成报告更具情绪感知能力，体现人性化特点。
- 引入定制化位置编码增强时间感知能力，提高生成报告在时序一致性上的表现。
- 设计了一套自动化报告生成框架，实现从数据预处理到高质量重投报告生成的完整流程。实验结果表明，该方法在生成效果上优于现有的Transformer变体，为解决行业痛点提供了高效、实用的技术方案。

1　数据处理

1.1　数据预处理

数据主要来源于电信客服中心的客服与客户对话记录。这些对话数据包含客户基本信息、投诉内容、客服的处理方案、客户反馈等内容。

本文的数据预处理环节包括数据清洗、文本规范化、分词与标注共3个部分。在数据清洗部分，需要检查并去除缺失值、空白字段和无效对话，同时过滤如寒暄、重复对话等非核心内容，以提高模型专注于关键投诉信息的能力，数据清洗结果见表1。在文本规范化部分，对文本进行规范化处理，如统一日期和时间的格式、标准化专有名词和缩略词等，文本规范化结果见表2。在分词与标注部分，使用中文分词工具——Jieba分词，进行对话文本分词，并标注如时间、地点、情绪词等的实体信息，以便为模型提供更精细的输入，分词与标注结果见表3。上述数据预处理环节可以为本文的模型提供高质量的输入数据，提升自动化生成的准确性和实用性。

1.2　情绪值设计

本文在处理情绪分值时，使用情绪分析工具VADER（valence aware dictionary and sentiment reasoner）来提取情绪值（emotionValue）[10]，其情绪得分E计算式如下所示。

$$E=\frac{\sum_{i=1}^{N}\text{emotionValue}_i}{N} \tag{1}$$

其中，N为对话情绪词总数。情绪值被划分为

表1　数据清洗结果

处理方式	原文	处理后
去除缺失值和空白字段	“...”“ ”等	（直接删除）
过滤非核心内容	“哎，您好”“嗯，再见。”等	（直接删除）
合并相关对话	“给的快代理店，他是没有写没有写到有qq卖的对吗”“嗯，就是给你的那个快递里面，他是没有写没有写到有qq卖的，是吧？对啊”	“给你的那个快递代理，有没有写到是否有qq卖的?”

表2　文本规范化结果

处理方式	原文	处理后
日期和时间格式化	“今天下午三点”	“2024-11-17 15:00”
专有名词/缩略词标准化	“转网”“合约”“拆机”等	“携号转网”“金融合约”“宽带拆机”等
去除多余标点符号	“哦，好好好好嗯，那没问题，再见。”	“没问题了。”

表3　分词与标注结果

处理方式	原文	处理后
中文分词	“请打开电信营业厅App搜索星号卡激活”	[“请”,“打开”,“电信营业厅”,“App”,“搜索”,“星号卡”,“激活”]
实体标注	“先生您的手机号177****0022无法激活”	[手机号:177****0022]
情绪特征提取	“这点事为什么处理这么久”	{“情绪”:“不接受!”}

“积极（emotionValue ≥0.5）”“消极（emotionValue ≤−0.35）”“中性（emotionValue为其他）”3个级别。经VADER计算后得到的情绪值示例如图1所示。然而，情绪值的范围较为有限，难以全面反映客户情绪的波动和实际态度。此外，对话内容中有些语句因语境等其他因素，虽然表面上没有明显的态度词，但从语气中可以听出是正面的或是负面的含义。因此，本文在数据预处理中做了情绪特征提取，如“这点事为什么处理这么久”，从语气和语境中能够得出客户对当前的情况不满意。如果不对这句话做预处理，直接用VADER进行分析，将得到情绪态度为“中性”，而实际应该为“消极”。

2　模型设计与改进

2.1　自注意力机制

自注意力机制是一种用于从输入数据中高效提取相关性特征的关键技术[11]。在电信行业的对话记录处理中，自注意力机制通过衡量输入特征间的相似性，动态地为相关部分分配权重，使模型能够聚焦于关键投诉内容并提升语义理解能力。自注意力机制计算式为：

$$\text{Attention}\left(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}\right)=\text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\text{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \quad (2)$$

其中，$\boldsymbol{Q}$（Query）、$\boldsymbol{K}$（Key）和$\boldsymbol{V}$（Value）表示通过线性变换从输入特征生成的3组向量，用于构建自注意力矩阵，即将文本信息转换成输入向量$\boldsymbol{X}=[x_1,\cdots,x_n]$，然后通过线性变换生成查询向量$\boldsymbol{Q}$、键向量$\boldsymbol{K}$和值向量$\boldsymbol{V}$。$\boldsymbol{Q}\boldsymbol{K}^{\text{T}}$能够捕捉查询向量和键向量之间的关系。为了避免数值过大导致梯度不稳定，点积结果通过$\sqrt{d_k}$进行缩放归一化[12]。自注意力计算过程示意图如图2所示。即输入token序列$l_1,l_2,\cdots,l_j$，经线性变换分别生成查询向量$\boldsymbol{q}_i$，

```
原文: 行吧，我可以接受
译文: Okay, I can accept
情绪分数: {'neg': 0.0, 'neu': 0.182, 'pos': 0.818, 'compound': 0.5423}
综合情绪值: 0.54
情绪态度: 积极
--------------------------------------------------
原文: 不接受!
译文: Don't accept it!
情绪分数: {'neg': 0.553, 'neu': 0.447, 'pos': 0.0, 'compound': -0.3561}
综合情绪值: -0.36
情绪态度: 消极
--------------------------------------------------
原文: 还是无法激活
译文: Still cannot be activated
情绪分数: {'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0}
综合情绪值: 0.00
情绪态度: 中性
--------------------------------------------------
```

图1　经VADER计算后得到的情绪值示例

键向量$\boldsymbol{k}_i$，值向量$\boldsymbol{v}_i(i=1,2,\cdots,j)$。以$\boldsymbol{q}_2$为例，计算其与序列中全部$\boldsymbol{k}_i(i=1,2,\cdots,j)$的点积，经缩放与归一化，最终将得到的结果与所有$\boldsymbol{v}_i$（$i=1,2,\cdots,j$）进行加权求和，输出向量$\boldsymbol{O}_2$。自注意力的优点在于能够动态地对输入特征进行加权，而非固定模式，因此可以捕捉长距离依赖关系，尤其适合处理与序列相关的对话记录的数据。

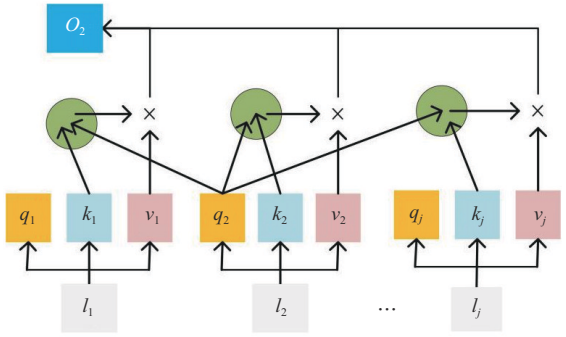


图2　自注意力计算过程示意图

2.2　Transformer模型

Transformer摒弃了循环，完全依赖于自注意力。自注意力机制通过学习输入序列中各词之间的依赖关系，生成高质量的文本输出。Transformer模型主要由编码器和解码器两部分组成，Transformer模型结构如图3所示，其中编码器包含多头注意力、前馈神经网络、位置编码等层次[13]。本文的输入数据由对话内容和元数据构成。为了适配Transformer模型的输入，需要将对话内容和元数据统一转换为嵌入表示。元数据包括情绪值和时间戳。对话内容包含多轮对话，通过拼接得到文本序列：

$$S=[r_1\oplus s_1\oplus t_1\oplus r_2\oplus s_2\oplus t_2\oplus\cdots\oplus r_n\oplus s_n\oplus t_n] \quad (3)$$

其中，r_n表示第n条对话的角色（客户或客服），t_n表示第n条对话的时间戳，s_n表示第n条对话的文本内容，$\oplus$表示字符串拼接操作。按照时间戳为对话记录排序，确保上下文语义的连贯性。此外，拼接对话文本的同时，在每条语句前添加角色标记（如[客户]和[客服]），以明确发言者身份。最后，将文本序列$S=[w_1,\cdots,w_n]$中的每个词w_i映射到嵌入空间，即$x_i=\mathrm{Embed}(w_i)$，则生成文本序列长度为n、词嵌入维度为d的对话嵌入矩阵：

$$\boldsymbol{X}=[x_1,x_2,\cdots,x_n],\boldsymbol{X}\in\mathbb{R}^{n\times d} \quad (4)$$

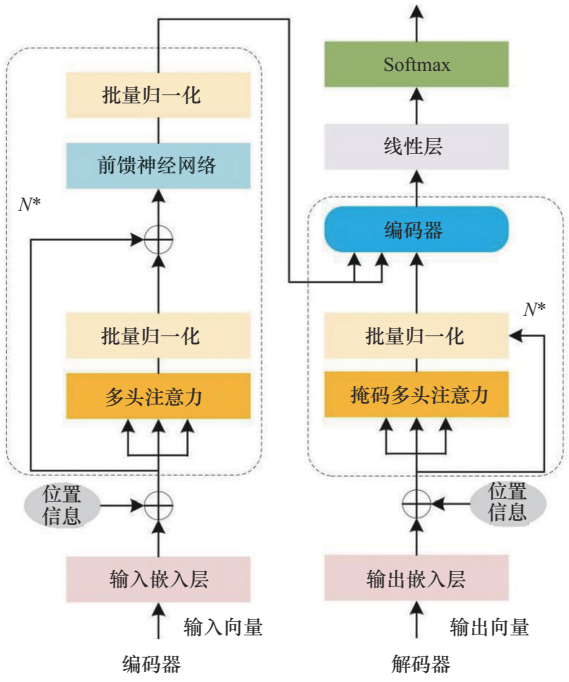


图3　Transformer模型结构

2.3　改进模型结构

2.3.1　情绪嵌入层

为了使模型能够理解客户的情绪变化，改进的模型在编码器中增加了情绪嵌入层。情绪嵌入是根据每句话的情绪值生成的向量，作为模型输入的一部分。情绪嵌入表达式为：

$$E_{\mathrm{embed}}=w_{\mathrm{e}}+\mathrm{emotionValue}+b_{\mathrm{e}},e_i=E_{\mathrm{embed}}(e_i) \quad (5)$$

其中，w_{e}和b_{e}为情绪嵌入的参数。由此得出情绪嵌入矩阵的表达式为：

$$\boldsymbol{E}=[e_1,e_2,\cdots,e_n],\boldsymbol{E}\in\mathbf{R}^{n\times d} \quad (6)$$

2.3.2　定制化位置编码

Transformer模型中的位置编码主要用于保留序列中的位置信息，以确保时序的一致性。在电信重投报告生成中，时序信息对于理解客

户的连续投诉至关重要。因此，改进的模型在标准位置编码的基础上加入了时序权重，使模型更具时间感知性[14]。改进的位置编码表达式为：

$$\begin{cases} \mathrm{PE}_{(\mathrm{pos},\,2i)} = \sin\left(\dfrac{\mathrm{pos}}{10\,000^{2i/d}}\right) + T_w \\ \mathrm{PE}_{(\mathrm{pos},\,2i+1)} = \cos\left(\dfrac{\mathrm{pos}}{10\,000^{2i/d}}\right) + T_w \end{cases} \tag{7}$$

其中，pos为位置，d为词向量维度，T_w为时间权重参数，式（7）中上下两个表达式分别表示为每个位置pos生成一个d维的向量，其中偶数维度（$2i$）使用正弦函数，奇数维度（$2i$+1）使用余弦函数，并引入了可学习的时间权重T_w以捕捉客户不同时点的诉求差异性。

2.3.3 Longformer模型

融合了文本向量和情绪嵌入向量的输入可能导致特征空间维度较高，进而增加了模型的计算复杂度和过拟合风险，并且Transformer模型的自注意力机制在处理长输入序列时具有$O(n^2)$的时间和空间复杂度，这可能会限制模型在长对话内容情况下的处理能力。Longformer是一种轻量级的Transformer变体，与Transformer有相同的网络结构，但该模型采用的是稀疏注意力机制，即仅在局部范围和选定全局标记之间计算注意力，从而大幅减少了计算量[15]，全局+局部注意力示意图如图4所示。因此，在将输入序列转化为嵌入向量后，本文先使用Longformer对融合了文本向量和情绪嵌入的嵌入向量进行处理。

综上，得到改进的Transformer模型，其结构如图5所示。输入层包括文本向量、情绪嵌入向量和定制位置编码。编码层采用改进的Transformer编码器块，每层包括多头注意力、情绪嵌入和位置编码。解码层基于编码层生成的语义表示，生成符合报告格式的文本内容。输出层生成的文本经过后处理，形成重投报告。该模型结合了情感嵌入和定制位置编码，以增强报告生成中的时间和情感意识。

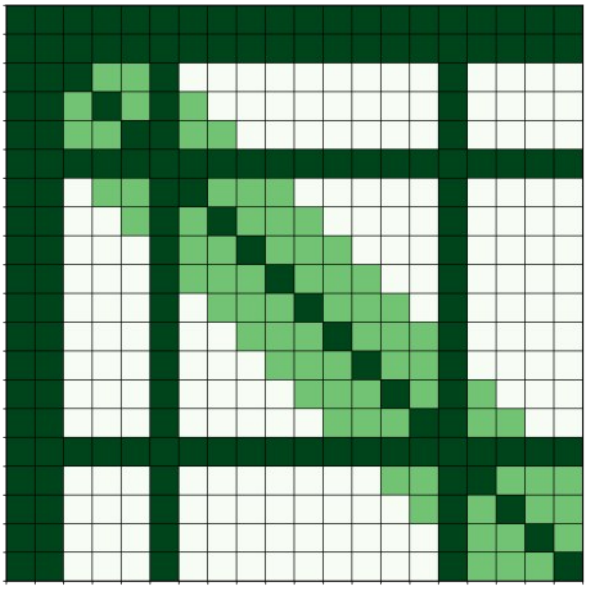


图4 全局+局部注意力示意图

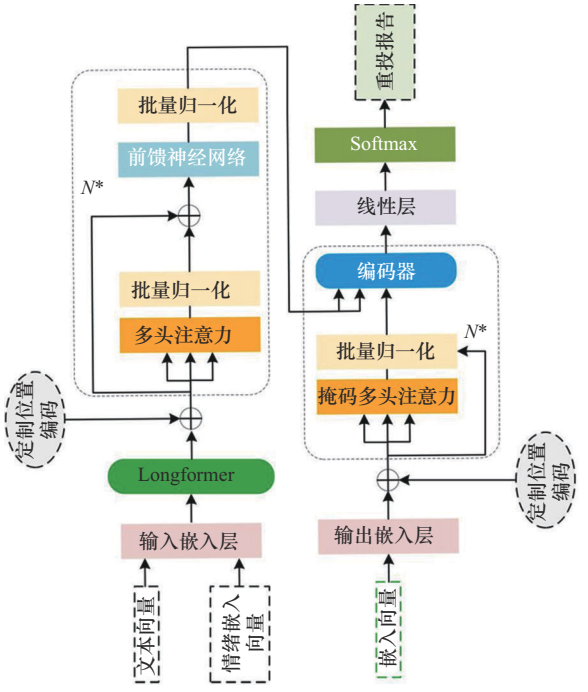


图5 改进的Transformer模型结构

3 实验结果与分析

3.1 实验设置

本实验基于Python3.7和PyTorch框架搭建实验环境，GPU为NVIDIA RTX 4090，模型超参数设置见表4。实验通过对原始的18 507条电信客

服对话进行数据清洗，获取其对话中的受理内容，包括客户姓名、手机号码、单位、归档时间等字段，以此作为重投数据集。为了确保自动生成的重投报告能够覆盖所有信息和时间戳，输入的数据中，除了重投数据，还需要包含前单受理内容的历史数据和其他补充信息，训练过程示意图如图6所示。

表4　模型超参数设置

参数名	参数意义	参数值设定
optimizer	优化函数	AdamW
max_seq_length	最大序列长度	2 048
embedding_dim	词嵌入维度	768
hidden_size	隐藏层维度	256
num_hidden_layers	隐藏层数量	12
attention_window	注意力窗口大小，捕获局部信息	256
learning_rate	学习率	1×10^{-5}
batch_size	批量大小	8
num_epochs	根据验证集表现决定是否早停	10
weight_decay	权重衰减，避免过拟合	0.01

3.2　评价指标

采用BLEU（bilingual evaluation understudy）分数和ROUGE（recall-oriented understudy for gisting evaluation）分数作为主要评估指标，以评估模型生成文本的准确性和一致性。

3.2.1　BLEU分数

BLEU分数是一种基于*n*-gram的自动化评估方法，用于测量生成文本和参考文本在词汇上的相似度[16]。BLEU分数的核心思想是计算生成文本与参考文本之间的*n*-gram精确匹配，并通过短语匹配来评估翻译或生成质量，其主要适用于机器翻译、文本生成等任务。BLEU分数的计算式为：

$$\mathrm{BLEU}=\mathrm{BP}\cdot\exp\left(\sum_{n=1}^{N}w_n\log p_n\right) \tag{8}$$

其中，p_n表示生成文本与参考文本的*n*-gram精确匹配率，w_n为*n*-gram的权重，BP（brevity penalty）为惩罚系数，防止生成文本过短。BLEU分数的范围为0～1，分数越高，表示生成文本与参考文本越接近。p_n和BP的计算式分别如下所示：

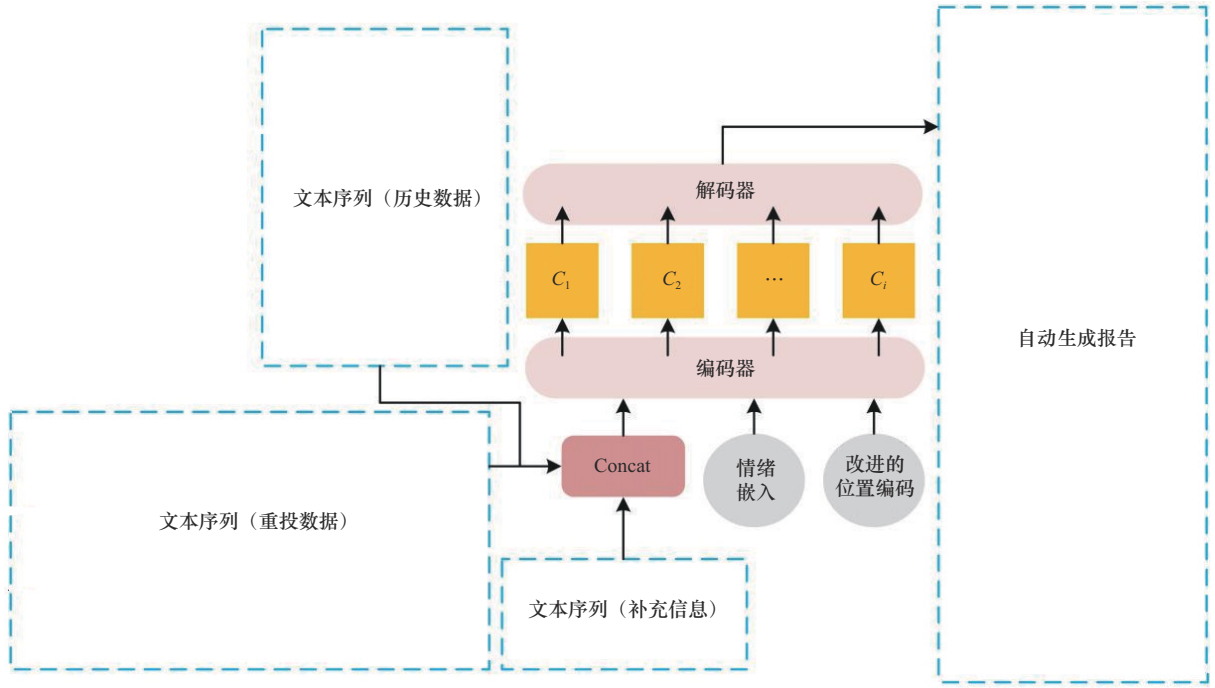


图6　训练过程示意图

$$p_n=\frac{\text{参考文本的}n\text{-gram}}{\text{生成文本中的所有}n\text{-gram}} \tag{9}$$

$$\mathrm{BP}=\exp\{\min(0,1-\frac{\mathrm{len}_R}{\mathrm{len}_C})\} \tag{10}$$

其中，len_R表示参考文本的长度，len_C表示生成文本的长度。在评估自动生成报告的质量时，本文采用BLEU作为评价指标。根据电信报告长文本的特点，在BLEU计算时使用最大n-gram为4。同时，为了避免零匹配问题，采用加1平滑方法，对1-gram、2-gram、3-gram和4-gram的精确度进行均匀加权。此外，还设置最大词数剪切，即每个n-gram的生成数量不超过参考文本中相应n-gram的最大出现次数。

3.2.2 ROUGE分数

ROUGE分数是另一种常用的文本评估指标，它主要关注生成文本对参考文本中n-gram或长段文本的召回情况，以评估生成文本的覆盖程度[17]。ROUGE包含多种变体，本文采用基于最长公共子序列（longest common subsequence，LCS）的ROUGE变体，即ROUGE-L。ROUGE-L关注的是生成文本和参考文本的词序匹配情况。ROUGE-L分数计算式如下所示：

$$\text{ROUGE-L}=\frac{\mathrm{LCS}(C,R)}{\mathrm{len}_R} \tag{11}$$

其中，LSC(C,R)表示生成文本C和参考文本R的LCS长度，len_R表示参考文本的长度。ROUGE-L通过计算生成文本与参考文本之间的LCS，衡量生成文本的全局语义一致性和流畅性。在计算时，本文选择了召回率（Recall）、精确率（Precision）和F1值这3个指标。F1值综合考虑了召回率和精确率，避免了单纯关注其中一个指标可能导致的偏差，最终以F1值作为评估标准，其计算式为：

$$\mathrm{F1}=\frac{2\times\mathrm{Recall}\times\mathrm{Precision}}{\mathrm{Recall}+\mathrm{Precision}} \tag{12}$$

3.3 实验结果与分析

核验发现，人工生成的报告存在时间戳不准确、责任单位错误及客服专员错误等问题，且人工生成对于套餐信息不敏感，关键信息容易被忽略，而自动生成的报告内容更加清晰。改进的Transformer模型在生成重投报告时表现优异，具有更高的语义一致性和内容准确性。同时，自动生成的报告能够更好地整合时间序列信息，使投诉处理流程更加准确。

模型不同超参数的BLEU分数见表5。实验结果表明，针对电信重投任务，较小的学习率可以确保模型训练的稳定性，而更大的注意力窗口可以捕捉更长距离的上下文信息，从而提升BLEU分数。

表5 模型不同超参数的BLEU分数

注意力窗口大小	学习率	BLEU分数
256	1×10^{-5}	0.352
256	1×10^{-4}	0.179
256	1×10^{-3}	0.031
128	1×10^{-5}	0.217
128	1×10^{-4}	0.164
128	1×10^{-3}	0.016

在编码层与解码层均采用6层深度叠加结构的改进型Transformer模型后，显著增强了模型的表征能力和序列处理能力。为评估其性能优势，将其与原Transformer、LSTM和Longformer模型进行对比，不同模型的性能对比实验结果见表6。实验结果表明，改进的Transformer模型通过引入情绪嵌入和定制化位置编码，有效弥补了传统模型在情绪感知和时序一致性方面的不足，显著提升了生成报告的质量。

表6 不同模型的性能对比实验结果

模型	BLEU分数	ROUGE分数
原Transformer	0.197	0.318
Longformer	0.225	0.321
LSTM	0.192	0.235
改进Transformer	0.352	0.482

为了分析不同模型组件对报告生成效果的贡献，本文设计了两个消融实验：去除情绪嵌入层和去除定制化位置编码。消融实验结果见表 7。表 7 的数据进一步验证了情绪嵌入和定制化位置编码对模型性能的提升作用。

表 7　消融实验结果

模型	BLEU 分数	ROUGE 分数
原 Transformer	0.197	0.318
加入情绪嵌入	0.276	0.307
加入定制化位置编码	0.243	0.325

（1）去除情绪嵌入层

在未引入情绪嵌入时，模型在捕捉客户情绪及投诉语义的核心要点上表现较弱，生成文本对客户态度的反应不够精准，语义完整性存在不足。情绪嵌入层通过提取对话中的情绪信息来增强模型对客户态度的理解，从而提高报告的情感层次和人性化。在加入情绪嵌入后，模型的 BLEU 分数从 0.197 提升至 0.276，增幅达 40.1%。这一变化表明，情绪信息的引入显著提升了模型生成文本的准确性与自然性，尤其在理解客户情绪和诉求方面具有关键作用。具体来说，情绪嵌入增强了生成文本的情感色彩，使得报告能够更贴合实际的客户需求和情感表达。

（2）去除定制化位置编码

定制化位置编码的引入进一步增强了模型对投诉时间线索的把握能力，使其在多轮对话中能够更准确地刻画客户诉求随时间变化的逻辑关系。在加入定制化位置编码后，ROUGE 分数从 0.318 提升至 0.325，这表明模型在理解投诉内容的时序结构和事件逻辑时发生了显著变化。定制化位置编码提升了报告生成内容的时间一致性和语义连贯性，使报告能够更好地反映出客户投诉的前后关系及问题解决的过程。

值得注意的是，情绪嵌入与位置编码的协同作用大幅提升了模型的综合性能。在两个组件的共同作用下，模型的 BLEU 分数达到 0.352，ROUGE 分数提升至 0.482，这表明，情绪嵌入增强了模型对客户情感的理解，而定制化位置编码提升了模型在处理长文本和多轮对话时的时间逻辑性。两者结合，使得生成的报告不仅在语义上更为准确，而且在时序上更加符合客户诉求的变化规律。通过将情绪嵌入与位置编码结合使用，模型能够综合考虑语义和时间两个维度，从而生成更加准确、流畅且符合实际需求的重投报告。消融实验柱状图如图 7 所示，图 7 直观展示了不同模型设置下的 BLEU 分数和 ROUGE 分数变化。由图 7 可以明显观察到，情绪嵌入和定制化位置编码不仅各自带来性能提升，两者的协同使用更是显著增强了生成报告的语义准确性与时间逻辑性。

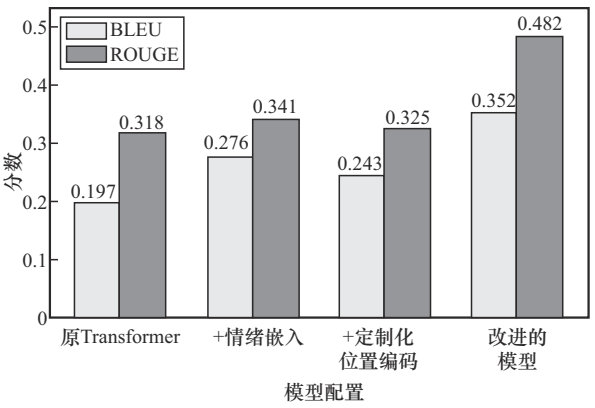


图 7　消融实验柱状图

应用改进后的模型生成重投报告后，报告处理效率显著提升，同时审核部门的满意度也得到明显提高，应用成效见表 8。具体而言，自动生成的报告大幅缩短了重投报告的流转处理时间。传统人工撰写重投报告平均花费约 19 min，而自动生成的报告（包括输入数据导入的时间）平均仅花费 2 min，效率提升了 89%。为了进一步提高报告的一次性通过率，即便增加人工核对环节，全流程时间平均仅为 8 min，相较传统方法仍然节省了 58% 的时间。此外，自动生成的重投

报告在总结用户重复投诉原因和最终处理方案方面更加精准。相比于人工撰写的重投报告一次性通过率仅为76%，自动生成报告通过率为82%，辅以人工校验后一次性通过率提升至90%。

表8 应用成效

生成方式	平均花费时间/min	提升效率	一次性通过率
人工生成	19	—	76%
模型自动生成	2	89%	82%
模型自动生成+人工核验	8	58%	90%

4 结束语

本文提出了一种基于改进Transformer模型的自动化生成方法，以解决电信行业重投报告生成中效率低、质量不一致的问题。通过在模型中引入情绪嵌入和定制化位置编码，本文的改进模型能够更好地捕捉客户的情绪变化和时序信息，在生成报告时体现对客户诉求的深度理解。实验结果表明，与传统方法相比，改进的Transformer模型在BLEU和ROUGE指标上均表现出显著优势，生成报告的语义流畅性和信息完整性得到大幅提升。应用于实际业务后，模型展现了显著的效率和质量优势。自动生成的重投报告大幅缩短了处理时间，同时显著提升了一次性通过率，有效降低了人工撰写中存在的主观性和误差对企业造成的风险，进一步提升了企业的服务效率与审核部门的满意度。未来的研究将致力于优化模型在长序列文本处理上的能力，并探索将更多特征（如客户历史行为数据）融入模型，以进一步提升报告生成的准确性和实用性，为电信智能客服发展提供更高效和可靠的解决方案。

参考文献：

[1] 曹璐. 基于价值链的移动运营服务质量管理研究[D]. 武汉: 武汉理工大学, 2012.
CAO L. Research on service quality management of mobile operation based on value chain[D]. Wuhan: Wuhan University of Technology, 2012.

[2] 李季, 张帅, 周静. 服务与需求的匹配度对客户流失的影响研究: 基于电信行业的客户数据实验[J]. 管理评论, 2020, 32(5): 192-204.
LI J, ZHANG S, ZHOU J. The impact of matching between service and demand on customer churn: evidence from telecommunication industry[J]. Management Review, 2020, 32(5): 192-204.

[3] 袁琳, 孙巍, 马晓敏, 等. 图模型框架下的报道性新闻自动摘要方法研究[J]. 图书情报工作, 2024, 68(17): 122-135.
YUAN L, SUN W, MA X M, et al. Research on automatic summary methods for reportable news under the graph model framework[J]. Library and Information Service, 2024, 68(17): 122-135.

[4] SUTSKEVER I, VINYALS O, LE Q V. Sequence to sequence learning with neural networks[J]. arXiv preprint, 2014: 1409.3215v3.

[5] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. arXiv preprint, 2017: 1706.03762.

[6] 陈嘉鸿, 黄国恒, 谭喆. 基于跨模态差异注意力的医学报告生成[J]. 广东工业大学学报, 2025, 42(1): 70-78.
CHEN J H, HUANG G H, TAN Z. Cross-modal discrepancy attention network for medical report generation[J]. Journal of Guangdong University of Technology, 2025, 42(1): 70-78.

[7] 谭金源, 刁宇峰, 杨亮, 等. 基于BERT-SUMOPN模型的抽取-生成式文本自动摘要[J]. 山东大学学报(理学版), 2021, 56(7): 82-90.
TAN J Y, DIAO Y F, YANG L, et al. Extractive-abstractive text automatic summary based on BERT-SUMOPN model[J]. Journal of Shandong University (Natural Science), 2021, 56(7): 82-90.

[8] PAREKHR, PATELN P, THAKKARN, et al. DL-GuesS: deep learning and sentiment analysis-based cryptocurrency price prediction[J]. IEEE Access, 2022,10: 35398-35409.

[9] GHOSH P, SAMANTA O, GOTO T, et al. Sales forecasting of overrated products: fine tuning of customer's rating by integrating sentiment analysis[J]. IEEE Access, 2024, 12: 69578-69592.

[10] BORGA, BOLDTM. Using VADER sentiment and SVM for predicting customer response sentiment[J]. Expert Systems with Applications, 2020,162: 113746.

[11] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint, 2018: 1810.04805.

[12] 张涛源, 谢新林, 谢刚, 等. 融合Transformer的带钢缺陷实时检测算法[J]. 计算机工程与应用, 2023, 59(16): 232-239.
ZHANG T Y, XIE X L, XIE G, et al. Real-time strip steel defect detection algorithm fused with Transformer[J]. Computer Engineering and Applications, 2023, 59(16): 232-239.

[13] LIU F C, GAO C Q, CHEN F, et al. Infrared small and dim target detection with transformer under complex backgrounds[J]. IEEE Transactions on Image Processing, 2023, 32: 5921-5932.

[14] 潘芳, 张会兵, 董俊超, 等. 基于高效Transformer的中文在线课程评论方面情感分析[J]. 计算机科学, 2021, 48(S1): 264-269.
PAN F, ZHANG H B, DONG J C, et al. Aspect sentiment analysis of Chinese online course review based on efficient Transformer[J]. Computer Science, 2021, 48(S1): 264-269.

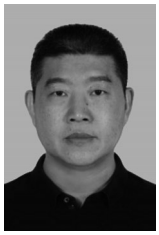
[15] BELTAGY I, PETERS M E, COHAN A. Longformer: the long-document transformer[EB]. 2020.

[16] 廖俊伟. 深度学习大模型时代的自然语言生成技术研究[D]. 成都: 电子科技大学, 2023.
LIAO J W. Research on natural language generation technology in the era of deep learning big model[D]. Chengdu: University of Electronic Science and Technology of China, 2023.

[17] 姜雨娇, 黄铝文, 荚子萌. 基于IMGRU-Seq2Seq的自动问答方法研究[J]. 计算机应用与软件, 2024, 41(6): 215-222, 256.
JIANG Y J, HUANG L W, JIA Z M. Automatic question answering method based on IMGRU-Seq2Seq[J]. Computer Applications and Software, 2024, 41(6): 215-222, 256.

[作者简介]

梁伟明（1971− ），男，中国电信股份有限公司广州分公司客户服务部总经理、马晓亮劳模和工匠人才创新工作室成员，主要研究方向为智能客服和自然语言处理。



肖军（1974− ），男，中数通信息有限公司业务总监、马晓亮劳模和工匠人才创新工作室成员，主要研究方向为基于大数据和人工智能的客服数字化建设。

马晓亮（1973− ），男，博士，中国电信股份有限公司广州分公司副总经理、正高级工程师，马晓亮劳模和工匠人才创新工作室领衔人，主要研究方向为人工智能、自然语言处理和数据安全保护等。

辛盛（1975− ），男，中数通信息有限公司技术总监、马晓亮劳模和工匠人才创新工作室成员，主要研究方向为人工智能和自然语言处理等。

徐荣彬（1988− ），男，中数通信息有限公司项目部经理、马晓亮劳模和工匠人才创新工作室成员，主要研究方向为数据分析与AI数字化工具。