融合梯度与注意力机制的可解释性在线协作讨论交互文本分类方法

邵 峥[1*] 李金鹏[1] 王 勇[1] 何常乐[1]
SHAO Zheng LI Jinpeng WANG Yong HE Changle

摘 要 深度学习模型在在线协作讨论交互文本分类任务中取得了显著性能，但其决策过程的不透明性严重制约了该类模型在教育领域中的可信应用。基于此，文章提出了一种融合梯度与注意力机制的可解释性在线协作讨论交互文本分类方法。以 CNN-BiGRU 网络为基础分类模型，通过将卷积层与循环层得到的梯度重要性与注意力权重进行非线性融合，以得到词级重要性分数，从而在保证分类精度的同时提供词粒度的可解释依据。从分类性能与解释质量两个维度进行评估，实验结果显示所提出的可解释性分类方法在 Macro-F_1 上达到 87.47%，优于基线模型。在可解释性上，该方法的 AOPC 值和 PAAC 值分别为 44.16% 和 38.79%，均优于其他可解释性方法。

关键词 可解释性；注意力机制；梯度；非线性融合

doi：10.3969/j.issn.1672-9528.2026.06.043

0 引言

近年来，深度学习在自然语言处理（NLP）领域取得了突破性进展，比如在文本分类[1]、情感分析[2]和关系抽取等任务中都展现出了超越传统机器学习方法的性能。然而，随着这些模型的规模不断扩大以及神经网络深度的增加，其表现出的“黑箱”行为也日益明显，内部决策机制难以被理解和解释，严重限制了其在高信任场景中的应用与推广。例如在协作讨论场景下，教师不仅需要了解模型根据学习者发布的协作文本类型所做出的决策，更要了解决策的依据，以便

1. 信阳学院计算机与人工智能学院 河南信阳 464000
[基金项目] 信阳学院校级科研项目资助（2025-XJLYB-004）

Piscataway:IEEE,2016:779-788.
[4] 龚航 , 刘培顺 . 夜间行驶车辆远光灯检测方法 [J]. 计算机科学 ,2021,48(12):256-263.
[5] 周国娟 . 联邦学习发展现状与展望 [J]. 网络安全技术与应用 , 2025(10):61-64.
[6] 陈丹霞 , 曾鹏 . 交通数据隐私保护技术研究及其在智能交通系统中的应用 [J]. 中国高新科技 ,2024(9):69-71.
[7]Wang C Y, Bochkovskiy A, Liao H Y M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway:IEEE,2003:7464-7475.
[8]Li Chuyi, Li Lulu, Jiang Hongliang, et al.YOLOv6: a single-stage object detection framework for industrial applications[PP/OL].(2022-09-07)[2026-05-28].https://doi.org/10.48550/arXiv.2209.02976.
[9]Hu Jie, Shen Li, Sun Gang. Squeeze-and-excitation networks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, UT, USA: IEEE, 2018: 7132-7141.
[10] 周秋红 . 基于改进 SSD 模型的道路病害检测研究 [J]. 黑龙江交通科技 , 2023, 46(4): 30-32.
[11]He Kaiming, Zhang Xiangyu, Ren Shaoqing, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1904-1916.
[12] 金昊婴 . 面向分布式机器学习训练的通信优化技术探析 [J]. 长江信息通信 , 2025, 38(8): 52-54.

【作者简介】
龚航（1995—），男，广东韶关人，硕士研究生，研究方向：人工智能算法与应用。

（收稿日期：2025-11-07 修回日期：2026-06-08）

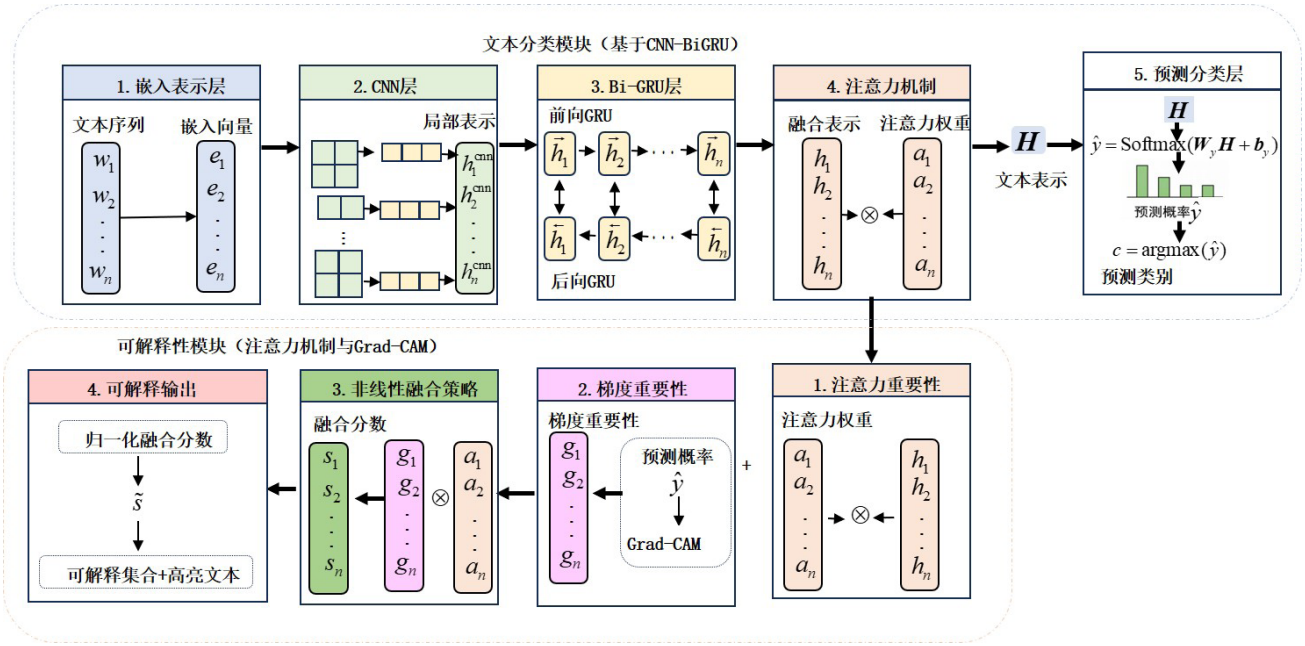

图 1 融合梯度与注意力机制的可解释性方法

对学习者给出反馈时能提供高信任度的支持[3]。因此，如何在保障文本分类模型预测精度的同时，兼顾模型解释结果的有效性与可靠性，是人工智能领域亟待解决的关键问题之一。可解释人工智能（XAI）通过揭示模型预测决策的依据，让用户更加理解和信任其决策结果[4]。而现有 XAI 方法一般分为两大类：一类是以 LIME[5] 和 SHAP[6] 为代表的模型无关型方法，通过拟合可解释的代理模型提供事后解释；另一类是以梯度、激活图或注意力权重为代表的模型特定型方法，可以直接生成解释。但是好的解释方法不仅要有较好的可理解性，还应能忠诚地反映模型的真实决策过程[7]。

注意力机制在模型决策过程中的权重分布可以作为可解释方法[8]。但其反映的是信息在文本序列间的分布，未能揭示输入特征对输出结果的因果贡献[9]。而基于梯度信息的方法虽然可以利用模型输出相对于输入特征的梯度，而量化对预测结果的贡献程度。例如，Selvaraju 等[10] 首次提出基于梯度的方法 Grad-CAM，利用卷积层的梯度信号，以生成类别特异性的空间激活图，后被有效迁移至文本分类模型。然而，Grad-CAM 方法容易受到噪声干扰，且对输入扰动较敏感，其解释的稳定性较难保证。

鉴于此，本文提出融合注意力机制与梯度的可解释性在线协作讨论交互文本分类方法。该方法在 CNN-BiGRU 分类模型上通过融合 Grad-CAM 梯度归因信号与注意力上下文权重，互补性地捕捉梯度的局部敏感性与注意力的全局上下文信息，以提高模型的精度与可解释性。同时，本文引入非线性融合策略，不仅提升分类精度，还能生成更加忠实、鲁棒的词级重要性解释。通过对比实验与消融实验，验证该方法在分类性能与可解释性上的有效性。

1 方法

本文提出一种融合注意力机制与梯度 Grad-CAM 的可解释性在线协作讨论交互文本分类框架，如图 1 所示。该框架由两个核心模块构成：文本分类模块与可解释性模块。文本分类模块采用 CNN-BiGRU 模型，通过对输入文本进行特征提取，分别捕获局部语义特征与序列上下文信息，进而生成分类预测。可解释性模块整合两类互补信息源：一是为前向传播过程中获取的注意力权重；二是基于 Grad-CAM 的梯度重要性评分。为进一步提升关键词元的判别能力，本文引入非线性融合策略，在词级层面将注意力信息与梯度信息整合为统一的重要性评分。

1.1 问题定义

在在线协作讨论交互场景中，给定文本序列 $X=\{w_1, w_2,\cdots, w_n\}$，其中 w_i 为第 i 个词元。文本分类模块的任务可形式化为学习映射函数：$f_\theta: X\to c$，其中 $f_\theta(\cdot)$ 是参数化分类模型，c 是类别集合。可解释性模块旨在通过融合注意力权重 a_i 与梯度重要性 g_i，构建词级重要性函数 $s_i=F(a_i, g_i)$，其中 s_i 表示词元 w_i 对预测结果的贡献程度，从而为模型预测结果分配词级重要性分数。

1.2 文本分类模块

（1）嵌入表示层

假设文本序列为 $X=\{w_1, w_2,\cdots, w_n\}$，其中序列长度为 n，先使用 one-hot 编码将序列向量化，并映射到嵌入层中生成嵌入向量 $\boldsymbol{e}_i\in\mathbb{R}^d$。最后将生成的嵌入向量拼接形成嵌入矩阵

$\boldsymbol{E}=\{\boldsymbol{e}_1,\boldsymbol{e}_2,\cdots,\boldsymbol{e}_n\}$，其中维度为 d，作为文本分类模型的输入。

（2）CNN 层

为捕获文本中的局部语义模式，卷积神经网络（convolutional neural network,CNN）通过滑动窗口在向量序列上进行卷积操作，从而提取不同位置的局部 n-gram 特征，得到对应的局部表示 $\boldsymbol{h}_t^{\text{cnn}}$，用公式表示为：

$$\boldsymbol{h}_i^{\text{cnn}}=\sigma(\boldsymbol{W}_c\cdot\boldsymbol{e}_{i:i+k-1}+\boldsymbol{b}_c) \tag{1}$$

式中：$\boldsymbol{h}_t^{\text{cnn}}\in\mathbb{R}^{d_c}$ 是 CNN 输出特征，$\boldsymbol{e}_{i:i+k-1}$ 是连续 k 个词向量拼接，$\sigma(\cdot)$ 是非线性激活函数，$\boldsymbol{W}_c$ 是卷积核权重参数，$\boldsymbol{b}_c$ 是偏置项。

（3）BiGRU 层

为充分捕获文本序列中的上下文依赖关系，本文使用双向门控循环单元（bidirectional gated recurrent unit, BiGRU）从前后两个方向对序列进行编码分别得到前向隐藏状态$\vec{\boldsymbol{h}}_i$和后向隐藏状态$\overleftarrow{\boldsymbol{h}}_i$。最后，在每个位置 i，将前后两个隐藏状态进行拼接，得到该位置的上下文表示 $\boldsymbol{h}_i^{\text{gru}}$，用公式分别表示为：

$$\vec{\boldsymbol{h}}_i=\text{GRU}_f(\boldsymbol{e}_i) \tag{2}$$

$$\overleftarrow{\boldsymbol{h}}_i=\text{GRU}_b(\boldsymbol{e}_i) \tag{3}$$

$$\boldsymbol{h}_i^{\text{gru}}=[\vec{\boldsymbol{h}}_i;\overleftarrow{\boldsymbol{h}}_i] \tag{4}$$

式中：$[\cdot;\cdot]$ 表示向量拼接操作，$\boldsymbol{h}_i^{\text{gru}}\in\mathbb{R}^{d_g}$ 表示上下文表示。

（4）注意力机制

为突出对分类任务具有关键作用的词元，本文在特征融合表示 $\boldsymbol{h}_i=[\boldsymbol{h}_i^{\text{cnn}};\boldsymbol{h}_i^{\text{gru}}]$ 的基础上引入注意力机制。具体而言，模型通过学习一个注意力分布，为每个词元分配对应的重要性权重 $\boldsymbol{a}_i$，以刻画其在整体语义表示中的贡献程度，用公式表示为：

$$\boldsymbol{a}_i=\frac{\exp(\boldsymbol{v}^{\text{T}}\tanh(\boldsymbol{W}_a\boldsymbol{h}_i+\boldsymbol{b}_a))}{\sum_{j=1}^{n}\exp(\boldsymbol{v}^{\text{T}}\tanh(\boldsymbol{W}_a\boldsymbol{h}_i+\boldsymbol{b}_a))} \tag{5}$$

式中：$\boldsymbol{W}_a$ 为注意力权重矩阵，$\boldsymbol{v}$ 为上下文向量，$\boldsymbol{a}_i$ 为第 i 个词的注意力权重，$\boldsymbol{h}_i$ 为融合后的词级表示，$\boldsymbol{b}_a$ 为偏置项。

在得到注意力权重后，对所有词级表示进行加权聚合，得到文本的整体表示 $\boldsymbol{H}$，用公式表示为：

$$\boldsymbol{H}=\sum_{i=1}^{n}\boldsymbol{a}_i\boldsymbol{h}_i \tag{6}$$

（5）预测分类层

在预测分类层中，首先将注意力加权得到的文本表示向量 $\boldsymbol{H}$ 输入到全连接层进行线性变换，然后通过 Softmax 函数将输出映射为类别概率分布。模型最终选择概率最大的类别作为预测结果，用公式表示为：

$$\hat{y}=\text{Softmax}(\boldsymbol{W}_y\boldsymbol{H}+\boldsymbol{b}_y) \tag{7}$$

$$c=\text{argmax}(\hat{y}) \tag{8}$$

式中：$\boldsymbol{W}_y$ 为权重矩阵，$\boldsymbol{b}_y$ 为偏置向量，$\hat{y}$为预测概率分布，c 为预测类别。

1.3 可解释模块

（1）注意力重要性

本文直接使用在分类模块得到的注意力权重 $\boldsymbol{a}_i\in[0,1]$ 作为全局重要性指标，用于衡量各词元对模型预测结果的影响。权重值越大，表示该词元在构建文本表示时所占的比重越高，对最终分类决策的影响也越显著。

（2）梯度重要性

Grad-CAM 方法使用获得的模型预测结果，计算目标类别输出对每个词级表示 $\boldsymbol{h}_i$ 的梯度，并结合对应的特征表示，得到词级的梯度重要性分数 g_i，计算公式如下所示。通过对词元的重要性计算，以揭示模型输出对输入特征的敏感性。

$$g_i=\text{ReLU}(\frac{\partial\hat{y}}{\partial\boldsymbol{h}_i}\cdot\boldsymbol{h}_i) \tag{9}$$

式中：g_i 是第 i 个词元的梯度重要性评分，$\hat{y}$是对目标类别的预测输出概率。

（3）非线性融合

本文使用非线性增强将注意力和 Grad-CAM 进行融合得到最终的词级重要性。为了解决 Grad-CAM 不稳定、易受到干扰的问题，本文把梯度重要性 g_i 作为基础项，利用注意力权重 $\boldsymbol{a}_i$ 对其进行非线性增强，让序列中更为重要的词获得更高的重要性分数，计算公式如下所示。此外，为了提升模型对关键特征词的识别能力，还引入可调系数 α 和 β 以控制融合强度与非线性程度，从而提升模型对关键特征的区分能力。

$$s_i=g_i\cdot(1+\alpha\boldsymbol{a}_i)^{\beta} \tag{10}$$

式中：s_i 为融合后第 i 个词的重要性分数，α 是注意力重要性调节系数，控制注意力机制对梯度的影响程度，β 是非线性增强系数，控制重要性放大的强度。

（4）可解释输出

融合后得到的重要性表示尺度不统一，因此，本文对重要性分数 s_i 进行归一化，用公式表示为：

$$\tilde{s}_i=\frac{s_i-\min(s)}{\max(s)-\min(s)} \tag{11}$$

归一化后的重要性分数$\tilde{s}_i\in[0,1]$，生成词级解释结果，用公式表示为：

$$\varepsilon=\{(\boldsymbol{w}_i,\tilde{s}_i)\}_{i=1}^{n} \tag{12}$$

式中：ε 为可解释集合，每个词都对应了一个重要性分数。

2 实验

2.1 数据集

本研究采用的数据集是从慕课网所开设的三门国家精品在线协作学习课程的问题讨论区中收集到的 24 387 条数据。同时，采用郑娅峰等 [11] 在相关研究中提出的交互文本分类体

系，该体系将文本分为六大类，分别是陈述类、协商类、提问类、管理类、情感类和其他类。本研究采用人工标注的方式对 24 387 条在线讨论交互文本进行分类，并对每一类交互文本随机抽取 2 000 条按照 8:2 的比例划分为训练集与测试集。具体细节如表 1 所示。

表 1 数据集具体细节

类别	类别名称	训练集	测试集
C1	陈述类	1 600	400
C2	协商类	1 600	400
C3	提问类	1 600	400
C4	管理类	1 600	400
C5	情感类	1 600	400
C6	其他类	1 600	400

2.2 参数设置

由于在线协作讨论文本的语料长度较短，本研究在所有模型中设置文本的最大长度为 40，最大词为 8 633，嵌入层维度为 128，卷积核大小为 3，隐藏层维度为 128，注意力调节系数 α 为 0.3，非线性增强系数 β 为 1.5，温度参数 τ 为 0.2。分类器使用 Adam 优化器结合交叉熵损失函数进行训练，学习率为 1e-3，丢弃率为 0.3，训练轮数为 20，批处理数为 32。

2.3 评估指标

本文采用分类模型评估常用的指标，包括准确度（Accuracy），精确度（Precision）和宏平均 F_1 值（Macro-F_1）。此外，本文采用基于 AUC 曲线下面积计算 AOPC（Area Over the Perturbation Curve） 值 [12] 和 PAAC（Percentage Area Above the Curve）值 [13]。其中，AOPC 值用于衡量在特征移除过程中模型性能下降的整体程度。AOPC 值越大，表示移除重要特征模型性能下降越明显，说明该解释方法具有更高的忠实性。而 PAAC 值反映模型在移除低重要性特征过程中的性能保持能力。PAAC 值越大，解释方法越有效，即解释的判别能力越强。

2.4 基线方法

本文使用 TextCNN、BiLSTM、BiGRU、CNN-BiGRU（无注意力）、与 CNN-BiGRU（有注意力）模型作为基线方法，以评估分类性能。同时，使用 Attention、Grad-CAM、SHAP、LIME 作为可解释性基线方法，以评估可解释性。

2.5 实验结果与分析

（1）分类性能对比

为了验证所提方法的有效性，将所有方法在测试集上的分类性能进行对比，结果如表 2 所示，加粗为最优。可以看到，本文方法的 Macro-F_1 达到了 87.47%，优于相同分类模型架构的 CNN-BiGRU（有注意力）3.28 个百分点。表明该方法不仅提供了更好的解释，其训练方法受到梯度影响也提升了分类精度。相比单一架构模型，本文方法在精确度与准确度上均有所提升。

表 2 各方法分类性能对比

单位：%

方法	Macro-F_1	Accuracy	Precision
TextCNN	76.50	78.06	78.41%
BiLSTM	77.58	78.23	78.91
BiGRU	78.14	80.06	80.65
CNN-BiGRU（无注意力）	82.56	83.37	84.12
CNN-BiGRU（有注意力）	84.19	85.02	85.74
本文方法	**87.47**	**88.59**	**88.92**

（2）可解释性对比

为了验证所提方法的有效性，将所有方法在测试集上的可解释性指标进行对比，结果如表 3 所示，加粗为最优。

表 3 各可解释性方法对比

单位：%

方法	AOPC	APPC
Attention	31.82	29.79
Grad-CAM	34.71	32.14
SHAP	37.46	35.22
LIME	34.37	31.84
本文方法	**44.16**	**38.79**

从表 3 中可知，注意力权重 AOPC 值最低，说明注意力权重不应被视为可靠的词汇重要性指标。梯度方法则优于注意力机制。而本文所提方法最优，AOPC 值为 44.16%，显著优于所有单一性的可解释性方法。SHAP 值稍弱于本文方法，但计算时间较长，在大规模学习场景下不具有实用性。

（3）消融实验

为了进一步展示本文方法在分类性能与可解释性上的有效性，进行模型组件的消融实验，结果如表 4 所示。从表 4 可知，每个组件的去除均导致分类性能的下降。其中，去除 BiGRU 导致 Macro-F_1 下降了 4.8 个百分点，说明序列上下文建模对协作讨论文本分类最为关键。而只有 CNN-BiGRU 分类模型后，AOPC 值下降了 15.51 个百分点，说明本文引入的可解释性方法是有效的。

表 4 消融实验结果

单位：%

方法	Macro-F_1	AOPC
本文方法	**87.47**	**44.16**
BiGRU+ 注意力 + 梯度	85.03	38.94
CNN+ 注意力 + 梯度	82.67	34.10
CNN-BiGRU+ 梯度	86.24	37.44
CNN-BiGRU+ 注意力	84.19	35.35
CNN-BiGRU	82.56	28.65

3 总结

本文针对在线协作讨论交互文本分类任务中深度学习模型可解释性不足的问题，提出了一种融合注意力机制与梯度 Grad-CAM 信息的可解释性文本分类方法。同时，引入非线性融合策略，实现了对模型决策依据的精细刻画。在真实的在线协作讨论交互文本数据集上的实验结果表明，所提出的可解释性分类方法在 Macor-F_1 值最优，同时，在 AOPC 值及 APPC 值上也均优于单一解释方法。说明本文方法不仅能提升分类模型的精确度，也能提供可靠的可解释依据。未来可考虑在以下方面继续研究：（1）结合句法结构探索更加细粒度的解释机制；（2）将本文方法扩展至更复杂的预训练模型，以验证其在大规模任务中的泛化能力。

参考文献：

[1] Niu Zhaoyang, Zhong Guoqiang, Yu Hui. A review on the attention mechanism of deep learning[J]. Neurocomputing, 2021, 452: 48-62.

[2] Lecun Y, Bengio Y, Hinton G. Deep learning[J]. Nature, 2015, 521: 436-444.

[3] Hmelo-silver C E. Problem-based learning: what and how do students learn?[J]. Educational Psychology Review, 2004, 16: 235-266.

[4] Adadi A, Berrada M. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI)[J]. IEEE Access, 2018, 6: 52138-52160.

[5] Montavon G, Samek W, Müller K R. Methods for interpreting and understanding deep neural networks[J]. Digital Signal Processing, 2018, 73: 1-15.

[6] Liu Gang,Guo Jiabao. Bidirectional LSTM with attention mechanism and convolutional layer for text classification[J]. Neurocomputing, 2019, 337: 325-338.

[7] Jacovi A, Goldberg Y. Towards faithfully interpretable NLP systems: on the concept of explanations as statements of responsibility[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.Brussels:ACL,2020: 4198-4205.

[8] Yu Yong, Si Xiaosheng, Hu Changhua, et al. A review of recurrent neural networks: LSTM cells and network architectures[J]. Neural Computation, 2019, 31(7): 1235-1270.

[9] Luo Siwen, Ivison H, Han S C, et al. Local interpretations for explainable natural language processing: a survey[J]. ACM Computing Surveys, 2024, 56(9): 1-36.

[10] Selvaraju R R, Cogswell M, DAS A, et al. Grad-CAM: visual explanations from deep networks via gradient-based localization[C]// 2017 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2017: 618-626.

[11] 郑娅峰 , 张巧荣 , 李艳燕 . 协作问题解决讨论活动中行为模式自动化挖掘方法研究 [J]. 现代教育技术 , 2020,30(2): 71-78.

[12] Edin J, Motzfeldt A G, Christensen C L, et al. Normalized AOPC: fixing misleading faithfulness metrics for feature attribution explainability[C]//Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics. Brussels:ACL,2025:1715-1730.

[13] Šimić I, Veas E, Sabol V. A comprehensive analysis of perturbation methods in explainable AI feature attribution validation for neural time series classifiers[J]. Scientific Reports, 2025, 15: 26607.

【作者简介】

邵峥（1995—），通信作者（email:1065605723@qq.com），女，河南驻马店人，硕士，助教，研究方向：自然语言处理。

李金鹏（1996—），男，河南驻马店人，硕士，助教，研究方向：智能编曲。

王勇（1997—），男，河南信阳人，硕士，助教，研究方向：信息安全。

何常乐（1997—），男，河南信阳人，硕士，助教，研究方向：机器学习。

（收稿日期：2026-05-09 修回日期：2026-06-18）