编者按：2023 年中国情报学年会暨情报学与情报工作发展论坛、第十三届全国情报学博士生学术论坛在湖南省长沙市成功举办。会议共征集论文近 600 篇，《数据分析与知识发现》期刊邀请部分优秀论文形成会议专题出版。

专　题

以可解释工具重探基于深度学习的谣言检测*

贺国秀　任佳渝　李宗耀　林晨曦　蔚海燕
（华东师范大学经济与管理学院 上海 200062）

摘要：【目的】探究基于内容的深度谣言检测模型能否真正识别谣言的关键语义。【方法】基于谣言检测任务的中英文基准数据集，本文分别利用基于局部代理模型的可解释工具 LIME 和基于合作博弈论的可解释工具 SHAP，分析 BERT 模型所识别出的关键特征，并判断其是否能反映谣言特性。【结果】可解释工具在不同模型与数据集上计算得出的关键特征差异性较大，无法辨别模型识别的重要特征和谣言之间的语义关系。【局限】本文验证的数据集和模型数量都十分有限。【结论】基于深度学习的谣言检测模型仅拟合了训练集的特征，面向多样的真实场景缺少足够的泛化性和可解释性。

关键词：谣言检测　可解释机器学习　深度学习　LIME　SHAP
分类号：TP393　G35
DOI：10.11925/infotech.2096-3467.2023.0684
引用本文：贺国秀，任佳渝，李宗耀等. 以可解释工具重探基于深度学习的谣言检测[J]. 数据分析与知识发现，2024，8（4）：1-13. (He Guoxiu, Ren Jiayu, Li Zongyao, et al. Revisiting Deep Learning-based Rumor Detection Models with Interpretable Tools[J]. Data Analysis and Knowledge Discovery, 2024, 8(4): 1-13.)

1　引　言

随着互联网技术和社交媒体的快速发展，信息传播的范围不断扩大，影响力不断增强。由于社交媒体的互动性和匿名性特点[1]，部分用户或组织发布了虚假或误导信息。这些信息的大量转发，会给社会带来巨大负面影响。例如，新型冠状病毒感染（以下简称“新冠”）疫情期间，大量涌现的疫情谣言向政府舆情治理带来极大挑战[2]。为遏制谣言的传播和发酵，各大社交媒体都建立了一系列信息甄别社区。例如，国内微博官方建立了微博社区管理中心，通过用户举报和人工事实核查的方法进行谣言的查验。但是人工审核方式费时费力，且存在严重滞后性，使得谣言的进一步传播具有可乘之机。

因此，如何高效进行谣言检测（Rumor Detection，RD）具有重要意义。众多学者就此展开研究，并在基准数据集上提出大量性能较高的模型。目前，谣言检测任务主要依赖于机器学习和深度学习技术。传统的机器学习方法主要基于人工特征工程，难以提取谣言的深层语义特征；而深度学习凭借其强大的语义理解能力能够自动捕捉谣言的深层语义特征，极大地提高了检测的准确率[3]，成为当下研究主流。

虽然有些模型已取得很高的精度，但是当下网

通讯作者（Corresponding author）：贺国秀（He Guoxiu），ORCID：0000-0002-1419-7495，E-mail：gxhe@fem.ecnu.edu.cn。

*本文系国家自然科学基金项目（项目编号：72204087）、上海市哲学社会科学规划青年课题（项目编号：2022ETQ001）和中央高校基本科研业务费专项资金资助项目的研究成果之一。

This work is supported by the National Natural Science Foundation of China (Grant No. 72204087), the Shanghai Philosophy and Social Science Planning Youth Project (Grant No. 2022ETQ001), the Fundamental Research Funds for the Central Universities.

络谣言仍屡禁不止。重新评估谣言检测模型的可靠性是进一步提高谣言检测效率的基础和关键。深度学习模型具有复杂的组成结构,目前仍被视为“黑盒”模型。在数据集上进行端到端的训练后,深度学习模型虽然可实现令人满意的性能,但是模型所依赖的特征和推理的路径对使用者来说仍不清楚,模型的可靠性和可信度有待商榷。因此,本文利用可解释工具——沙普利加性解释(Shapley Additive Explanations, SHAP)和事后局部解释(Local Interpretable Model-Agnostic Explanations,LIME)揭示谣言检测模型提取的关键特征,对现有的谣言检测模型进行可靠性评估,以探究谣言检测模型的性能瓶颈。

首先,基于当前流行的机器学习和深度学习理论构建谣言检测模型,并在基准谣言数据集上进行训练和验证;其次,使用LIME和SHAP对检测效果最好的模型进行解释,得到模型所捕捉的关键特征;最后,通过对比不同模型和数据集的结果评估谣言检测模型的可靠性,探讨未来谣言检测任务的研究方向。

2 研究现状

2.1 谣言检测方法研究综述

大多数谣言检测问题可以转化为分类问题。目前,基于特征工程的统计机器学习方法和基于语义内容理解的深度学习方法是谣言检测方法研究的主流。

基于统计机器学习方法的检测模型通过借鉴传播学、语言学、人口统计学等学科理论,在特征选择上不断拓宽。Castillo等[4]通过分析推特(Twitter)数据集的文本特征和用户特征,建立基于支持向量机(Support Vector Machine, SVM)的谣言检测模型。Shu等[5]通过引入用户个人信息和分享行为(如转发、评论等)特征增强了谣言检测能力。此外,通过计算情感倾向得分提取情感特征[6]以识别谣言的方法也取得了一定成效。

在基于深度学习的谣言检测研究中,Ma等[7]首次引入卷积神经网络(Convolutional Neural Network, CNN)。Shen等[8]提出一种基于双向LSTM(Bidirectional-LSTM,Bi-LSTM)和注意力机制的谣言检测模型。Bain等[9]提出基于双向图卷积网络(Graph Convolutional Network, GCN)的谣言检测模型,通过同时使用自下而上和自上而下的传播,分别由自下而上的图卷积网络和自上而下的图卷积网络捕获与谣言有关的特征。

随着预训练语言模型的发展,研究者们将这些模型应用于谣言检测任务。Anggrainingsih等[10]利用BERT的句子嵌入提取推文句子的上下文语义,并揭示推文的具体语言模式。Zhong等[11]使用预训练语言模型进行语义角色的标记,以理解文本及其相互作用,然后使用图模型对网络进行编码以检测谣言。

综上所述,无论是统计机器学习方法还是深度学习方法,在谣言检测任务上,研究重点都离不开特征提取与选择。相较于机器学习方法所提取的相对表层的特征,深度学习方法凭借其巨大的参数量能够捕捉深层次的语义信息。在单文本谣言检测场景中,模型能否真正识别谣言的关键语义尤为重要。因此,探究谣言检测模型所捕捉的谣言特征具有重要意义。

2.2 模型可解释性研究综述

深度学习模型以其复杂的网络结构和强大的表示能力,在谣言检测任务中实现了显著的性能提升。由于其高度非线性化的操作,往往令使用者无法探知模型的决策路径,更无法判断其决策是否具有自然语言含义。此外,虽然深度学习模型在实验环境的训练和测试集上表现出良好的性能,但是面对现实复杂的情况时,缺乏可信决策依据的模型往往无法实现预期的性能,这对于模型的大规模应用是一种阻碍。因此,有效解释深度学习模型对谣言检测具有重要意义。

“可解释性”概念由Kim等[12]首次提出。Miller[13]对“可解释性”做了进一步阐释,“可解释性是人们可以理解决策原因的程度”,这表明可解释性的目标不是完全理解神经网络中所有的内部特征,而是尽可能将可以解释的特征信息分离出来,进行定量分析,以此建立可信任的关系。常见的可解释性方法有4种,分别为基于输入特征、基于激活响应、基于模型内部结构以及基于全局理解。其中,基于输入特征的可解释性方法有LIME[14]、Anchor[15]等;基于激活响应的可解释性方法有Grad-CAM[16]、

Guided Backpropagation等；基于模型内部结构的可解释性方法有 Network Dissection[17]、DeepTaylor[18]等；基于全局理解的可解释性方法有 SHAP[19]、DeepLIFT[20]等。

本文选取两种常用的可解释性方法（LIME与SHAP）进行实验。其中，LIME（Local Interpretable Model-Agnostic Explanation）是一种常用的局部可解释方法。首先，该方法随机采样生成一组与输入样本类似的新数据点；其次，通过模型对新数据点预测结果，构造一个局部的线性模型来解释模型的预测结果。LIME是一种通用性比较强的方法，适用于不同类型的数据，可以在不需要了解模型内部复杂结构的情况下，通过可视化显示哪些输入特征对模型做出决策的影响最大，提供一种直观的、对模型分类决策的解释方法。

考虑到LIME对模型的解释是局部的，本文同时选用SHAP作为全局解释的实验方法。SHAP（Shapley Additive Explanations）方法是 Lundberg等[19]于2017年提出的可解释模型。SHAP源于合作博弈论，其实质是在Shapley Value基础上总结与改良的可加性解释模型。该方法先将每个样本中的所有特征视作贡献者，通过总体和个体的比较计算具体贡献值，再将所有特征的贡献值进行加总，得到模型的预测概率。相比于LIME等其他方法，SHAP既能够进行单个样本的局部解释，也能够对所有样本进行全局解释[21]。SHAP保证了解释的一致性和唯一性，即对于同一模型和同一输入，SHAP值是唯一的。这使得SHAP能够提供一种稳健的方法，有助于理解模型的决策过程。

目前，在经济学、医学、公共管理等领域，已有部分研究使用SHAP解释方法，如易明等[22]融合XGBoost与SHAP对政务新媒体公共价值进行解释。在谣言检测上，曾子明等[23]使用SHAP对多特征融合的谣言传播者识别模型进行解释。然而，由于谣言检测涉及非结构化文本数据，大多数研究者或在使用解释模型时仅挑选了个别样本进行研究，缺乏针对文本总体的研究[24]；或仅选取文本长度、标点数量等文本表层特征，而忽略了深层次的语义信息。本文针对谣言语义特征，既关注具体样本的解释，又关注全局样本的关键特征。

3　谣言检测模型的可解释性方法设计

本文收集谣言检测领域论文中常用的英文及中文数据集，采用多种机器学习和深度学习模型，如多层感知机（MLP）、卷积神经网络（CNN）和BERT等，进行训练和评估。通过比较各模型在这些数据集上的评价指标，选取性能最好的模型。为了评估所训练模型是否能够真正地识别谣言语义，本文运用两种解释性方法（LIME和SHAP）对模型进行解释分析。

本文选取的每个基准数据集均包含谣言和非谣言信息，以及与之相关的元数据。首先，对这些数据集进行预处理——文本清洗、分词、提取词嵌入等。其次，采用MLP、CNN和BERT等深度学习作为谣言检测模型，在相应的数据集上进行训练。记录并比较各模型的准确率、精确率、召回率、F_1值等评价指标。通过对比不同模型的性能，找出更适合谣言检测任务的模型。再次，利用LIME和SHAP方法对模型进行可解释分析，得到关键特征排序。最后，通过对不同模型和数据集的全局对比和个案分析，归纳谣言检测模型捕捉到的关键特征的异同，评估深度谣言检测模型的可靠性和泛化性。本文的研究框架如图1所示。

3.1　谣言检测模型构建

谣言检测问题的本质是一个二分类问题[25]。在单文本场景下，其他如用户特征、互动特征等数据缺失，则无法参与建模，因此，谣言识别模型需关注文本本身的语义特征。语义特征表示文本中的语义信息，已有的语义信息提取方法包含基于词袋的文本向量表示方法和从大规模语料库中训练的词嵌入向量[26]。本文使用基于词频-逆文档频率（Term Frequency-Inverse Document Frequency，TF-IDF）的向量空间模型和BERT预训练模型等将文本转换为机器学习和深度学习能够接受的形式。TF-IDF方法是目前最普遍采用的文本表示方法，考虑了文本间词汇共现情况，能够判断出词语在样本集中的重要程度[27]。BERT模型基于Transformer编码器结构，增加注意力机制表示文本深层语义[28]，已经被证明在多个任务和应用场景中展现优异性能。为进一步改善BERT在中文数据集上的性能，多位学者基

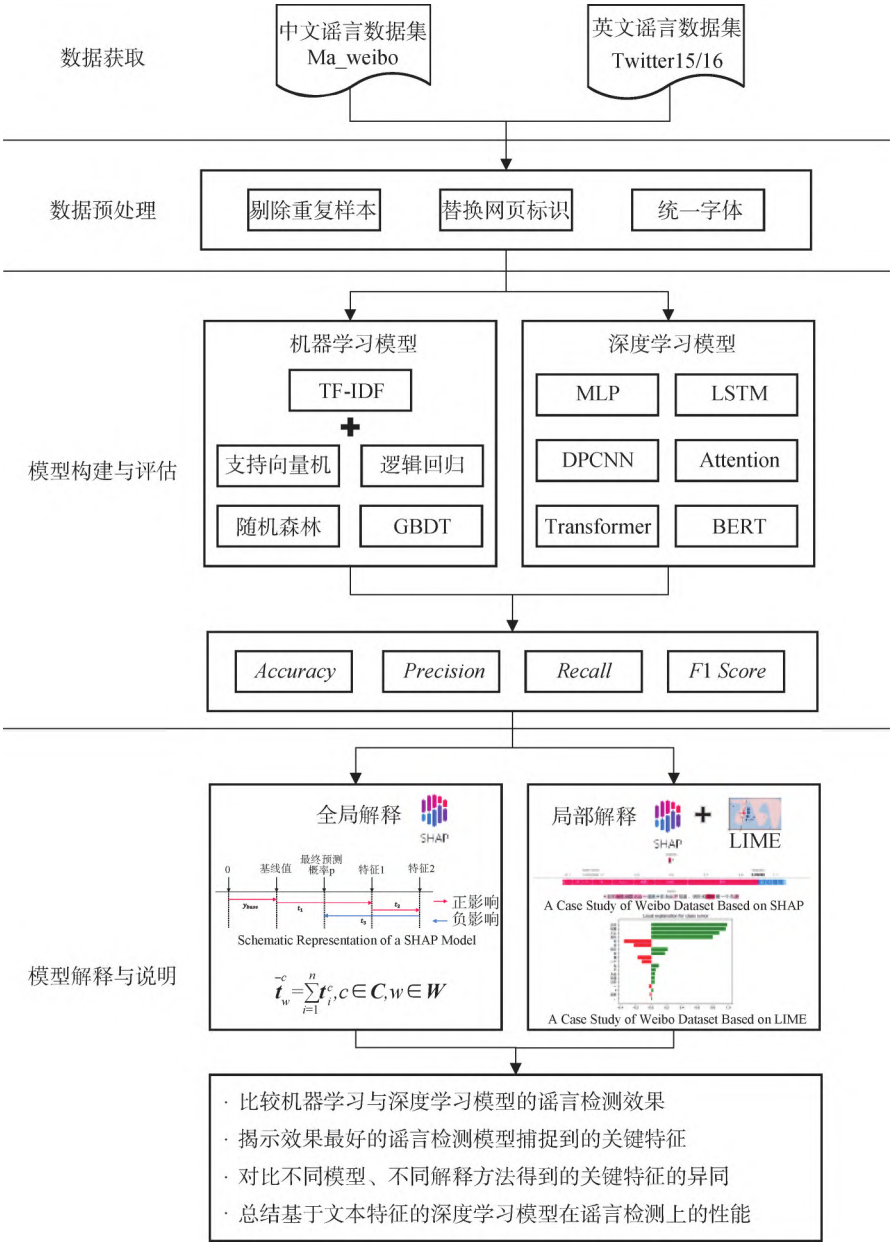

图 1　研究框架

Fig.1　Research Framework

于中文语料库训练了中文BERT模型。考虑到传统BERT以字为单位使得解释较困难，本文采用苏剑林预训练并开源了以词为单位的WoBERT模型[29]，结合中文谣言领域数据微调后生成词嵌入表示。

在机器学习方法上，选择支持向量机、逻辑回归（Logistic Regression，LR）、随机森林（Random Forest，RF）和梯度提升树（Gradient Boosting Decision Tree，GBDT）作为文本分类器。在深度学习方法上，选取多层感知机（Multilayer Perceptron，MLP）、文本卷积神经网络（Convolutional Neural Network，CNN）[30]、深度金字塔卷积神经网络（Deep Pyramid Convolutional Neural Network，DPCNN）[31]、双向长短期记忆网络（Bidirectional Long Short-Term Memory，Bi-LSTM）[32]、基于注意力机制的双向长短

期记忆网络(Bidirectional Long Short-Term Memory with Attention, BiLSTM-Attention)、和 Transformer 等模型，这些模型都使用 pytorch 库自带的 Embedding 方法得到最适合当前数据集的词嵌入表示。而经过微调后的 BERT 词嵌入表示接一个全连接层输出分类概率结果。

3.2　谣言检测模型解释方法

已有研究从统计学角度比较了谣言和非谣言的差异。Vosoughi 等[33]研究发现谣言比正常信息的表述更加猎奇，并且情感倾向更偏向惊奇和厌恶。Solovev 等[34]研究发现在新冠疫情流行期间关于新冠疫情的谣言表达出了更强的厌恶和生气情绪。

鉴于基于机器学习和深度学习的谣言检测模型具备挖掘文本语义的能力，并且具有良好的分类效果，本文提出假设：基于文本内容的机器学习和深度学习模型能够识别谣言的重要特征，不同模型在不同数据集上捕捉的关键语义应当具有一定相似性。本文借助 LIME 和 SHAP 分别对模型的局部和全局进行解释，并使用两种方法计算谣言的关键语义。

(1) 局部解释方法

LIME 核心原理是局部敏感性，即在输入空间的局部区域内，任何复杂的模型都可以被简化为一个线性模型。对于一个输入样本，LIME 会在其周围生成一些类似的数据样本，即“伪数据”，用于训练可解释的模型。对于生成的每个“伪数据”，LIME 会计算其与原始样本之间的距离，并将其转换为权重，越近的数据样本，则权重越大。在生成的“伪数据”集合上，LIME 使用一个可解释的模型(如线性模型)进行训练，得到每个特征的权重。根据每个特征的权重，解释该样本的预测结果。

$$\phi(x) = \arg\min_{g \in G} L(f, g, \pi_x) + \Omega(g) \tag{1}$$

其中，f 是原始模型；g 是可解释的模型；π_x 表示以 x 为中心生成的“伪数据”集合；L 是损失函数；$\Omega(g)$ 是正则化项。通过优化公式(1)，可以得到一个可解释的模型 $\phi(x)$，用于解释原始模型在该位置的预测结果。

(2) 全局解释方法

SHAP 可以对模型局部进行解释，其每条样本的特征值对于模型全局也具备解释意义。

给定一条由 n 个词组组成的文本序列 $\boldsymbol{X} = \{x_1, x_2, \cdots, x_n\}$，对应的标签为 $\boldsymbol{C} = \{0, 1\}$，其中 0 代表谣言，1 代表非谣言。通过 SHAP 工具计算序列词组在每个类别 c 的特征值 $\boldsymbol{T}^c = \{t_1^c, t_2^c, \cdots, t_n^c\}$，$c \in \boldsymbol{C}$，该特征值在两个类别上互为相反数，其中，正值表示对某一类别具有正向贡献，负值表示对某一类别具有负向贡献。整个模型的基线为所有样本标签的均值，定义为 y_{base}，则对于第 i 条样本，最终两个类别的预测概率值 p_i^c 等于各词组的特征值之和与基线的和，如公式(2)所示。

$$p_i^c = y_{base} + t_1^c + t_2^c + \cdots + t_n^c, c \in \boldsymbol{C} \tag{2}$$

针对某一类别每条样本的特征影响力，如图 2 所示。

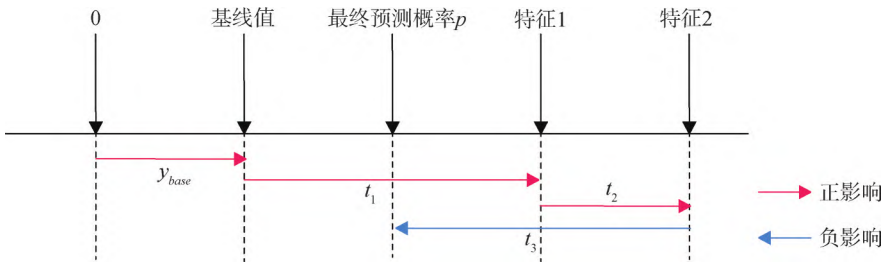

图 2　SHAP 示意图

Fig.2　Schematic Representation of a SHAP Model

得到每条样本的词组特征值后，采用平均值法则根据每个类别中词组的特征平均值大小进行排序，词组的特征平均值的计算如公式(3)所示。

$$\bar{t}_w^c = \sum_{i=1}^{n} t_i^c,\ c \in \boldsymbol{C},\ w \in \boldsymbol{W} \tag{3}$$

其中，$\boldsymbol{W}$ 为数据集的所有词组集合。

4 实验设计

4.1 数据来源及预处理

本文采用两个谣言数据集，即英文数据集（Twitter15/16）和中文数据集（Ma_weibo）。其中，Twitter15/16数据集是基于Twitter15（Liu等[35]）和Twitter16（Ma等[7]）两个参考数据而构建的数据集，包含2014年至2016年间针对给定事件的谣言和非谣言的二元分类英文推文。数据集内容涵盖了政治、娱乐、科技等多个领域，共计2 310条。Ma_weibo数据集是由Ma等[36]于2015年创建的一个中文微博谣言检测数据集，谣言数据从新浪社区管理中心获取，这些谣言被报告为各种不实信息。非谣言数据通过爬取未被报告为谣言的普通主题帖子进行数据收集，这数据涉及社会、政治、经济等多个领域，共包含2 313条谣言和2 351条非谣言数据。

为保证数据集的一致性和可比性，本文对所有数据集进行统一的预处理。首先，移除文本中的特殊字符和表情符号，将HTML标签和URL统一转换为“URL”字符串。对于英文数据集，使用Python的NLTK包进行分词；对于中文数据集，使用Jieba包对文本进行分词。将数据集按照8∶1∶1的比例划分为训练集、验证集和测试集。其中，训练集用于模型训练，验证集用于模型超参数调优，测试集用于评估模型性能。经过上述预处理，得到标准化且具有统一格式的数据集，有利于下文中谣言检测模型的训练和评估。

4.2 评价指标

为全面评估谣言检测模型性能，本文选取的主要评价指标有：准确率（Accuracy）、精确率（Precision）、召回率（Recall）和*F*1值（*F*1 Score）。此外，对于可解释性模型，本文采用Shapley Value与特征权重分别体现SHAP和LIME在模型中不同特征的贡献度。

分类结果的混淆矩阵如表1所示。真正例（True Positive，TP）表示将正样本正确预测为正样本；假正例（False Positive，FP）表示将负样本错误预测为正样本；假负例（False Negative，FN）表示将正样本错误预测为负样本；真负例（True Negative，TN）表示将负样本正确预测为负样本。

表1 混淆矩阵

Table 1 Confusion Matrix

	实际为谣言	实际为非谣言
预测为谣言	TN	FN
预测为非谣言	FP	TP

（1）准确率：所有预测正确的结果占所有结果的比重，如公式（4）所示。在下文中简写为*Acc.*

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

（2）精确率：正确预测为正类的结果占所有预测为正类结果的比重，如公式（5）所示。在下文中简写为*Pre.*

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

（3）召回率：正确预测为正类的结果占所有实际为正类的结果的比重，如公式（6）所示。在下文中简写为*Rec.*

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

（4）*F*1值：精确率和召回率的调和均值，如公式（7）所示。

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{7}$$

4.3 实验设置

本文采用Python的Sklearn包构建机器学习模型，采用Pytorch包构建所涉及的深度神经网络模型。所有实验在Linux操作系统上完成，所使用的GPU型号为NVIDIA A800。对于机器学习模型，TF-IDF的max_feature设置为500，其他机器学习模型使用默认参数；对于深度学习模型，其学习率根据模型设置为0.01~0.05，batch_size设置为16、32或64，embedding_size设置为128，hidden_size设置为32~64，迭代次数设置为50个epoch。深度学习的损失函数均为交叉熵，优化器为Adam。

5 研究结果与分析

5.1 谣言检测模型评估

在数据集Ma_weibo和Twitter15/16上，本文分别使用机器学习和深度学习模型实现谣言检测，在测试集上的实验结果如表2所示。

实验结果表明，在中文数据集中，除了基于TF-

表2　谣言检测模型在测试集上的实验结果

Table 2　Experimental Results of Rumor Detection Models on the Test Dataset

模型		Ma_weibo(%)				Twitter15/16(%)			
		Acc.	*Prec.*	*Rec.*	*F*1	*Acc.*	*Prec.*	*Rec.*	*F*1
机器学习模型	TF-IDF+SVM	87.14	86.52	88.00	87.25	84.25	89.21	82.12	87.25
	TF-IDF+LR	85.28	84.02	87.14	85.53	83.23	89.36	88.34	89.14
	TF-IDF+GBDT	71.71	72.89	69.14	70.97	80.65	78.24	70.12	68.63
	TF-IDF+RF	81.57	84.42	77.43	80.77	82.08	80.12	80.58	80.43
深度学习模型	MLP	88.00	87.78	88.28	88.03	88.44	89.13	88.65	89.13
	CNN	87.86	**89.31**	86.00	87.62	82.08	75.63	83.12	85.31
	RNN	85.28	87.08	82.85	84.92	71.68	70.09	72.36	75.38
	LSTM	86.29	86.71	85.71	86.20	63.01	70.59	65.10	60.00
	DPCNN	84.14	85.25	82.57	83.89	63.31	70.48	68.33	59.95
	Self-Attention	86.14	88.69	82.86	85.67	62.8	70.34	66.25	69.34
	BiLSTM-Attention	86.43	88.06	84.29	86.13	63.34	69.68	66.89	60.88
	Transformer	86.57	87.21	85.71	86.46	62.99	71.18	60.44	60.46
	BERT	**92.00**	89.30	**95.43**	**92.65**	**88.93**	**89.44**	**88.79**	**89.27**

IDF+GBDT 模型，其他模型都实现了 80% 以上的谣言识别准确率；在英文数据集上，模型都是实现了 60%~89% 的谣言识别准确率。这表明机器学习和深度学习在封闭的基准数据集上展现了极高的应用潜力。其中，BERT 模型与机器学习模型和其他深度学习模型相比具有更强的语义特征提取能力。对于中文数据集和英文数据集，基于 BERT 的谣言检测模型的准确率、召回率和 *F*1 值三个指标优于其他模型。另外，除了 BERT 模型外，其他深度学习模型相较于传统机器学习模型在谣言检测任务上没有显著优势，某些机器学习模型的指标甚至优于一些深度学习模型。为了深入探究不同模型之间捕捉特征的差异，下文使用 LIME 和 SHAP 对机器学习和深度学习性能最好的模型进行可解释分析。

5.2　解释模型

在中文数据集上，TF-IDF+SVM 以及 BERT 模型分别在机器学习模型和深度学习模型中实现了最优性能。在英文数据集上，TF-IDF+LR 和 BERT 模型分别在机器学习和深度学习中表现出最佳的性能。因此，本文选择这三个分类模型作为被解释模型。

（1）模型全局解释

首先，使用 SHAP 对模型进行全局解释。由于 SHAP 模型运算量极大，为了降低计算开销，本文在测试集中随机抽取 250 条非谣言样本和 250 条谣言样本进行解释。

在中文数据集中，借助 SHAP 工具计算得到的前 15 个谣言和非谣言关键词组如表 3 所示。TF-IDF+SVM 模型识别的谣言关键特征中，主要是日常生活在微博社区中常见的词语（如“超话”“鸡蛋”“肉松”“孩子”等），除此以外，还包含少部分无实际意义的连接词（如“只有”“最后”等）；BERT 模型识别的谣言关键特征中，同样地，主要是日常生活在微博社区中常见的词语，只有“据说”一词或能跟谣言联系起来。在非谣言的重要特征中，机器学习识别的特征依然主要是常见词语，而或有谣言意味的“网传”一词却对非谣言具有正向贡献。BERT 模型识别的谣言关键特征除了无意义的虚词、助词以外，还出现了“秦始皇”“溥仪”等词。

进一步观察数据集，可以发现，有关“肉松是棉花做的”这一话题的谣言数量在训练集中多达 124 条，占比为 5.3%，在用于解释的测试集中为 13 条，占比为 5.2%；有关“台风山竹”这一话题的谣言数量在训练集中有 42 条，占比为 1.8%，而在测试集中有 5 条，占比为 2%，这表明 TF-IDF+SVM 模型识别的谣言特征在数据集中频数较高。然而，这两个话题均属于该数据集涉及年份的热点新闻，并不具备普适性。同样地，在深度学习所识别的关键词中，“溥仪”（1 次）、“秦

始皇”(1次)、“尼姑”(4次)具有高度偶然性,并不能归纳为谣言或者非谣言的普遍关键语义。

表3 中文数据集关键特征

Table 3 Important Features of the Weibo Dataset

模型	谣言 Top 15	非谣言 Top 15
TF-IDF+SVM	超话,鸡蛋,肉松,疫情,最后,大家,孩子,事件,去世,棉花,少年,不要,山竹,塑料袋,只有	全文,医院,人民,死亡,网传,北京,视频,还是,上海,可怕,下来,乔任梁,女子,里面,转发
BERT	取消,恐惧,火车站,车子,尼姑,升空,据说,民警,江中,打响,哈尔滨,打工,印度,燃烧,禽兽	转正,没事,也门,政治,秦始皇,哈萨克斯坦,没什么,交换,离不开,皇后,仿佛,道歉,溥仪,苦恼,耐心

在Twitter数据集中,借助SHAP工具计算得到的前15个谣言和非谣言关键词组如表4所示。机器学习分类器和深度学习分类器捕捉到的谣言特征均包含外国用户偏好谈论的政治类词汇“soviet”(苏维埃)、“putin”(普京)、“obama”(奥巴马)等,这可能与国外政治领域信息公开程度有关。

表4 英文数据集关键特征

Table 4 Important Features of the Twitter Dataset

模型	谣言 Top 15	非谣言 Top 15
TF-IDF+SVM	putin, bi, soviet, protesters, amazon, potentially, soldier, atheist, town, worker, blasts, fatally, brown, obama, officers	boxed, which, snow, cutting, corporal, tell, building, fallen, seen, hailed, ottawa, appearance, mar, walker, killing
BERT	hailed, walks, google, asshole, tesla, discover, millions, opposes, for, false, locked, cannot, informed, starbucks, whether	soviet, protesters, amazon, potentially, blasts, penis, taliban, yourselves, boycott, employees, banks, already, disney, hacker, boo

然而,除了政治因素,机器学习分类器和深度学习分类器在捕捉特征时,并没有很好地识别出其他具有泛化性的特征,且捕捉到的大部分关键词语在谣言与非谣言均有分布。在谣言和非谣言的关键特征中均有大量中性词,如“amazon”(亚马逊)、“google”(谷歌)、“building”(建筑),其本身并不具有明显的情感倾向或观点。同样地,“taliban”(塔利班)、“Ottawa”(渥太华)、“tesla”(特斯拉)等与数据集关联较大,是与某些特定新闻事件相关的词,不适合作为普遍的谣言关键特征。

此外,在非谣言的特征词汇中,出现了诸如“which”“for”和“cannot”等常用词汇。这些词汇在英文语言中的使用非常普遍,并不具有明确的指向性,因此,其很难为非谣言的识别提供有力证据。一些副词(如“potentially”和“already”等)具有模糊性,可以在各种语境中使用,未必与谣言或非谣言构成直接联系。

从这些关键词中可以发现,模型并没有较好地区分谣言和非谣言的词汇特征。在这些关键词中,大多数是随机的、意义指向不明的,这使得其很难作为泛化的谣言或非谣言标志。

综上所述,在模型的全局解释上,机器学习和深度学习识别的谣言与非谣言特征具有较大差异,有时甚至出现相反结果。对比中英文社交媒体的谣言特征,可以发现二者几乎完全无共性特征。另外,模型未识别出含有情感倾向的词语,无法通过判断情绪差异区分谣言与非谣言。因此,基于语义特征的谣言检测模型仅拟合了训练样本的特征,无法获得普遍且有意义的自然语言特征。

(2)模型局部解释

进一步地,本文使用SHAP和LIME对具体样本进行个案解释。在中文数据集和英文数据集上分别挑选一条“谣言”样本和“非谣言”样本进行绘图。其中,目标解释模型均为在基准数据集上性能最好的BERT模型。

基于SHAP的微博数据集个案分析如图3所示。其中,图3(a)中的样本为谣言,图3(b)中的样本为非谣言。在谣言样本中,点击“谣言”标签,即“Output 0”,呈现红色的词语表示该词语对“谣言”这个分类具有正向贡献,即“日本”“台风”“意外”等词的出现会增加该样本是谣言的概率;非谣言样本中,点击“非谣言”标签,即“Output 1”,呈现红色的

“哈萨克斯坦”“出现”等词对“非谣言”这个分类具有正向贡献。在现实生活中，图3(a)样本中“台风”“地震”“火山”同时发生的情况较为罕见，怀疑其为“谣言”是有一定依据的，SHAP为这三个词赋予了较高的特征值；然而，图3(b)样本中特征值最大的“哈萨克斯坦”为地名，仅从地名中难以判断该样本是否为谣言。这体现出模型识别的特征具有一定偶然性。

图3　基于SHAP的微博数据集个案分析

Fig.3　A Case Study of Weibo Dataset Based on SHAP

使用LIME对这两条样本进行解释，如图4所示。其中，图4(a)的谣言文本中，绿色代表正值，表示该特征对“谣言”分类具有正向贡献；红色代表负值，表示该特征对“谣言”分类具有负向贡献。由此可见，“台风”“火山”“地震”“意外”等词的出现会对模型将这句话判别为谣言产生正面作用。图4(b)的非谣言文本中，绿色代表正值，表示该特征对“非谣言”分类具有正向贡献，即“哈萨克斯坦”“出现”促使模型将样本判别为非谣言；红色代表负值，表示该特征对“非谣言”分类具有负向贡献。上述解释结果与SHAP的解释结果高度一致，说明模型捕捉的特征较稳定。

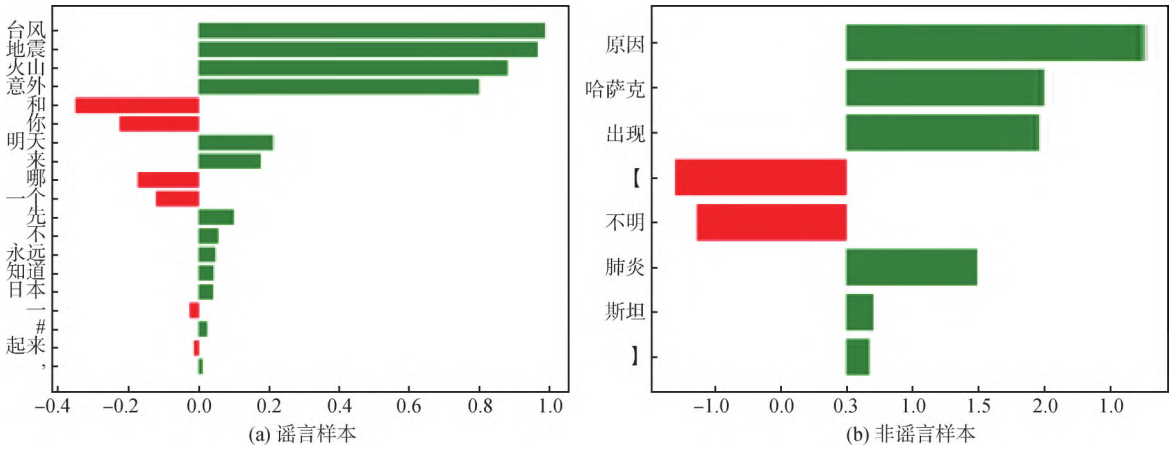

图4　基于LIME的微博数据集个案分析

Fig.4　A Case Study of Weibo Dataset Based on LIME

同样地，基于SHAP的Twitter数据集个案分析如图5所示，对于英文数据集来说，图5(a)为谣言，图5(b)为非谣言。在图5(a)样本中，“animals”(动物)、“eternity”(永恒)、“christ”(基督)等词的出现会增加该样本是谣言的概率。在图5(b)样本中，呈现红色的“melbourne”(墨尔本)、“center”(中心)等词的出现会增加该样本为非谣言的概率。从常识的角度来看，单词如“animals”“eternity”“christ”“melbourne”和“center”等并不能单独确定一条信息是否是谣言，因为这些单词在各种语境下都可能出现。例如，“melbourne”仅是一个城市名，而“animals”只是指动物，这些单词本身并无好坏之分，也不具有谣言或非谣言的语义。

使用LIME进行解释，基于LIME的Twitter数据集个案分析如图6所示，同样地，发现LIME的解释结果和SHAP的解释结果高度一致。本文选取了其他多组案例进行两种解释方法的对比，发现两种方法也都显示了高度相似的结果。这说明SHAP和LIME解释方法计算得出的关键特征是可信的，而通过观察发现的这些关键特征并没有明显的相似性。

综上所述，基于文本语义的谣言检测模型虽然

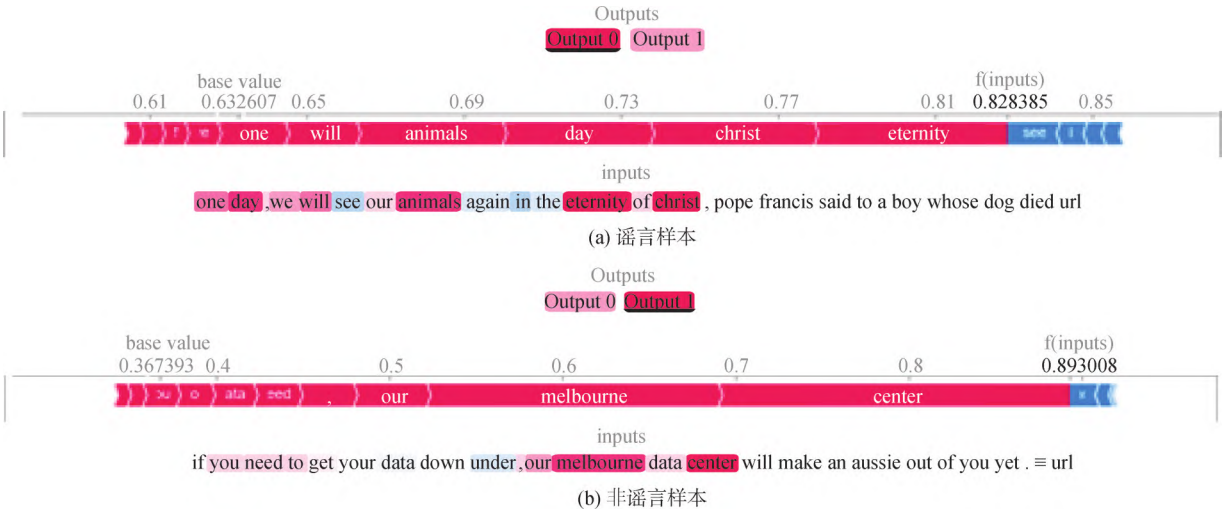

图5 基于SHAP的Twitter数据集个案分析

Fig.5 A Case Study of Twitter Dataset Based on SHAP

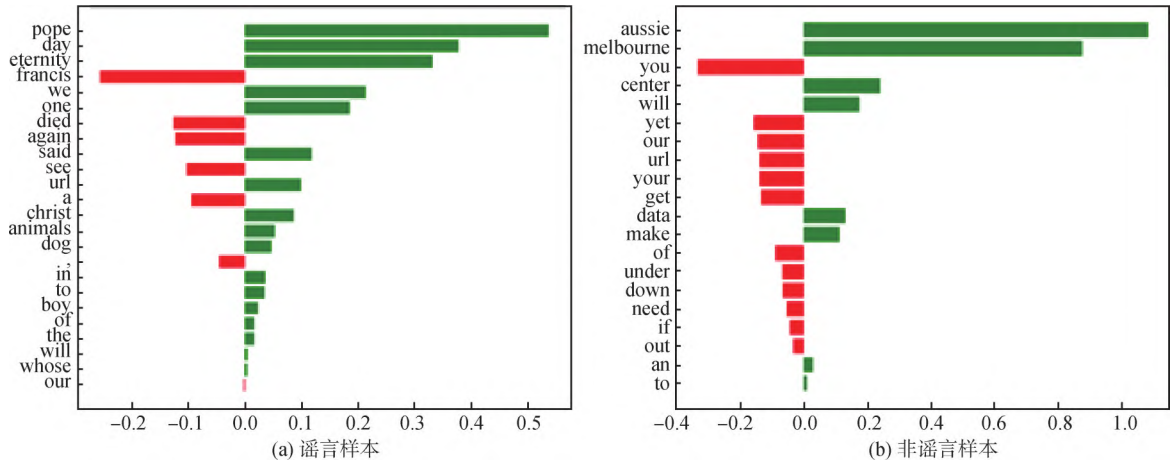

图6 基于LIME的Twitter数据集个案分析

Fig.6 A Case Study of Twitter Dataset Based on LIME

能在封闭的基准数据集上取得较好的分类效果，但其捕捉到的特征更依赖于数据集特有的外在特征，基于内容的深度学习模型无法真正理解谣言的内在特征。

6 结　语

本文以谣言检测领域中常用的英文数据集Twitter15/16和中文数据集MA_weibo为基础，对多种机器学习和深度学习模型进行评估。通过大量实验对比，发现BERT模型在谣言检测方面优于机器学习模型和其他深度学习模型，这说明其具有更强的语义特征提取能力。

为了探究训练出的模型能否真正地对谣言数据进行有效预测，本文采用两种解释性方法——LIME和SHAP，对模型进行解释分析。首先，基于整体数据对模型进行解释，发现模型习得的重要特征词之间存在较大差异，且这些重要特征词的出现没有明显的规律，这表明仅基于文本语义特征的谣言检测模型无法捕捉到普遍的、有意义的特征。进一步，从个案分析发现这些关键特征并没有明显的相似性，因此，无法简单地将这些个案的分类结果归因于模型中习得的特征。

综上所述，虽然基于文本内容的谣言检测模型在分类效果方面表现良好，但其无法真正理解谣言的内在特征，因此，已有基于文本的模型无法真正应对实际场景的挑战。

在理论上，本文可解释地研究了已有基于深度学习的谣言检测模型，分析了深度学习模型在做出预测时主要依赖的特征，为模型研究提供了更深视角，为指导基于更多特征的模型构建提供了理论基础；在实践上，本文的研究结果警示读者不能仅因良好的分类效果而盲从基于内容的深度学习模型，谣言检测模型在多元化场景的泛化性仍有待提高。另外，本文也为解释其他深度学习模型提供了范例。

参考文献：

[1] 李宗建, 程竹汝. 新媒体时代舆论引导的挑战与对策[J]. 上海行政学院学报, 2016, 17(5): 76-85.(Li Zongjian, Cheng Zhuru. Challenges and Countermeasures of Public Opinion Guidance in the New Media Time[J]. The Journal of Shanghai Administration Institute, 2016, 17(5): 76-85.)

[2] 李露琪, 侯丽, 邓胜利. 突发公共卫生事件网络虚假信息传播行为影响因素研究——以新冠疫情期间微博虚假信息为例[J]. 图书情报工作, 2022, 66(9): 4-13. (Li Luqi, Hou Li, Deng Shengli. Research on Influencing Factors of Network Misinformation Transmission Behaviors in Public Health Emergencies: A Case Study of Weibo Misinformation During the COVID-19[J]. Library and Information Service, 2022, 66(9): 4-13.)

[3] 黄学坚, 马廷淮, 王根生. 基于分层语义特征学习模型的微博谣言事件检测[J]. 数据分析与知识发现, 2023, 7(5): 81-91. (Huang Xuejian, Ma Tinghuai, Wang Gensheng. Detecting Weibo Rumors Based on Hierarchical Semantic Feature Learning Model [J]. Data Analysis and Knowledge Discovery, 2023, 7(5): 81-91.)

[4] Castillo C, Mendoza M, Poblete B. Predicting Information Credibility in Time-Sensitive Social Media[J]. Internet Research, 2013, 23(5): 560-588.

[5] Shu K, Zhou X Y, Wang S H, et al. The Role of User Profiles for Fake News Detection[C]//Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. New York: ACM, 2019: 436-439.

[6] 祖坤琳, 赵铭伟, 郭凯, 等. 新浪微博谣言检测研究[J]. 中文信息学报, 2017, 31(3): 198-204.(Zu Kunlin, Zhao Mingwei, Guo Kai, et al. Research on the Detection of Rumor on Sina Weibo[J]. Journal of Chinese Information Processing, 2017, 31(3): 198-204.)

[7] Ma J, Gao W, Wong K F. Detect Rumors in Microblog Posts Using Propagation Structure via Kernel Learning[C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2017: 708-717.

[8] Shen S S, Lee H Y. Neural Attention Models for Sequence Classification: Analysis and Application to Key Term Extraction and Dialogue Act Detection[C]//Proceedings of the 17th Annual Conference of the International Speech Communication Association. San Francisco: ISCA, 2016: 2716-2720.

[9] Bian T, Xiao X, Xu T Y, et al. Rumor Detection on Social Media with Bi-Directional Graph Convolutional Networks[C]// Proceedings of the AAAI Conference on Artificial Intelligence. Menlo Park: AAAI, 2020: 549-556.

[10] Anggrainingsih R, Hassan G M, Datta A. BERT Based Classification System for Detecting Rumours on Twitter[OL]. arXiv Preprint, arXiv:2109.02975.

[11] Zhong W J, Xu J J, Tang D Y, et al. Reasoning over Semantic-Level Graph for Fact Checking[OL]. arXiv Preprint, arXiv: 1909.03745.

[12] Kim B, Doshi-Velez F. Interpretable Machine Learning: The Fuss, the Concrete and the Questions[C]//Proceedings of the 32nd International Conference on Machine Learning. New York:

ACM, 2017: 1-13.

[13] Miller T. Explanation in Artificial Intelligence: Insights from the Social Sciences[J]. Artificial Intelligence, 2019, 267: 1-38.

[14] Ribeiro M T, Singh S, Guestrin C. "Why should I Trust You?" Explaining the Predictions of Any Classifier[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2016: 1135-1144.

[15] Ribeiro M T, Singh S, Guestrin C. Anchors: High-Precision Model-Agnostic Explanations[C]//Proceedings of the 32nd AAAI Conference on Artificial Intelligence. Menlo Park: AAAI, 2018: 1527-1535.

[16] Selvaraju R R, Cogswell M, Das A, et al. Grad-Cam: Visual Explanations from Deep Networks via Gradient-based Localization[C]//Proceedings of the 16th IEEE International Conference on Computer Vision. Piscataway: IEEE, 2017: 618-626.

[17] Bau D, Zhou B L, Khosla A, et al. Network Dissection: Quantifying Interpretability of Deep Visual Representations[C]// Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2017: 3319-3327.

[18] Montavon G, Lapuschkin S, Binder A, et al. Explaining Nonlinear Classification Decisions with Deep Taylor Decomposition[J]. Pattern Recognition, 2017, 65: 211-222.

[19] Lundberg S M, Lee S I. A Unified Approach to Interpreting Model Predictions[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: ACM, 2017: 4768-4777.

[20] Shrikumar A, Greenside P, Kundaje A. Learning Important Features through Propagating Activation Differences[C]// Proceedings of the 34th International Conference on Machine Learning. New York: ACM, 2017: 3145-3153.

[21] 刘天畅, 王雷, 朱庆华. 基于SHAP解释方法的智慧养老服务平台用户流失预测研究[J]. 数据分析与知识发现, 2024, 8(1): 40-54. (Liu Tianchang, Wang Lei, Zhu Qinghua. Predicting User Churn of Smart Home-Based Care Services Based on SHAP Interpretation[J]. Data Analysis and Knowledge Discovery, 2024, 8(1): 40-54.)

[22] 易明, 姚玉佳, 胡敏. 融合XGBoost与SHAP的政务新媒体公共价值共识可解释性模型——以"今日头条"十大市级政务号为例[J]. 图书情报工作, 2022, 66(16): 36-47.(Yi Ming, Yao Yujia, Hu Min. An Interpretable Model for New Government Media Public Value Consensus Integrating XGBoost and SHAP: Taking the Top 10 Municipal Government Accounts of the Jinri Toutiao as an Example[J]. Library and Information Service, 2022, 66(16): 36-47.)

[23] 曾子明, 张瑜, 李婷婷. 多特征融合的突发公共卫生事件潜在谣言传播者识别[J]. 图书情报工作, 2022, 66(13): 80-90.(Zeng Ziming, Zhang Yu, Li Tingting. Detection of Potential Rumor Spreaders in Public Health Emergencies Based on Multi-Feature Fusion[J]. Library and Information Service, 2022, 66(13): 80-90.)

[24] Ayoub J, Yang X J, Zhou F. Combat COVID-19 Infodemic Using Explainable Natural Language Processing Models[J]. Information Processing & Management, 2021, 58(4): 102569.

[25] Mikolov T, Chen K, Corrado G, et al. Efficient Estimation of Word Representations in Vector Space[OL]. arXiv Preprint, arXiv: 1301.3781.

[26] 李慧, 柴亚青. 基于属性特征的评论文本情感极性量化分析[J]. 数据分析与知识发现, 2017, 1(10): 1-11. (Li Hui, Chai Yaqing. Analyzing Sentiment Polarity of Comments Based on Attributes[J]. Data Analysis and Knowledge Discovery, 2017, 1 (10): 1-11.)

[27] Devlin J, Chang M W, Lee K, et al. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding[C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: Association for Computational Linguistics, 2019: 4171-4186.

[28] Rumelhart D E, Hinton G E, Williams R J. Learning Representations by Back-Propagating Errors[J]. Nature, 1986, 323(6088): 533-536.

[29] 苏剑林. 提速不掉点:基于词颗粒度的中文WoBERT[EB/OL]. [2020-9-18] https://kexue. fm/archives/7758(Jianlin Su . WoBERT: Word-based Chinese BERT model.[EB/OL]. [2020-9-18] https://kexue.fm/archives/7758)

[30] Chen Y H. Convolutional Neural Network for Sentence Classification[D]. Waterloo: University of Waterloo, 2015.

[31] Johnson R, Zhang T. Deep Pyramid Convolutional Neural Networks for Text Categorization[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2017: 562-570.

[32] Schuster M, Paliwal K K. Bidirectional Recurrent Neural Networks[J]. IEEE Transactions on Signal Processing, 1997, 45 (11): 2673-2681.

[33] Vosoughi S, Roy D, Aral S. The Spread of True and False News Online[J]. Science, 2018, 359: 1146-1151.

[34] Solovev K, Pröllochs N. Moral Emotions Shape the Virality of COVID-19 Misinformation on Social Media[C]//Proceedings of the ACM Web Conference 2022. New York: ACM, 2022: 3706-3717.

[35] Liu X M, Nourbakhsh A, Li Q Z, et al. Real-Time Rumor Debunking on Twitter[C]//Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. New York: ACM, 2015: 1867-1870.

[36] Ma J, Gao W, Wei Z Y, et al. Detect Rumors Using Time Series of Social Context Information on Microblogging Websites[C]//

Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. New York: ACM, 2015: 1751-1754.

作者贡献声明：

贺国秀：提出研究思路，设计研究方案，实验结果分析，修改论文；
任佳渝、李宗耀、林晨曦：程序开发、数据分析、论文起草；
蔚海燕：实验结果分析，修改论文。

利益冲突声明：

所有作者声明不存在利益冲突关系。

支撑数据：

[1] 贺国秀 . 以可解释工具重探基于深度学习的谣言检测研究数据集 .DOI:10.57760/sciencedb.13348.

收稿日期：2023-07-31
收修改稿日期：2023-09-09

Revisiting Deep Learning-based Rumor Detection Models with Interpretable Tools

He Guoxiu　Ren Jiayu　Li Zongyao　Lin Chenxi　Yu Haiyan
(Faculty of Economics and Management, East China Normal University, Shanghai 200062, China)

Abstract: [**Objective**] This study explores whether content-based deep detection models can identify the semantics of rumors. [**Methods**] First, we use the BERT model to identify the key features of rumors from benchmark datasets in Chinese and English. Then, we utilized two interpretable tools, LIME, based on local surrogate models, and SHAP, based on cooperative game theory, to analyze whether these features can reflect the nature of rumors. [**Results**] The key features calculated by the interpretable tools on different models and datasets showed significant differences, and it is challenging to decide the semantic relationship between the features and rumors. [**Limitations**] The datasets and models examined in this study need to be expanded. [**Conclusion**] Deep learning-based rumor detection models only work with the features of the training set and lack sufficient generalization and interpretability for diverse real-world scenarios.

Keywords: Rumor Detection　Interpretable Machine Learning　Deep Learning　LIME　SHAP