融合证据分析的贝叶斯神经网络虚假信息检测方法

陈君海， 项凤涛， 黎拓新， 罗翔宇
（国防科技大学智能科学学院，湖南 长沙 410073）

摘　要：社交媒体的普及带来了虚假信息扩散速度加快、影响范围扩大等问题，虚假信息的广泛传播不仅会扰乱社会秩序，还可能引发群体性事件，对国家安全和社会稳定构成潜在威胁，研究高效的虚假信息检测工具与技术变得尤为重要。基于此，提出了融合证据分析的贝叶斯神经网络虚假信息检测（evidence-aware Bayesian neural networks for fake news detection，EBNN-FND）方法，该方法借助贝叶斯神经网络框架，对检测模型与数据的不确定性进行量化分析，从而提升预测结果的可靠性。模型设计了文本嵌入模块、特征处理模块、观点-证据交互模块和特征混合模块，能够充分整合信息文本与关联证据信息的特征。在公共数据集上的实验表明，EBNN-FND 模型在虚假信息检测任务中的表现显著优于现有基线模型，具有高效性与稳定性，不仅为虚假信息检测领域提供了新的研究视角，也为解决信息传播过程中的不确定性问题提供了一种可行的技术方案。

关键词：贝叶斯神经网络；变分推理；无偏估计；虚假信息检测
中图分类号：TP391.4
文献标志码：A
doi: 10.11959/j.issn.2096−6652.202533

Evidence-aware Bayesian neural networks for fake news detection

CHEN Junhai, XIANG Fengtao, LI Tuoxin, LUO Xiangyu
College of Intelligence Science and Technology, National University of Defense Technology, Changsha 410073, China

Abstract: The popularity of social media has led to accelerated fake news propagation and expanded influence. The extensive spread of fake news not only disrupts social order but may also trigger mass incidents, posing a potential threat to national security and social stability. Consequently, the development of efficient fake news detection tools and techniques has become increasingly critical. To address this challenge, an evidence-aware Bayesian neural networks for fake news detection (EBNN-FND) method was proposed. This model quantifies uncertainties in both the detection model and the data, thereby improving the reliability of prediction results. The EBNN-FND model consists of four modules: a text embedding module, a feature processing module, a news-evidence interaction module, and a feature fusion module. Thereby, it can effectively integrate the features of news context and related evidence. Experiments on public datasets demonstrate that the EBNN-FND model significantly outperforms existing baseline models in fake news detection tasks, showcasing its efficiency and robustness. It not only provides a new research perspective for the field of rumor detection but also offers a viable technical solution to address uncertainty issues in information dissemination.

Key words: Bayesian neural networks, variational inference, unbiased estimation, fake news detection

0　引言

随着网络通信技术的不断发展，数字化社交媒体已成为公众获取信息的重要渠道，社交平台用户数量的快速增长以及社交平台的不断扩张大幅提高了信息传播的效率，扩大了传播范围。与此同时，

收稿日期：2025−04−07；**修回日期：**2025−07−17
通信作者：项凤涛，xiangfengtao@nudt.edu.cn
基金项目：国家自然科学基金项目（No.62473371）
Foundation Item: The National Natural Science Foundation of China (No.62473371)

虚假信息传播引发的危害也日益显现。虚假信息作为一种未经验证且通常具有欺骗性质的内容，其可能涵盖捏造的事实、夸大的描述以及具有欺骗性的情境等。如2023年6月，一则关于“幼师体罚儿童并强迫喂食药物”的虚假信息在互联网上迅速传播，由于内容与儿童安全紧密相关，该虚假信息迅速在家长群体中扩散，并引发了不小的恐慌。2024年9月，一则关于某超市即将停业的不实消息被广泛传播，这一不实新闻引起了消费者对该超市商品的抢购潮，突发的抢购行为给超市的运营造成了巨大压力，还有一些不法分子在超市内擅自拆封并使用未付款商品，造成现场秩序严重混乱，严重干扰了超市的正常运营，并带来了较大的经济损失。这两个案例清晰地表明，若不加以控制，虚假信息的无序传播将对社会秩序构成严重威胁，并可能在公众中引发广泛的恐慌情绪。因此，采取有效措施遏制虚假信息的传播对于维护社会和谐与稳定至关重要。

人工的虚假信息检测方法大多通过综合评估信息内容、上下文信息及信息源等的可信度来验证信息的真实性。尽管这些方法取得了一定的成效，但仍存在若干难题：首先，虚假信息的定义模糊，缺乏一个简洁且被广泛认可的定义，这阻碍了鉴别标准的统一；其次，虚假信息的传播模式日趋复杂，使得真假信息的辨识愈发困难；最后，尽管专业审核人员能够进行精确的虚假信息检测，但由于专家数量有限，这种方法难以广泛推广。因此，迫切需要开发出既可靠又高效的相关工具和技术。近些年来，越来越多的深度学习方法被应用到虚假信息检测当中，但由于检测模型泛化性能不佳以及预测结果可信度较低，虚假信息检测技术的准确率难以得到进一步提升，并且这一问题在未知测试集上更明显。为了解决这一问题，本文引入贝叶斯神经网络（Bayesian neural network，BNN）以提升预测结果的可信度与准确率。BNN是一种基于贝叶斯方法的模型，它将深度神经网络的参数视为随机变量，并依据训练数据推导后验分布，从而生成预测分布。同时，BNN的输出能够量化不确定性，为不确定性评估提供了有效的量化手段[1]。

在深度学习领域，神经网络往往无法有效评估不确定性，因为其参数计算主要依赖于固定点估计方法[2]。贝叶斯方法作为一种基于先验概率推导后验概率的推理框架，通过迭代过程利用现有数据逼近目标预测结果，其坚实的数学基础赋予了预测结果更高的说服力与可信性。贝叶斯方法与神经网络结合后，具有以下显著优势[3]：（1）BNN根植于坚实的数理基础，确保了其构造过程在数学上的可验证性；（2）BNN通过概率分布描述模型参数与偏差，从而提供模型输出的解释能力，避免了重复测试与交叉验证，提升了模型的可解释性；（3）BNN能够量化预测的不确定性，并整合专家先验知识以优化学习过程，同时，BNN支持使用不同形状的概率分布作为先验与似然分布，增强了模型的灵活性。

综上所述，本文主要贡献如下。

（1）提出了一种新颖的虚假信息检测模型，即融合证据分析的贝叶斯神经网络虚假信息检测（evidence-aware Bayesian neural networks for fake news detection，EBNN-FND）方法。该模型通过BNN框架量化检测模型和数据中包含的不确定性，提升了模型对文本信息的特征提取能力，不仅克服了传统方法的局限性，还提升了预测结果的可靠性。

（2）引入外源证据信息，通过深入挖掘多元输入数据中虚假信息与相关证据之间的关联，增强了模型的预测准确率和稳定性。

（3）实验结果表明，在公共数据集上的测试中，该模型在虚假信息检测任务中表现出优越的性能。

1　相关工作

1.1　虚假信息检测

近年来，虚假信息在各类社交媒体平台上的泛滥，使得相关研究的数量也日益增多[4-6]。传播的信息通常包含以下3个核心要素。

（1）观点内容：通常是多模态信息，包括文本、声音、视频等多种格式[7-9]。

（2）发布者与评论者：这两类主体构成了社交信息传播网络的基础，分别对应传播网络的起点和节点。在社交媒体上，最早发布观点的称为发布者，而通过转发、评论等方式参与互动的用户被称为评论者。平台的历史记录功能可以追溯发布者和评论者的过往活动，从而评估其发言的可信度。

（3）外部知识：外部知识包括内容发布的时间、背景以及广泛的事实库。

虚假信息检测方法还可根据是否整合外源信息分为以下两类。

（1）基于模式的虚假信息检测方法[10-12]：该方法仅依赖观点内容本身，无须外源信息补充，可通过提取输入信息的特征来完成虚假信息检测。

（2）基于证据的虚假信息检测方法[13-15]：相比之下，该方法不仅利用观点本身，还需要结合外部信息（如证据）进行辅助检测。具体来说，它通过检索与观点关键词高度相关的条目，并分析这些内容之间的关联性，辅助判断信息的真实性。

最初的虚假信息检测主要基于机器学习方法，如支持向量机、朴素贝叶斯和随机森林等，这些方法通常需要人工提取特征，从信息中提前提取基于内容的、基于用户的以及基于交互的特征来训练分类器。许多研究通过组合多个分类器来提高检测效果，包括决策树[16-17]、贝叶斯网络[16]、随机树[17]、逻辑回归[18]和支持向量机[16-17]等。随着深度学习方法的兴起，基于深度神经网络的虚假信息检测方法逐渐得到应用，并显著提升了检测性能。卷积神经网络（convolutional neural network，CNN）、循环神经网络（recurrent neural network，RNN）以及长短期记忆网络（long short-term memory，LSTM）等方法被广泛应用于文本内容检测与传播结构检测，这些方法不需要额外的人工特征提取过程。Ma 等[19]首次将 RNN 应用于虚假信息检测领域，并将事件建模为一个持续的信息流，该信息流由初始帖子与相关帖子的集合构成。Ruchansky 等[20]提出了一个结合文本内容、用户行为和发帖人行为三方面特征的混合模型。Jin 等[21]融合了文本信息、视觉和社交数据，提出了一种基于多模态融合的模型，该模型利用 LSTM 整合了帖子中的文本内容和社交背景，再将该联合表征与通过预训练的 VGG-19（visual geometry group-19，VGG-19）网络提取的视觉特征进行整合。Kumar 等[22]发现，结合 CNN、双向 LSTM 与注意力机制的集成网络模型在虚假信息检测任务中表现出卓越的性能。预训练语言模型的发展对自然语言处理和深度学习领域产生了深远影响，这些模型通过在海量文本数据上的预训练，积累了丰富的语言知识和深层语义信息，在包括虚假信息检测在内的多种任务中展现出卓越的性能。由于这些模型在广泛的语料库上进行了训练，它们能够捕捉到复杂的语言模式和语义特征，从而针对虚假信息检测任务实现有效的微调。微调过程优化了模型参数，显著提升了虚假信息检测的准确性和检测效率。

基于上述方法，本文提出了 EBNN-FND。该方法通过 BNN 提取文本的深度特征，并结合外源证据信息，提供更可靠且更合理的虚假信息鉴别过程。

1.2 BNN

尽管深度学习已在诸多领域取得显著成效，但仍存在局限性。当训练数据稀疏时，模型容易发生过拟合，并对自身结果表现出“过度自信”，该现象通常表现为模型无法准确识别可能犯错的情境，以及缺乏充分考虑与数据和模型相关不确定性的能力[23]。在高风险决策领域，不确定性对于优化深度学习模型至关重要，为应对这一挑战，BNN 凭借其在处理模型不确定性方面的优势，近年来获得广泛关注，并成为研究的热点。

BNN 利用概率论技术从数据集中提取知识，通过结合专家知识的先验信息和数据的似然值来计算后验概率，在处理模型不确定性的同时推断模型的未知参数。与传统神经网络相比，BNN 的优势在于能够推断模型参数的先验分布和后验分布。通常，参数的先验分布采用高斯分布或其他连续分布，通过贝叶斯推理，逐步迭代计算后验分布，使其逐渐贴合真实分布，从而反映数据对参数权重的约束。这种用概率分布表示模型参数的方法，天然提供了一种计算不确定性的途径[24]。一般来说，BNN 可以处理两种不确定性[25]：（1）偶然不确定性，源于数据本身的性质和噪声，无法通过模型训练减少；（2）认知不确定性，源于模型结构和参数的差异，可通过贝叶斯方法有效控制，同时避免过拟合。

BNN 的后验分布通常难以精确计算，因为后验分布往往具有复杂的高维结构，直接计算非常困难。因此，通常采用采样或近似方法来拟合后验分布。其中，采样的方法有马尔可夫链蒙特卡罗（Markov chain Monte Carlo，MCMC）采样，常见的近似拟合方法有变分推理（variational inference，VI）与蒙特卡罗（Monte Carlo，MC）Dropout 方法。从计算效率角度出发，VI 通常比 MCMC 采样更高效，更适配本文的虚假信息检测问题；从模型灵活性出发，VI 可以灵活选择近似分布族，调整精度和计算成本；从模型优化出发，VI 通过优化过程和参数共享等手段有效控制模型复杂度；从模型适配性出发，VI 利用反向传播进行优化，易于集成到深度学习框架，便于端到端训练，具有较好

适配性。综上所述，本文选择VI作为BNN后验分布的近似计算方法。

2　方法

针对虚假信息检测问题，各种大语言模型已经取得了不错的表现，但同时也伴随着庞大的参数量，对设备算力有较高的要求，这使得模型在实际应用中无法得到有效推广，无法在各种边缘设备上进行有效部署。因此，本文尝试提出参数量级相对较小的虚假信息检测方法，构建了EBNN-FND方法来对信息进行真假鉴别。本节详细描述了EBNN-FND模型，并对各模块的结构和功能进行了阐述。

2.1　方法概述

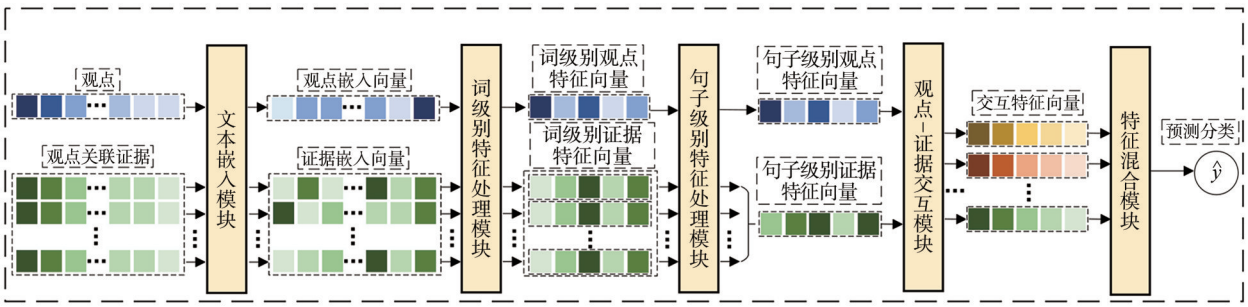
图1展示了EBNN-FND的整体框架结构。该模型由以下4个主要模块构成：（1）文本嵌入模块，该模块依据不同需求，采用基于Transformer的或基于文本卷积神经网络（TextCNN）的文本编码层将输入文本转换为向量形式；（2）特征处理模块，其包含词级别和句子级别两个处理单元，并创新地引入了贝叶斯过程对多头注意力层和前馈网络层进行贝叶斯化处理，从而提取观点和相关证据的嵌入向量中的深层特征；（3）观点-证据交互模块，该模块通过融合去偏方法，实现句子级别观点与证据特征向量的交互，生成具有判别性的交互特征；（4）特征混合模块，针对多维度特征处理的差异性，该模块通过集成多种交互特征实现优势互补，最终基于混合特征向量完成分类预测。上述模块中，特征处理模块是本文的研究重点，其余模块的设计旨在优化模型对虚假信息检测任务的适应性，从而提升整体检测性能。

2.2　文本嵌入模块

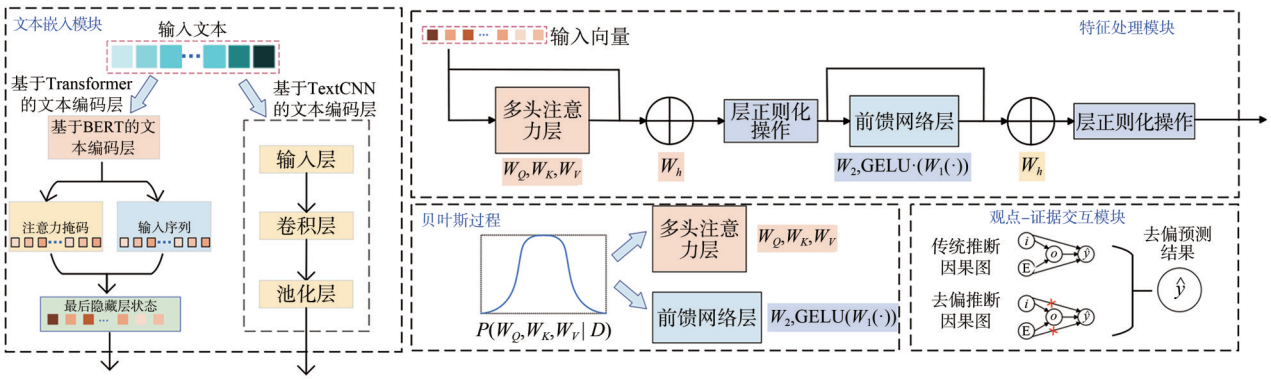
文本嵌入模块将输入文本序列 $\boldsymbol{s}_l=(w_1,w_2,\cdots,w_i,\cdots,w_n)$ 转换为目标嵌入向量表征 $\boldsymbol{m}=(x_1,x_2,\cdots,x_i,\cdots,x_n)$，其核心功能是将高维的文本数据映射到低维的向量空间，这一过程不仅保留了原始文本的重要语义信息，同时也便于计算机进行高效的处理和分析。本文选择了如下两种策略构建文本嵌入模块。

2.2.1　基于Transformer的文本编码层

在虚假信息检测任务中，预训练语言模型有助于深入理解和表达文本的语义内容。来自Transformer的双向编码器表征（bidirectional encoder representation from Transformer，BERT）模型[26]能够



（a）EBNN-FND模型整体框架



（b）EBNN-FND各子模块框架

图1　EBNN-FND的整体框架结构

为虚假信息文本生成细致的字符级向量表示，使用预训练的BERT模型对输入文本内容进行字符级别的嵌入操作，将文本特征有效映射到向量空间。为了在虚假信息检测领域应用，本文将BERT模型最后一层输出的隐藏状态向量作为文本的嵌入向量。在后续的Politifact和Snopes数据集实验及LIAR数据集的部分实验中，采用该方法作为模型的文本嵌入模块。

2.2.2 基于TextCNN的文本嵌入层

尽管BERT模型在文本处理任务中表现优异，但其具有庞大的预训练参数量，运行时间较长。相比之下，基于TextCNN的方法虽然不具备BERT的泛化能力，但在某些数据集上同样表现良好，且显著减少了运行时间。TextCNN[27]使用不同尺寸的卷积核提取句子中的关键信息，有效捕捉了局部相关性。在后续LIAR数据集的部分实验中，采用该方法作为文本嵌入模块，并将第二层卷积层输出的特征图作为文本的嵌入向量。

2.3 特征处理模块

特征处理模块旨在挖掘特征向量中的深层次信息和规律，以提升模型的预测准确性和泛化能力。本文的特征处理模块分为词级别与句子级别两个部分，二者的主要区别在于处理的数据粒度和提取的特征类型。词级别特征处理模块侧重于捕捉词汇的内在属性和语义信息，为句子级别处理提供基础特征；句子级别特征处理模块则关注识别词汇的组合模式、语义关系及其在更大语境中的功能与意义。两个模块相辅相成，通过多层次的特征处理，可以更全面地理解和分析文本数据。

Xue等[28]在其研究中发现，仅对浅层的Transformer解码器进行贝叶斯化（Bayes former，BF）处理能在语音识别任务中提高准确率。基于此发现，本文构建的特征处理模块包含6个Transformer解码器，其中前两个经过贝叶斯处理，包括对前馈网络（fully forward connected network，FFN）层贝叶斯化与多头注意力（multi head attention，MHA）层贝叶斯化。本节将详细探讨贝叶斯化如何实现特征提取目标。

尽管Transformer模型在众多任务上表现出色，但其固定点估计方法未能考虑模型相关的不确定性，导致在训练数据不足时容易出现过拟合和泛化能力不足的问题。为此，本文采用BNN将模型参数ω视为后验概率分布$p(\omega|D)$，以进行预测任务：

$$p(\hat{y})=\int p(\hat{y}|\omega)p(\omega|D)\mathrm{d}\omega \tag{1}$$

其中，$\hat{y}$代表预测标签，D为训练数据集，$p(\omega|D)$为模型参数的后验分布。

鉴于后验分布的复杂性，精确求解极具挑战性，因此本文使用VI来近似获得后验分布，并通过优化变分参数θ为后续的概率推断奠定基础：

$$\theta^{\mathrm{opt}}=\underset{\theta}{\operatorname{argmin}}\,\mathrm{KL}\left[q(\omega|\theta)\middle\|P(\omega|D)\right] \tag{2}$$

$$\mathrm{KL}\left[q(\omega|\theta)\middle\|P(\omega|D)\right]=\int q(\omega|\theta)\log\frac{q(\omega|\theta)}{P(\omega|D)}\mathrm{d}\omega=\log P(D)+\mathrm{KL}\left[q(\omega|\theta)\middle\|P(\omega)\right]-\int q(\omega|\theta)\log P(D|\omega)\mathrm{d}\omega \tag{3}$$

其中，θ^{opt}为最佳的变分参数，$q(\omega|\theta)$为变分后验分布，$P(\omega|D)$为真实后验分布。KL散度用于衡量两个分布的差异，通过最小化$q(\omega|\theta)$与$P(\omega|D)$的KL散度来获得θ^{opt}。$P(\omega)$为模型参数的先验，$P(D|\omega)$为似然值，$P(D)$为与模型无关的常量，因此式（2）等价于最小化目标函数$\mathcal{F}(D,\theta)$：

$$\mathcal{F}(D,\theta)=\mathrm{KL}\left[q(\omega|\theta)\middle\|P(\omega)\right]-\int q(\omega|\theta)\log P(D|\omega)\mathrm{d}\omega \tag{4}$$

该目标函数也被称为证据下界[29]，包含复杂度成本与似然成本两项。本文为了最小化目标函数$\mathcal{F}(D,\theta)$，结合蒙特卡洛采样和反向传播来学习模型参数分布[30]，并通过采样降低计算成本：

$$\mathcal{F}(D,\theta)\approx\frac{1}{K}\sum_{i=1}^{K}\log q\left(\omega^{(i)}|\theta\right)-\log P\left(\omega^{(i)}\right)-\log P\left(D\middle|\omega^{(i)}\right) \tag{5}$$

其中，$\omega^{(i)}$表示从变分后验$q(\omega|\theta)$中随机采样的第i个样本，K为采样总数。在先验分布的设置上，本文参考Ma等[2]和Xue等[31]的工作，将$q(\omega|\theta)$和$P(\omega)$设为对角高斯分布：

$$\begin{cases}q(\omega|\theta)=\mathcal{N}(\Theta;\mu,\sigma)\\P(\omega)=\mathcal{N}(\Theta;\mu^{r},\sigma^{r})\end{cases} \tag{6}$$

其中，$\mathcal{N}(\cdot)$表示高斯分布，μ和σ分别表示对应高斯分布的均值和标准差。为避免训练的不稳定性，本文对μ和σ采用重参数[32]技巧：

$$\Theta=\mu+\sigma\odot\varepsilon_k,\quad \varepsilon_k\sim\mathcal{N}(0,1) \tag{7}$$

其中，ε_k是一个服从标准正态分布的随机变量。基于上述假设，式（4）中的KL散度可拆为参数的KL散度$\mathrm{KL}_{\mathrm{weights}}$和偏置的KL散度$\mathrm{KL}_{\mathrm{bias}}$，因此总的KL散度为两部分之和：

$$\mathrm{KL}_{\mathrm{weights}}=\frac{1}{2}\left(E\left(\mu_{\mathrm{weights}}^2-2\log\sigma_{\mathrm{weights}}+\sigma_{\mathrm{weights}}^2-1\right)\right) \tag{8}$$

$$\mathrm{KL}_{\mathrm{bias}}=\frac{1}{2}\left(E\left(\mu_{\mathrm{bias}}^2-2\log\sigma_{\mathrm{bias}}+\sigma_{\mathrm{bias}}^2-1\right)\right) \tag{9}$$

$$\mathrm{KL}=\mathrm{KL}_{\mathrm{weights}}+\mathrm{KL}_{\mathrm{bias}} \tag{10}$$

其中，μ_{weights}和$\sigma_{\mathrm{weights}}$分别表示与参数相关的高斯分布的均值和标准差，$\mu_{\mathrm{bias}}$和$\sigma_{\mathrm{bias}}$分别表示与偏置相关的高斯分布的均值和标准差。均值$\mu$和标准差$\sigma$的梯度计算可使用标准反向传播算法：

$$\begin{cases}\mu=\mu-\eta\times\dfrac{\partial L}{\partial\mu}\\ \sigma=\sigma-\eta\times\dfrac{\partial L}{\partial\sigma}\end{cases} \tag{11}$$

其中，L为变分推理计算过程中得到的损失函数，$\eta\sim\mathcal{N}(0,1)$。

基于上述原理，本文对Transformer层进行贝叶斯化处理。鉴于特征处理模块由6个Transformer解码器构成，且Xue等[31]的实验表明，仅对浅层Transformer解码器进行贝叶斯化效果更佳。本文对前两个解码器进行贝叶斯化，包括FFN层贝叶斯化与MHA层贝叶斯化两个部分，BNN层结构示意如图2所示。

2.4　观点-证据交互模块

观点-证据交互模块旨在融合观点内容与证据碎片，分析其中的因果关联以识别虚假信息中的可疑部分，并提取更可信的交互特征向量，从而有效提高虚假信息检测的准确性和泛化能力。以下是两种实现观点和证据特征向量进行交互的方法。

2.4.1　特征融合交互方法

本方法采用了一种直接且有效的策略来整合观点与相关证据的特征，通过横向拼接信息特征与证据特征，实现特征整合：

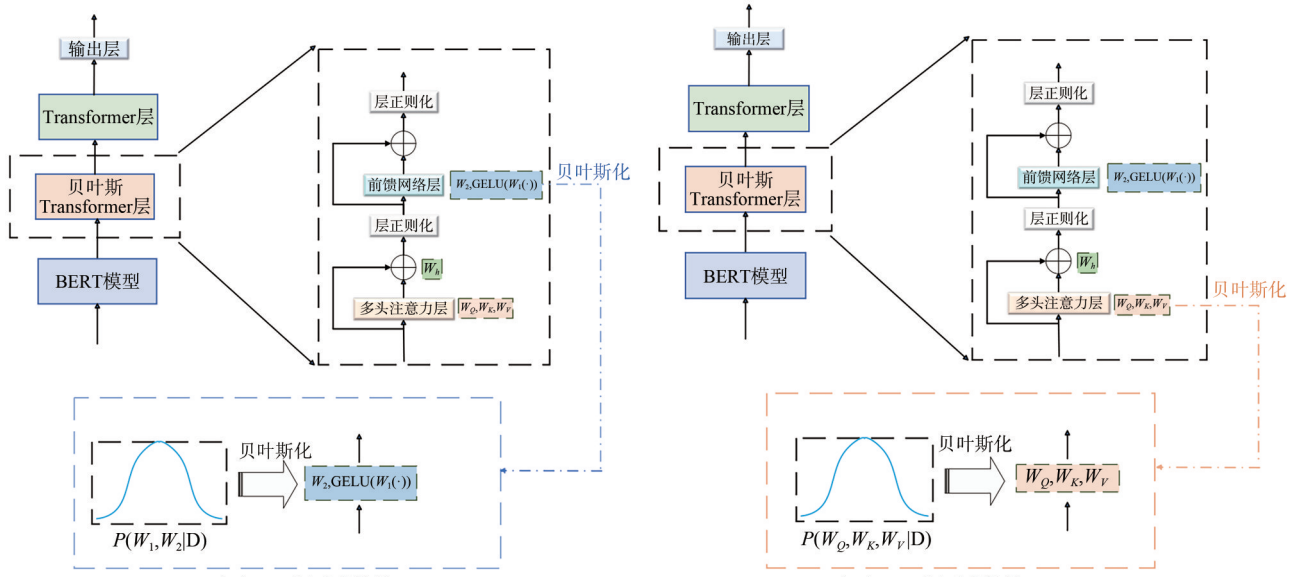
$$\boldsymbol{F}_{\mathrm{concat}}=\left[\boldsymbol{F}_{\mathrm{News}};\boldsymbol{F}_{\mathrm{evidence}}\right] \tag{12}$$

其中，$\boldsymbol{F}_{\mathrm{concat}}$表示拼接后的信息特征，$\boldsymbol{F}_{\mathrm{News}}$和$\boldsymbol{F}_{\mathrm{evidence}}$分别对应观点和证据的特征向量。

2.4.2　去偏交互方法

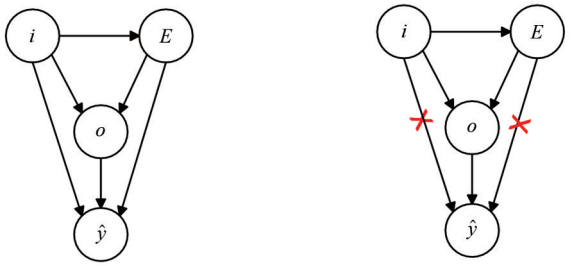
本文参考Wu等[15]提出的因果去偏方法，引入了如下的去偏方法提升模型检测性能。假设观点i与对应的证据集$E=\{e_1,e_2,\cdots\}$通过交互后得到的交互特征为o，则希望仅通过o对预测结果作出推断。本文通过因果图对虚假信息检测模型的特定偏差进行定量分析，并设计对应的除偏策略，基于证据的虚假信息检测因果图如图3所示。

基于证据的虚假信息检测模型通过观点与证据特征之间的交互处理，即$i\to o$与$E\to o$，利用交互特征o预测情报的真伪，即$o\to\hat{y}$。因为数据是人为筛选的，所以难免带有偏差，由于数据偏差，模型可能错误学习观点内容与预测标签之间的虚假相



（a）FFN层示意结构　　（b）MHA层示意结构

图2　BNN层结构示意



（a）传统推断模型的因果图 （b）移除了虚假关联后的因果图

图3 基于证据的虚假信息检测因果图

关性，即$i\to\hat{y}$；考虑证据是根据观点进行检索的，观点与证据内容高度相关，可能存在相似主题或关键词，即$i\to E$；这些相似的主题和关键词可能导致证据内容与预测标签之间也产生虚假相关性$E\to\hat{y}$。在上述虚假相关性的影响下，通常会导致$p_{\text{train}}(\hat{y}|i)\neq p_{\text{test}}(\hat{y}|i)$以及$p_{\text{train}}(\hat{y}|E)\neq p_{\text{test}}(\hat{y}|E)$，其中，$p_{\text{train}}(\hat{y}|i)$和$p_{\text{train}}(\hat{y}|E)$分别表示训练阶段基于观点和证据的预测结果，$p_{\text{test}}(\hat{y}|i)$和$p_{\text{test}}(\hat{y}|E)$分别表示测试阶段基于观点和证据的预测结果。降低虚假关联$i\to\hat{y}$和$E\to\hat{y}$的影响，有助于模型得出更准确的预测结果。

（1）有偏训练

本模块在有偏数据集上进行训练，将观点和证据作为模型输入，并传递至各层。最终，将观点i、证据E与交互特征o输入融合模块进行信息聚合，输出每个类别的预测分布：

$$y_{i,E}=h(i,E) \tag{13}$$

其中，$y_{i,E}$为输出的概率分布，$h(\cdot)$为混合函数。模型能够估计观点和证据对预测结果的总因果效应，为优化模型参数ω，采用交叉熵损失作为训练目标，并将其最小化：

$$L_{\Theta}(y,y_{i,E})=-\sum_{j}y_j\log y_{j,E} \tag{14}$$

其中，y是真实标签集，y_j是第j条样本的真实标签，$y_{j,E}$是第j条样本的预测结果。

（2）反事实训练

反事实推理是在假设某些变量不起作用的条件下推断可能的结果，以区分剩余变量对结果的贡献。在虚假信息检测中，相应的反事实推论要求模型在仅有观点的情况下预测真实性，需进行因果干预以消除交互特征o的传入链接，并指定受阻证据的值。为表示被阻断的证据信息，本文仿照Wu等[15]的方法，对训练集上的证据向量进行平均处理，得到平均证据特征E^*，并用其代替原始证据特征E，与i一起输入融合模块，得到平均交互特征o^*。假设平均证据特征E^*保留有主题或关键词，但不具备证据关联的因果检索能力，模型的反事实输出如下：

$$y_{i,E^*}=h(i,E^*) \tag{15}$$

在此干预下，原有因果关系$i\to o$与$E\to o$被干扰，阻断了证据信号的传递。

（3）去偏操作

为降低模型输出的偏差，模型需要同时计算有偏训练结果$y_{i,E}$和反事实预测结果y_{i,E^*}，并通过逐元素相减的方法获得相对无偏的输出$\hat{y}$，表达如下：

$$\hat{y}=y_{i,E}-\lambda\cdot y_{i,E^*} \tag{16}$$

其中，λ是为了避免减法不足或过多而引入的超参数。

2.5 特征混合模块

在进行虚假信息检测任务时，通常会涉及多方面的复杂因素，如本文用于实验的LIAR数据集，涵盖了观点真实标签、发言人信息、工作职称、所属阵营、观点内容、关联证据以及发言人过往发言可信度等多方面信息，因此需要采用更合适的方法将上述多维度的数据特征进行整合。基于该目的，本文引入了特征混合模块以进一步整合多维度数据特征，提升模型的鲁棒性、准确性、泛化能力以及信息理解能力等。Xu等[33]在其研究中证明了模糊逻辑在处理多元信息时的有效性，因此本模块利用模糊逻辑将上下文数据的输出特征映射到模糊隶属度空间中，以更有效地捕捉底层模式特征并降低噪声影响。特征混合模块中的高斯隶属度函数用参数σ与m来控制函数的形状，则每一个隶属函数$\mu_{\text{G}}(\cdot)$可以表示如下：

$$\mu_{\text{G}}(x;m,\sigma)=\mathrm{e}^{-\frac{1}{2}\left(\frac{(x-m)}{\sigma}\right)^2} \tag{17}$$

其中，x为$\mu_{\text{G}}(\cdot)$的自变量，m控制着曲线的中心，σ控制着曲线的发散程度。为了得到实例i对类别j的隶属度值$\mu_{i,j}$，应求得所有特征的模糊隶属度值的平均值$\mu_{p,q}$如下：

$$\mu_{p,q}=\frac{1}{n}\sum_{k=1}^{n}\mu_{\text{G}}(x_{p,k};m_{q,k},\sigma_{q,k}) \tag{18}$$

其中，n为特征维度，$x_{p,k}$为第p个实例的第k个特征。若特征k属于类别q，则其隶属度值对应的隶

属度函数μ_{G}的参数由$m_{q,k}$和$\sigma_{q,k}$构成。

3　实验结果与分析

本节详细介绍了实验中使用的数据集、基线模型以及实验的参数设置，对结果进行了详细的分析，并进行了消融研究，以说明本文所提EBNN-FND模型的有效性。

3.1　实验设置

3.1.1　实验数据集

本文采用了3个现实世界的真实数据集：Hansen整理的Politifact和Snopes数据集[34]以及Wang整理的LIAR数据集[35]。Politifact数据集由专业记者和事实核查员对政治人物的公开声明进行审查，确保数据的准确性和可靠性。Snopes数据集基于网络虚假信息的广泛收集与分类构建，并通过文本清洗、去重和标注等预处理步骤，以确保记录的准确性和一致性。LIAR数据集主要源自Politifact网站，涵盖了观点真实标签、发言人信息、工作职称、所属阵营、观点内容、关联证据以及发言人过往发言可信度等详细信息。Snopes、Politifact数据集统计见表1，LIAR数据集统计见表2。

表1　Snopes、Politifact数据集统计

数据集	观点总数	虚假信息数	真实信息数
Snopes	5 069	3 642	1 427
Politifact	13 581	6 342	7 239
Snopes Hard	357	252	105
Politifact Hard	942	508	434

表2　LIAR数据集统计

参数	取值
数据集	Liar
总数	12 836
极度虚假	1 050
基本虚假	2 525
半虚假	2 641
大致真实	2 623
几乎真实	2 108
真实	1 889

在数据预处理阶段，Snopes数据集的标签限于“真实”和“虚假”，直接适用于二分类实验，相比之下，Politifact数据集则包含6种标签，都为统一标签，其中“极度虚假”“基本虚假”和“半虚假”这3个标签统一归类为“虚假”，其他标签则归类为“真实”。此外，为了增加实验难度，还从原始数据集中筛选出BERT模型难以准确分类的样本，创建了两个更具挑战性的子数据集：Snopes Hard和Politifact Hard[15]。与原始数据集相比，这些Hard数据集更具挑战性。对于包含6种分类结果的Liar数据集，本文未做进一步处理。为了在表述过程中不产生歧义，本文将未进行分类的被测文本称为“观点”，含有虚假信息的称为“虚假信息”，反之称为“真实信息”。

本文在Snopes和Politifact数据集上，将训练、测试、验证集之比设为7∶2∶1，在LIAR数据集上，将训练、测试、验证集之比设为8∶1∶1。实验所用证据信息均来自数据集内部，选取标准为与观点最相关的前10条证据。

3.1.2　实验参数设置

本文采用PyTorch深度学习框架（版本2.0.0），并符合一般边缘设备的算力要求，在NVIDIA RTX 4060 Laptop GPU上进行实验。可调参数汇总见表3。

表3　可调参数汇总

可调参数	取值
观点内容长度	60
批输入大小	64
Clip	0.8
输入嵌入维度	800
隐藏层维度	800
注意力头数目	8
训练采样数目	1
测试采样数目	100
优化器	SGD
学习率	0.05

3.1.3　评价指标

本文采用多种评估指标评估模型性能，包括准确率、F1-Macro和F1-Micro分数。在虚假信息检测中，高精确率确保了大多数被预测为“虚假信息”的样本的正确性，高召回率确保了大多数“虚假信息”能被识别。F1分数作为精确率和召回率的调和平均值，提供了一个更均衡的性能度量，在样本分布不均的数据集上，F1-Macro因其综合考虑了每个类别的表现，是更可靠且有效的评估指标。因此，在虚假信息检测任务中，F1-Macro指标能更加全面地反映一个模型的能力。

3.1.4 基线模型

为了更有效地评估EBNN-FND方法的性能，本文将其与以下基线方法进行了对比分析，并在Politifact和Snopes数据集上进行了实验。Snopes和Politifact基线模型对比见表4。

表4 Snopes和Politifact基线模型对比

模型	Snopes		Politifact	
	F1-Macro	F1-Micro	F1-Macro	F1-Micro
TRF	0.550	0.271	0.310	0.304
GLSTM	0.529	0.253	0.288	0.294
BERT	0.599	0.691	0.621	0.621
MAC	0.658	0.681	0.573	0.612
DeClarE	0.594	0.603	0.601	0.629
GET	0.636	0.660	0.618	0.636
EBNN-FND	0.661	0.715	0.666	0.668

（1）基于词频的随机森林（term-frequency based random forest，TRF）[36]：TRF基于拼接的信息文本和相关证据片段，为每个样本构建词频统计，并以基尼不纯度为指标进行训练。

（2）GLSTM（GloVe-based LSTM model）[36]：LSTM利用预训练的GloVe模型对文本进行嵌入，将信息和证据向量映射至同一空间，并采用注意力加权的双向LSTM进行编码。

（3）BERT[26]：BERT通过在大规模文本数据上无监督学习，捕获语言模式和关系。BERT的优势在于其双向Transformer架构能够全面考虑上下文信息，生成准确的词嵌入。

（4）用于事实核查的分层多头注意力网络（hierarchical multihead attentive network for fact-checking，MAC）[15]：MAC是一个综合多级注意力机制的层次化MHA网络。

（5）基于证据感知深度学习的虚假新闻与错误声明检测系统（debunking fake news and false claims using evidence-aware deep learning，DeClarE）[37]：DeClarE是一个端到端的神经网络模型，自动评估自然语言声明的可信性，不需要手工设计的特征或词典。

（6）基于图的语义结构挖掘框架（graph-based semantic structure mining framework，GET）[14]：GET利用图框架探索复杂的语义结构，通过信息与证据图的信息传播，捕获长距离的语义依赖。

在LIAR数据集上，本文提出的EBNN-FND模型与以下基线模型在准确率上进行了比较，其中，多源多类虚假新闻检测（multi-source multi-class fake news detection，MMFD）是由Karimi等[38]提出的改进检测方法，LIAR基线模型对比见表5。

表5 LIAR基线模型对比

参考文献	基线模型	准确率
文献[39]	GNN	0.268
文献[35]	CNN+LSTM	0.274
文献[38]	MMFD	0.348
文献[40]	BERT	0.406
文献[41]	LSTM	0.399
	LSTM+Attention	0.415
文献[42]	CNN+LSTM	0.437
本文方法	EBNN-FND w/o FZ	0.448
	EBNN-FND	0.460

3.2 实验分析

3.2.1 整体性能

表4和表5分别展示了EBNN-FND模型与基线模型在Snopes、Polotifact数据集上的F1-Macro和F1-Micro对比结果，以及在LIAR数据集上的准确率对比结果。总的来说，在Snopes和Politifact数据集上，EBNN-FND模型在所有评估指标上均优于其他模型；在LIAR数据集上，EBNN-FND模型的准确率也优于其他基线模型。实验结果表明，结合Bayesformer和观点-证据交互模块的EBNN-FND模型提升了虚假信息检测性能，最佳结果以粗体下划线标出。具体而言，在Snopes数据集上，EBNN-FND模型的F1-Macro指标为66.1%，较其他基线模型高出0.3%~13%；F1-Micro指标为71.5%，较其他基线模型高出5%~44%。在Politifact数据集上，F1-Macro指标为66.6%，比其他基线模型高出5%~35%；F1-Micro指标为66.8%，比其他基线模型高出3%~36%。在LIAR数据集上，准确率指标为46%，较其他基线模型高出2%~20%。

此外，本文还进行了特征混合模块有效性的论证测试，即EBNN-FND w/o FZ模型。该测试用例与EBNN-FND相比，缺少了特征混合模块，直接将各输出横向拼接后输入了全连接层以获得最终的预测结果。可以发现，EBNN-FND w/o FZ的准确率虽然较EBNN-FND有些许下滑，但相较于其他基线模型均有所提升。该实验用例的引入既证明了EBNN-FND框架中特征提取模块的有效性，也证明了特征混合模块的有效性。

以上结果表明，LSTM、TextCNN与Transformer等技术有助于模型捕捉长距离依赖关系，并通过注意力机制识别输入序列中与任务相关的部分，进而提升性能。

3.2.2　消融实验

为证明特征处理模块的有效性，本文提出EBNN-FND模型的3种变体，分别为FFN层贝叶斯化、MHA层贝叶斯化和无贝叶斯化方法，无贝叶斯化方法未对任何层进行贝叶斯化。在无额外证据输入时，EBNN-FND模型在数据集上的F1-Macro和F1-Micro指标对比如图4所示。

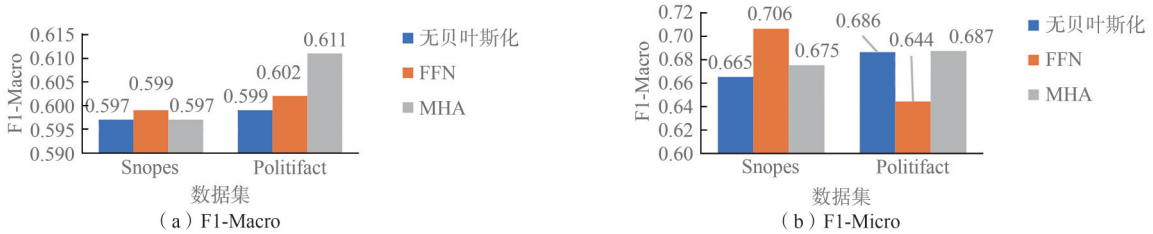
由图4可知，在Politifact数据集上，MHA变体在F1-Macro和F1-Micro指标上分别达到了61.1%和68.7%，表现优于无贝叶斯化和FFN变体；在Snopes数据集上，FFN方法在F1-Macro和F1-Micro指标上分别达到59.9%和70.6%，优于无贝叶斯化和MHA变体。这表明贝叶斯化处理的位置对不同数据集的影响各异。因此，在EBNN-FND模型中，本文对FFN层和MHA层均实施了贝叶斯化处理，即FFN+MHA。尽管MHA变体在Politifact数据集的F1-Micro指标达到了68.6%，超过了EBNN-FND模型的66.8%，但在F1-Macro指标上仅为61.1%，低于EBNN-FND模型的66.6%。这一差异可能源于MHA变体未能整合证据模块，从而学习到了观点与标签间的虚假关联。由于F1-Macro指标综合考虑了各类别的表现，MHA变体在此指标上表现不佳，而观点-证据交互模块通过消除虚假关联，准确捕捉证据与观点间的因果关系，证实了其有效性。

此外，在F1-Macro指标下进行对比，图4的各项数据均小于表4中EBNN-FND的各项数据，具体来说，图4（a）中Snopes数据集中的最大值为59.9%，小于表4中EBNN-FND模型在Snopes数据集中的66.1%，图4（a）中Politifact数据集中的最大值为61.1%，小于表4中EBNN-FND模型在Politifact数据集中的66.6%。这表明了在仅有观点输入的条件下，EBNN-FND的F1-Macro指标小于融合了证据与观点输入的结果，佐证了观点-证据交互模块的有效性。

为了进一步验证EBNN-FND方法的有效性，本文调整了训练数据集的保留比例，不同EBNN-FND变体在两个困难数据集上的对比如图5所示。图5中的mark表示训练数据集的保留比例，如mark=0.5意味着使用了原训练集的50%进行训练。

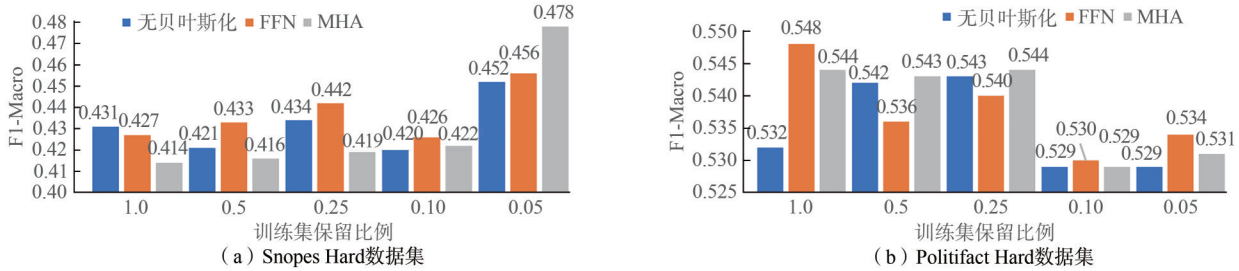
在数据集保留比例为0.1与0.05时，FFN和MHA变体在两个数据集上相较于无贝叶斯化变体均表现更佳。通常情况下，FFN和MHA变体的性能普遍优于无贝叶斯化变体。这表明，无论是针对FFN层，还是MHA层的贝叶斯化处理，均能在一定程度上提高模型处理困难样本的分类性能，这可能归因于贝叶斯化方法能够更有效地捕捉数据不确定性，增强模型的泛化能力和鲁棒性。

在考虑证据输入的情况下，EBNN-FND不同变体在LIAR数据集上的对比见表6，EBNN-FND不同变体在各输入源对比见表7。其中，CB表示CNN-双向长短期记忆神经网络（bidirectional long



（a）F1-Macro　　（b）F1-Micro

图4　EBNN-FND模型在各数据集上的F1-Macro和F1-Micro指标对比



（a）Snopes Hard数据集　　（b）Politifact Hard数据集

图5　不同EBNN-FND变体在两个困难数据集上的对比

short-term memory，BiLSTM）。本文设计了 EBNN-FND 模型的 5 种变体，分别为 TextCNN+BNN、BNN+TextCNN、BNN+BNN、TextCNN+TextCNN+BNN 与 TextCNN+TextCNN+CB+BNN。在这些变体中，外源证据的文本信息和数字信息模型输出均会输入观点−证据交互模块，并与观点模型输出进行交互，产生不同的观点−证据特征交互向量，随后输入特征混合模块。若外源证据文本信息对应多个模型，则分别输入各模型，并将得到的观点−证据交互特征向量进行混合。以 TextCNN+TextCNN+CB+BNN 变体为例，观点信息由 TextCNN 处理后，与 TextCNN、BNN 和 CB 外源证据信息进行特征交互，形成 3 种特征向量，再输入特征混合模块。

表 6　EBNN-FND 不同变体在 LIAR 数据集上的对比

模型	准确率	F1-Macro	F1-Micro	精确率	召回率
TextCNN+BNN	0.227 3	0.169 2	0.227 3	0.322 3	0.207 8
BNN+TextCNN	0.264 4	0.241 0	0.264 4	0.291 8	0.246 3
BNN+BNN	0.196 5	0.063 1	0.196 5	0.070 7	0.166 3
TextCNN+TextCNN+BNN	0.252 6	0.215 0	0.252 6	0.280 1	0.233 8
TextCNN+TextCNN+CB+BNN	0.446 7	0.454 6	0.446 7	0.481 0	0.446 3
EBNN-FND	0.460 1	0.458 1	0.460 1	0.533 5	0.457 1

表 6 结果显示，EBNN-FND 在各项性能指标上均优于其他变体，原因如下：（1）EBNN-FND 不仅利用外源证据文本信息，还整合了数字信息，增强了模型对证据的全面理解和分析能力，通过结合结构化数字信息和非结构化文本信息，多维度提取特征，提高模型在虚假信息检测中的准确性和鲁棒性；（2）TextCNN 相较于 BNN，更适合捕捉观点文本特征，其不同大小的卷积核能提取 *N*-gram 特征，强化局部特征提取能力；（3）BNN 相较于 TextCNN，更适合捕捉外源证据文本特征，BNN 基于 Transformer，具有强大的文本表示能力，能建模不确定性，捕捉长距离依赖关系，并通过注意力机制全局编码证据文本，深入理解和利用上下文信息。此外，BNN 模型的灵活性和可扩展性也使其在处理多样化证据文本时表现更佳。

3.2.3　超参数的影响

在观点−证据交互模块中的去偏操作中，本文引入了超参数 λ 作为调整系数，不同条件下，EBNN-FND 在两个数据集上的性能比对如图 6 所示。

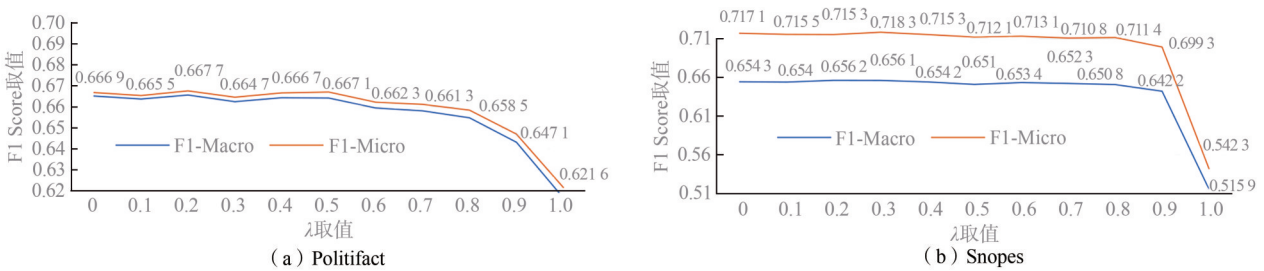
从图 6 可以看出，若将 F1-Macro 作为主要指标进行选择，在 Snopes 数据集中，$\lambda=0.2$ 时出现最大值，其 F1-Macro 指标为 65.62%；在 Politifact 数据集中，$\lambda=0.2$ 时出现最大值，其 F1-Macro 指标为 66.57%。此外，随着 λ 值的不断增大，各指标均呈现下降趋势。

4　结束语

针对虚假信息检测中模型预测结果可信度不足、数据与模型不确定性难以量化的问题，本文提

表 7　EBNN-FND 不同变体在各输入源对比

模型	观点	外源证据文本信息	外源证据数字信息
TextCNN+BNN	TextCNN	BNN	无
BNN+TextCNN	BNN	TextCNN	无
BNN+BNN	BNN	BNN	无
TextCNN+TextCNN+BNN	TextCNN	TextCNN+BNN	无
TextCNN+TextCNN+CB+BNN	TextCNN	TextCNN+BNN	CB
EBNN-FND	TextCNN	BNN	CB



（a）Politifact　　（b）Snopes

图 6　不同 λ 条件下，EBNN-FND 在两个数据集上的性能比对

出了一种参数量级相对较小、融合贝叶斯推理与证据分析的改进方法，建立了EBNN-FND模型。该模型通过BNN框架对检测过程中的参数不确定性进行概率建模，同时引入观点−证据交互模块消除数据偏差对因果推断的影响，实现了观点文本与外部证据的深度特征融合。实验结果表明，EBNN-FND模型在Politifact、Snopes和LIAR数据集上的各项检测性能均优于现有基线模型，验证了其在虚假信息检测任务中的有效性和稳定性，为复杂信息传播环境下的虚假信息治理提供了兼具理论严谨性和实用性的技术方案，为处理信息传播中的不确定性问题提供了新的解决思路。

未来，笔者将从以下3个方面深化研究以优化EBNN-FND性能：第一，设计更高效的不确定性量化框架，增强模型对稀疏数据的泛化能力；第二，融合多模态信息源，提升复杂虚假信息模式的捕获精度；第三，构建分布式推理架构，扩展模型对海量社交媒体数据的实时处理能力。相关研究将为动态信息环境下的虚假信息治理提供兼具理论严谨性与工程实用性的技术方案。

参考文献：

[1] CARANNANTE G, BOUAYNAYA N C. Bayesian deep learning detection of anomalies and failure: application to medical images[C]// Proceedings of the 2023 IEEE 33rd International Workshop on Machine Learning for Signal Processing (MLSP). Piscataway: IEEE Press, 2023: 1-6.

[2] MA R H, ZHANG H, ZHANG J, et al. Bayesian uncertainty modeling for P300-based brain-computer interface[J]. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 2023, 31: 2789-2799.

[3] XUAN J Y, LU J, ZHANG G Q. A survey on Bayesian nonparametric learning[J]. ACM Computing Surveys, 2020, 52(1): 1-36.

[4] 许旻辰, 屈丹, 司念文, 等. 社交媒体虚假信息检测技术研究综述[J]. 计算机工程, 2025, doi: 10.19678/j.issn.1000-3428.0070287.
XU M C, QU D, SI N W, et al. A survey of the technologies for detecting disinformation in social media[J]. Computer Engineering, 2025, doi: 10.19678/j.issn.1000-3428.0070287.

[5] 张元园, 袁嘉雱. 基于社交媒体的谣言检测研究综述[J]. 数据通信, 2024(1): 28-33.
ZHANG Y Y, YUAN J J. A review of rumor detection research based on social media[J]. Data Communications, 2024(1): 28-33.

[6] 张欣, 孙靖超. 基于大语言模型的虚假信息检测框架综述[J]. 计算机科学与探索, 2025, 19(6): 1414-1436.
ZHANG X, SUN J C. Review of false information detection frameworks based on large language models[J]. Journal of Frontiers of Computer Science and Technology, 2025, 19(6): 1414-1436.

[7] 袁唯淋, 赵卫伟, 胡振震, 等. 智能情报融合综述: 对抗视角下的开源情报融合分析[J]. 智能科学与技术学报, 2024, 6(3): 284-300.
YUAN W L, ZHAO W W, HU Z Z, et al. Research on intelligence fusion: a holistic analysis of open-source intelligence fusion from the perspective of confrontation[J]. Chinese Journal of Intelligent Science and Technology, 2024, 6(3): 284-300.

[8] 刘宇栋, 黄千里, 王恒, 等. 基于数据增强的多模态虚假信息检测框架研究[J]. 信息安全研究, 2025, 11(4): 377-384.
LIU Y D, HUANG Q L, WANG H, et al. Research on data-enhanced multi-modal false information detection framework[J]. Journal of Information Security Research, 2025, 11(4): 377-384.

[9] 徐绪堪, 李溢, 李一铭. 融合传播路径和多模态特征的社交媒体信息可信度模型构建[J]. 情报理论与实践, 2025, 48(5): 168-176.
XU X K, LI Y, LI Y M. Construction of a social media information credibility model that integrates propagation paths and multi-modal features[J]. Information Studies (Theory & Application), 2025, 48(5): 168-176.

[10] LIU Q, YU F, WU S, et al. Mining significant microblogs for misinformation identification[J]. ACM Transactions on Intelligent Systems and Technology, 2018, 9(5): 1-20.

[11] CHANDRA S, MISHRA P, YANNAKOUDAKIS H, et al. Graph-based modeling of online communities for fake news detection[EB]. arXiv preprint, 2020, arXiv: 2008.06274.

[12] ZHANG X Y, CAO J, LI X R, et al. Mining dual emotion for fake news detection[C]//Proceedings of the Web Conference 2021. New York: ACM, 2021: 3465-3476.

[13] 张仰森, 彭媛媛, 段宇翔, 等. 基于评论异常度的新浪微博谣言识别方法[J]. 自动化学报, 2020, 46(8): 1689-1702.
ZHANG Y S, PENG Y Y, DUAN Y X, et al. The method of sina weibo rumor detecting based on comment abnormality[J]. Acta Automatica Sinica, 2020, 46(8): 1689-1702.

[14] XU W Z, WU J F, LIU Q, et al. Evidence-aware fake news detection with graph neural networks[C]//Proceedings of the ACM Web Conference 2022. New York: ACM, 2022: 2501-2510.

[15] WU J F, LIU Q, XU W Z, et al. Bias mitigation for evidence-aware fake news detection by causal intervention[C]//Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2022: 2308-2313.

[16] CASTILLO C, MENDOZA M, POBLETE B. Information credibility on twitter[C]//Proceedings of the 20th International Conference on World Wide Web. New York: ACM, 2011: 675-684.

[17] MA J, GAO W, WEI Z Y, et al. Detect rumors using time series of social context information on microblogging[C]//Proceedings of the 24th ACM International on Conference on Information and Knowledge Manageme. New York: ACM, 2017: 1751-1754.

[18] GIASEMIDIS G, SINGLETON C, AGRAFIOTIS I, et al. Determining the veracity of rumours on twitter[C]//International Conference on Social Informatics. Cham: Springer, 2016: 185-205.

[19] MA J, GAO W, MITRA P, et al. Detecting rumors from microblogs with recurrent neural networks[C]//Proceedings of the 25th International Conference on World Wide Web. New York: ACM, 2016: 1861-1872.

[20] RUCHANSKY N, SEO S, LIU Y. CSI: a hybrid deep model for fake news detection[C]//Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. New York: ACM, 2017: 797-806.

[21] JIN Z W, CAO J, GUO H, et al. Multimodal fusion with recurrent neural networks for rumor detection on microblogs[C]//Proceedings of the 25th ACM International Conference on Multimedia. New York: ACM, 2017: 795-816.

[22] KUMAR S, ASTHANA R, UPADHYAY S, et al. Fake news detection using deep learning models: a novel approach[J]. Transactions on Emerg-

ing Telecommunications Technologies, 2020, 31(2): e3767.
[23] ABDULLAH A A, HASSAN M M, MUSTAFA Y T. A review on Bayesian deep learning in healthcare: applications and challenges[J]. IEEE Access, 2022, 10: 36538-36562.
[24] GHAHRAMANI Z. Probabilistic machine learning and artificial intelligence[J]. Nature, 2015, 521(7553): 452-459.
[25] HÜLLERMEIER E, WAEGEMAN W. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods[J]. Machine Learning, 2021, 110(3): 457-506.
[26] DEVLIN J, CHANG M W, LEE K, et al. Bert: pre-training of deep bidirectional transformers for language understanding[EB]. arXiv preprint, 2019, arXiv: 1810.04805.
[27] KIM Y. Convolutional neural networks for sentence classification[C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg: ACL, 2014: 1746-1751.
[28] XUE B Y, HU S K, XU J H, et al. Bayesian neural network language modeling for speech recognition[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2022, 30: 2900-2917.
[29] DUAN H P, YANG L X, FANG J, et al. Fast inverse-free sparse Bayesian learning via relaxed evidence lower bound maximization[J]. IEEE Signal Processing Letters, 2017, 24(6): 774-778.
[30] BLUNDELL C, CORNEBISE J, KAVUKCUOGLU K, et al. Weight uncertainty in neural networks[J]. 32nd International Conference on Machine Learning, ICML 2015, 2015, 2: 1613-1622.
[31] XUE B Y, YU J W, XU J H, et al. Bayesian transformer language models for speech recognition[C]//Proceedings of the ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2021: 7378-7382.
[32] KINGMA D P, WELLING M. Auto-encoding variational bayes[EB]. arXiv preprint, 2014, arXiv: 1312.6114.
[33] XU C, KECHADI M T. Fuzzy deep hybrid network for fake news detection[C]//Proceedings of the 12th International Symposium on Information and Communication Technology. New York: ACM, 2023: 118-125.
[34] HANSEN C, HANSEN C, CHAVES LIMA L. Automatic fake news detection: are models learning to reason? [C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). Stroudsburg: ACL, 2021: 80-86.
[35] WANG W Y. "Liar, liar pants on fire": a new benchmark dataset for fake news detection[EB]. arXiv preprint, 2017, arXiv: 1705.00648.
[36] HANSEN C, HANSEN C, LIMA L C. Automatic fake news detection: are models learning to reason? [EB]. arXiv preprint, 2021, arXiv: 2105.07698.
[37] POPAT K, MUKHERJEE S, YATES A, et al. DeClarE: debunking fake news and false claims using evidence-aware deep learning[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2018: 22-32.
[38] KARIMI H, ROY P, SABA-SADIYA S, et al. Multi-source multi-class fake news detection[C]// Proceedings of the 27th International Conference on Computational Linguistics. Stroudsburg: ACL, 2018: 1546-1557.
[39] GUO J Y, DU L, BI W D, et al. Homophily-oriented heterogeneous graph rewiring[C]//Proceedings of the ACM Web Conference 2023. New York: ACM, 2023: 511-522.
[40] LIU C, WU X H, YU M, et al. A two-stage model based on BERT for short fake news detection[C]//International Conference on Knowledge Science, Engineering and Management. Cham: Springer International Publishing, 2019: 172-183.
[41] LONG Y, LU Q, XIANG R, et al. Fake news detection through multi-perspective speaker profiles[C]//Proceedings of the Eighth International Joint Conference on Natural Language Processing. Stroudsburg: ACL, 2017: 252-256.
[42] XU C, KECHADI M T. An enhanced fake news detection system with fuzzy deep learning[J]. IEEE Access, 2024, 12: 88006-88021.

[作者简介]

陈君海（2001− ），男，国防科技大学智能科学学院硕士生，主要研究方向为人工智能、真假鉴别。

项凤涛（1986− ），男，博士，国防科技大学智能科学学院副教授，主要研究方向为智能辅助决策、不确定性推理、智能控制。

黎拓新（2002− ），男，国防科技大学智能科学学院硕士生，主要研究方向为模式识别人工智能、迁移学习。

罗翔宇（2002− ），男，国防科技大学智能科学学院硕士生，主要研究方向为人工智能、时空学习。