基于注意力机制多特征融合的虚假信息检测

地力夏提·阿布都热依木[1,2,3]，马博[1,2,3*]，杨雅婷[1,2,3]，王磊[1,2,3]

（1.中国科学院新疆理化技术研究所， 新疆 乌鲁木齐市 8300111； 2. 中国科学院大学，北京 100049； 3. 新疆理化技术研究所 新疆民族语音语言信息处理实验室，新疆 乌鲁木齐 8300111）

摘要：在虚假信息识别任务中，面对图文结合的虚假内容，基于单模态的模型难以进行准确识别。如何充分利用多模态的信息，准确快速地识别突发事件中的虚假信息是一个挑战。为此，本文提出了基于注意力机制多特征融合的虚假信息检测方法（att-MFNN）。该模型中文本特征和情感特征基于注意力机制融合，与视觉特征组成多模态特征并送入虚假信息识别器和事件分类器中。通过引入事件分类器学习不同事件中共同特征，提高对于新事件的识别性能。att-MFNN 在微博（weibo）和推特（twitter）数据集准确度达到了 89.22%和 87.51%，并且 F1、Precision、Recall 指标均优于现有的模型。

关键词：虚假信息检测；多特征融合；注意力机制；情感提取

中图分类号：TP391　　文献标志码： A

前言

网络和社交媒体的快速发展，降低了传播信息的成本，使得人们之间的交流更加频繁。但这也给虚假信息的迅速传播提供了机会。微博，推特等社交媒体因及时和全面的提供世界各地的新闻，也已经成为各种事件中重要的新闻媒体和舆论平台。而在突发事件发生时，虚假新闻的数目也会明显增多。例如新冠肺炎疫情在世界各地爆发期间，社交媒体中关于肺炎起源，预防，诊断和治疗和传播的虚假信息随处可见。虚假信息在打破信息生态系统的真实性平衡的同时会影响人们理解和回应真实信息的方式[1]。因此，若不及时识别出虚假信息，可能会造成大规模的负面影响，甚至会造成不可挽回的损失。

近年来，提出了许多虚假新闻识别的方法，主要分为基于单模态和基于多模态的方法。而基于多模态的虚假信息检测则开始成为技术重点。现有的多模态深度学习模型由于其优越的特征提取能力，在性能上优于传统的深度学习模型。Jin[2]等人首次提出了基于循环神经网络（Recurrent Neural Network, RNN）[3]的多模态模型，提供了通过不同模态的特征提取融合的方式进行虚假信息检测的思路。Wang[4]等人利用引入对抗学习消除不同事件的中的特殊特征，提高了对新事件有关的虚假信息检测准确度。Zhang[5]等人利用帖子文本和评论内容提出用户情感、发布者情感、两者差值组成的双重情感结合文本特征进行谣言检测。Bian等人[6]基于双向图卷积神经网络（Bi-GCN），自顶向下和自底向上挖掘虚假信息的传播和散布模式进行识别。文献[7]中通过专家标注的大量文档，学习文档的写作风格，基于风格特征进行虚假信息检测。

为进一步提高多模态特征的假新闻检测准确率，本文提出了一种端到端模型，即基于注意力机制多特征融合的神经网络（Attention based Multi-Feature Fusion Neural Network for

基金项目：中国科学院,青年创新促进会项目，科发人函字[2019]26 号；国家自然科学基金项目,（U2003303）；中国科学院西部青年学者项目（A 类）,（2019-XBQNXZ-A-004）；国家重点研发计划,（2018YFC0823002）；新疆天山创新团队项目,（2020D14045）

***通讯作者：**mabo@ms.xjb.ac.cn

Fake News Detection,att-MFNN）。att-MFNN 主要由四个模块组成:多模态特征提取器、多特征融合器、事件分类器和虚假信息识别器。多模态特征提取器采用预训练的基于变换器的双向编码器表示模型(Bidirectional Encoder Representations from Transformers, *BERT*)[8]模型提取文本特征，采用预训练的 VGG-19[9]模型提取图像特征，通过情感特征提取器提取情感特征。提取的特征通过注意力机制融合后输入虚假信息检测器中识别虚假信息。为提升模型的对于突发事件有关的谣言检测能力，模型中加入了事件分类器，学习不同事件之间的共同特征。att-MFNN 在 Weibo 和 Twitter 数据集中准确度达到了 89.22%和 87.51%，性能指标均优于基线模型。

本文主要贡献如下：

1）提出了将文本特征、视觉特征和情感特征通过注意力机制融合的领域自适应神经网络用于虚假信息检测。

2）在两个数据集中，att-MFNN 在准确度、F1 值、精确度、召回率等性能指标上均优于基线模型。

在后面部分中：在第一节我们回顾了相关工作。第二节我们介绍了本文中 att-MFNN 及其不同组件的具体细节。在第三节中，我们介绍了使用的数据集、实验设置和基线模型，并展示了实验结果和并对结果进行讨论和分析。最后，在第四节对全文进行了总结。

1 相关工作

现有的大多数关于虚假新闻检测的研究都是基于特征的，这些特征可以从文本、社交背景和图像中提取。为了提取文本特征用于虚假信息检测，Ma 等人[10]引入了基于循环神经网络 RNN 的模型捕获相关帖子的社交上下文特征随时间的变化。之后在此工作上，chen 等人[11]将注意力机制（Attention Mechanism）[12]加入 RNN 中以选择性地提取时间表征。Wang[3]等人工作中提出基于卷积神经网络（Convolutional Neural Networks,CNN）[13]改进的 Text-CNN 模型提取文本特征。在本文中，为提取帖子中的文本特征，使用预训练模型 BERT 提取文本特征。

最近的研究中也有侧重于情感特征的虚假信息检测。文献[14]中指出，虚假新闻的内容和它们的情感词之间存在联系，他们构造了一个情感特征(负面和正面词的计数之比)来帮助检测虚假新闻。Guo 等人[15]从情感角度探索虚假信息的传播模式，观察发现，虚假内容编造者偏好使用带有强烈情感的语句和容易引起关注的热点话题传播虚假信息。因此，该文献提出了基于文本内容和评论情感的双重情感融合模型。之后，Zhang[4]等人发现虚假内容的文本和用户评论之间存在着明显的情感反差，所以将文本情感、用户情感、文本情感和用户情感差值组成双重情感特征模型，进行虚假信息检测。在本文中，我们基于情感特征提取器，提取不同的情感子特征，拼接后形成最终的情感特征。

视觉特征是多模态虚假信息检测中的重要指标。之前的研究[2,3,16]证明了从帖子中图片提取的视觉特征包含着关键的信息，可以通过深度神经网络提取高度复杂的图像表示，得到显著改善的结果。在本文中，我们采用 VGG19 提取图像特征，并与文本、情感特征进行融合。

现有的多模态模型融合特征时多数采用拼接或相加的方式，导致模态信息冗余，不能有效地结合不同模态的优势。Xu 等人[17]通过基于注意力机制提取模态间的交互信息，证明了不同模态的特征的组合可以获得更全面的特征表示。Guo 等人[15]设计了三种门（Gate）在不同级别进行特征融合。因此，本文中提出了将文本特征和情感特征基于注意力机制融合的方法以获取更加丰富的多模态特征。

为增加模型的鲁棒性，在其他自然语言处理任务中会引入生成对抗学习（Generative Adversarial Networks，GAN）[18]。Li 等人[19]中，在多模态融合过程中加入双鉴别器，识别多个模态共同的特征并进行学习。在虚假信息检测领域，Wang 等人[3]引入了事件鉴别器来

去除事件的特殊特征。因此，受文献[3,19]的启发，在本文中，我们也加入了一个事件分类器，目的是从多模态特征中去除事件独有的特征，引导模型学习事件的通用特征。

2 模型介绍

图 1 为 att-MFNN 模型的总体架构，总体分为多模态特征提取器、多特征融合器、事件分类器和虚假信息识别器。多特征提取器又细分为文本特征提取器、情感特征提取器、视觉特征提取器。提取的文本和情感特征通过多特征融合器进行融合，再与视觉特征拼接，形成最终的多模态特征。再将多模态特征送入事件鉴别器和虚假信息识别器中，获取分类结果。

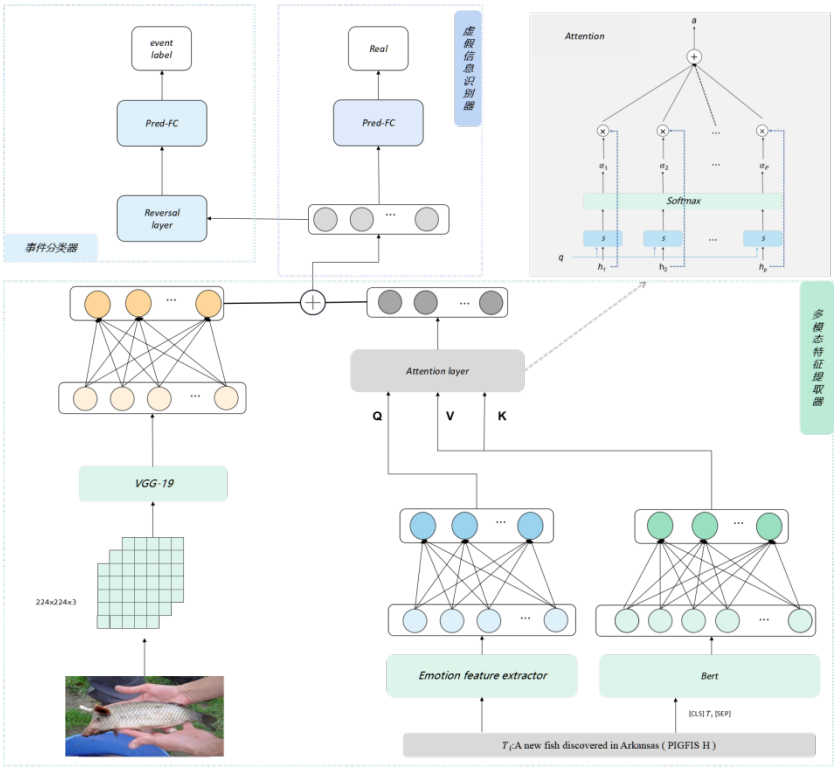

图 1 基于注意力机制多特征融合的神经网络框架图

Fig.1 Multi-feature fusion based neural network for fake news detection(att-MFNN)

2.1 多模态特征提取

2.1.1 文本特征提取

为了捕捉潜在的语义和上下文含义，本文中我们使用包含 12 个编码器层的 $BERT_{base}$ 预训练模型进行文本特征提取。我们将输入的含有 n 个单词的文本可以表示为 $T = [T^0, T^1, \ldots, T^n]$，$T^0$ 表示[CLS]的嵌入。将 T_i 输入 $BERT_{base}$ 预训练模型得到特征向量 T_f，并进行均值池运算获得文本特征 $R_t \in R^B$，其中 B 表示从 BERT 获得的文本特征的维数,随后输入全连接层，以确保文本特征的最终输出(表示为 $R_{tf} \in R^p$)具有与视觉特征相同的维度(表示为 p)。因此,

$$R_{tf} = \sigma_t(W_{tf} \cdot R_t) \tag{1}$$

其中 $W_{tf} \in R^{B \cdot p}$ 是文本特征提取器中全连接层的权重矩阵，σ_t 是文本特征提取器中使用的 Leaky RELU 激活函数。

2.1.2 情感特征提取

文本情感特征由情感类别，情感词汇，情感强度以及辅助特征组成。

（1）情感类别：我们使用情感分类器（推特数据集使用 NVIDIA 的开源模型情感分类

模型[2]，微博数据集使用百度 AI 平台的情感识别[3]）来获取情感类别特征。将文本 T 输入到情感分类器f中，得到预测值$f(T)$，提取情感特征$Emo^c \in R^{d_f}$。情感类别记为$E = (e_1, e_2, \cdots, e_{d_f})$

$$Emo^c = f(T) \tag{2}$$

（2）情感词汇：一个文本通过几个特定的词语来传达特定的情感，通过情感词典计算出该文本的情感词汇特征。记情感词列表为ε_e，若单词T_i在此列表中,表示与情感词列表匹配成功记为 1，否则为 0。

$$\text{match}(T_i) = \begin{cases} 1, & T_i \in \varepsilon_e \\ 0, & \text{其他} \end{cases} \tag{3}$$

公式（4）中，$s(T,e)$表示文本 T 中每个单词的情感分数之和，$neg(T_i)$和$deg(T_i)$表示T_i的否定值和程度值。

$$s(T,e) = \sum_{i=1}^{L} s(T_i,e) = \frac{match(T_i)*neg(T_i)*deg(T_i)}{L} \tag{4}$$

公式（5）中最后将不同的情感类别拼接形成该文本的情感词汇特征。

$$Emo^{Lexion} = s(T,e_1) \oplus s(T,e_2) \oplus \cdots \oplus s\left(T,e_{d_f}\right) \tag{5}$$

（3）情感强度：类似情感词汇特征，公式（6）中$intensity(T_i)$表示若T_i在情感词典列表中，则按照表中值进行计算，否则$intensity(T_i) = 0$。并计算出情感程度值$I(T,e)$

$$I(T,e) = \sum_{i=1}^{L} I'(T_i,e) = \sum_{i=1}^{L} intensity(T_i) * s(T_i,e) \tag{6}$$

如公式（7）所示，最后将不同类别的情感程度值组成情感强度特征。

$$\text{Emo}^{\text{Intensity}} = \text{I}(\text{T},\text{e}_1) \oplus \text{I}(\text{T},\text{e}_2) \oplus \cdots \oplus \text{I}\left(\text{T},\text{e}_{\text{d}_\text{f}}\right) \tag{7}$$

（4）情感辅助特征：许多研究中，在数据处理过程中会将文本中的表情，标点符号清洗掉，其实这些符号在社交媒体中也是重要的情绪表达。分别统计不同类别的表情，符号的出现的频率。通过情感词典或公共工具包可以计算一条文本正/负极性程度，将以上辅助信息统一称为情感辅助特征，记为$Emo^{aux} \in R^a$。

最后上述四种特征的拼接后得到文本情感特征$Emo^T \in \text{R}^{\text{d}_\text{e}}$，如公式（8）所示：

$$\text{Emo}^{\text{T}} = [\text{Emo}^{\text{Category}}, \text{Emo}^{\text{Lexion}}, \text{Emo}^{\text{Intensity}}, \text{Emo}^{\text{aux}}] \tag{8}$$

获得的情感特征维数为 d_e, 情感特征提取器最后一层添加全连接层，作用是将情感特征输出成维度为R_{ef}（与文本特征维度一致）的最终形式。因此，

$$\text{R}_{\text{ef}} = \sigma_\text{e}(\text{W}_{\text{ef}} \cdot \text{R}_\text{e}) \tag{9}$$

其中$\text{W}_{\text{ef}} \in \text{R}^{\text{d}_\text{e}\cdot\text{p}}$是情感特征提取器中全连接层的权重矩阵，$\sigma_\text{e}$是情感特征提取器中使用的 Leaky RELU 激活函数。

2.1.3 视觉内容特征提取

CNN 已经成功应用于各种视觉理解问题。在本文中，我们使用预训练的 VGG-19，从帖子所含图像中提取视觉特征。将获得的图像特征的维数表示为 d_v。为了使视觉特征的最终输出(表示为 R_{vf})与文本和情感特征的维度一致，我们在 VGG-19 的最后一层添加了一个全连接层。因此，

$$R_{vf} = \sigma_v(W_{vf} \cdot R_v) \tag{10}$$

其中，$\text{W}_{\text{vf}} \in \text{R}^{\text{d}_\text{v}\cdot\text{p}}$是视觉特征提取器中全连接层的权重矩阵，$\text{R}_\text{v}$是 VGG-19 最后一层的输出，$\sigma_\text{v}$表示是视觉特征提取器中的 Leaky RELU 激活函数。

---

[2] https://github.com/NVIDIA/sentiment-discovery

[3] https://ai.baidu.com/tech/nlp_apply/emotion_detection

2.2 多特征融合器

（1）特征拼接

将R_{tf}, R_{ef}和R_{vf}三种特征连接成一个维数为 3p 的向量，记为$R_f \in R^{3p}$。此外，我们将多模态特征提取器表示为 $E(P;\theta_e)$，其中 P 表示向量化的输入，θ表示多模态提取器的参数集合，E 表示整体映射函数。因此，我们得到:

$$R_f = E(P;\theta_e) \tag{11}$$

（2）注意力机制融合

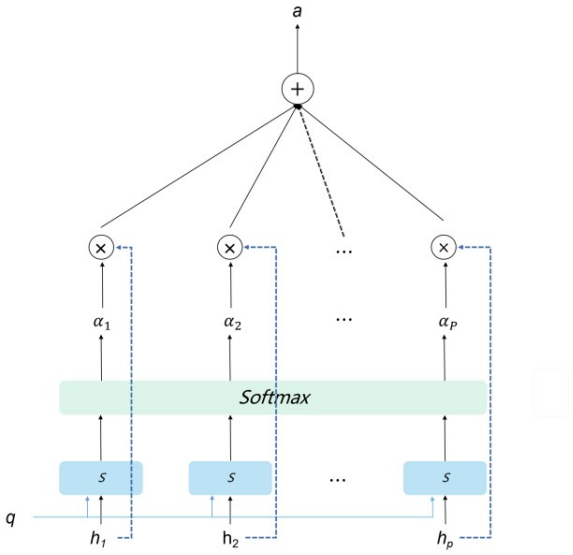

图 2 基于注意力机制的特征融合

Fig.2 Feature fusion based attention

将文本特征$R_{tf} = [h_1, h_2, \cdots, h_p]$，情感特征$R_{ef}$，分别设置为$Key = Value = R_{tf}$和$q = R_{ef}$后，按照如下方式进行特征融合：

（1）根据 *Query* 和 *Key* 计算二者的相似度，得到注意力得分。

$$s_i = F(Q, k_i) \tag{12}$$

（2）用 *softmax* 函数对注意力得分进行数值转换，进行归一化得到所有权重系数和为 1 的概率分布。

$$\alpha_i = softmax(s_i) = \frac{exp(s_i)}{\sum_{j=1}^{N} exp(s_j)} \tag{13}$$

（3）根据权重系数对 *value* 进行加权求和。

$$\mathrm{Attention}((\mathrm{K},\mathrm{V}),\mathrm{Q}) = \sum_{\mathrm{i}=1}^{\mathrm{N}} \alpha_i v_i \tag{14}$$

将文本特征R_{tf}和情感特征R_{ef}输入注意力层，得到文本情感特征$R_{tef} \in R^p$：

$$R_{tef} = A(R_{tf}, R_{ef}) \tag{15}$$

将R_{tef}和R_{vf}两种特征连接成一个维数为 $2p$ 的向量，记为$R_f \in R^{2p}$。此外，我们将多模态特征提取器表示为$E(P;\theta_e)$，其中 P 表示向量化的输入，θ表示多模态提取器的参数，**E** 表示整体映射函数。因此，我们得到:

$$R_{af} = E(P;\theta_e) \tag{16}$$

2.3 虚假新闻识别器

本文中，虚假新闻识别器由两个全连接层和 softmax 函数构成，并将检测器表示为

$C(R_f;\theta_c)$，其中θ_c表示检测器的参数，C 表示检测器的映射函数。对于输入的帖子 Post，假新闻检测器的输出$\hat{y}_i$表示该帖子为虚假信息的概率，因此定义如下：

$$\hat{y}_i = C(E(P_i;\theta_e);\theta_c) \tag{17}$$

用y_i表示帖子的真实标签，当帖子标签为虚假时为 1，真实时为 0。所有帖子真实标签y_i的集合，记为 Y。为了计算分类损失，本文中采用交叉熵损失函数如公式（18）所示：

$$L_C(\theta_e,\theta_c) = -E_{(p,y_i)\in(P,Y)}[y\log(\hat{y}_i) + (1-y)\log(1-\hat{y}_i)] \tag{18}$$

公式（19）分类损失最小化：

$$(\theta_e^*,\theta_c^*) = arg\min_{\theta_e,\theta_c} L_C(\theta_e,\theta_c) \tag{19}$$

2.4 事件分类器

为了提高未知事件和突发事件中的识别能力，我们引入事件分类器 $D(R_f;\theta_d)$，θ_d表示事件分类器的参数 D 为映射函数。我们将所有事件的集合记为 M，将多模态特征 R_f输入到分类器中，会将帖子 p 分类到 M 个事件中的一个。用 Z 来表示事件的标签集合，并且用交叉熵损失函数定义事件分类器的损失，如下：

$$L_d(\theta_e,\theta_d) = -E_{(p,y_i)\in(P,Z)}[\sum_{m=1}^{M} z\ log\ D(E(P_i;\theta_e);\theta_d))] \tag{20}$$

2.5 模型参数

多模态提取器倾向于通过最大化的事件分类损失 L_d来提取事件通用的特征，而事件分类器倾向于通过最小化事件分类损失 L_d来从多模态特征中发现事件独有特征。

定义最终损失为公式（21）：

$$L(\theta_e,\theta_c,\theta_d) = L_C(\theta_e,\theta_c) - \lambda L_d(\theta_e,\theta_d) \tag{21}$$

系数$\lambda \in R$用于平衡两个分类器的损失，使其达到两种损失对抗的效果。我们采用 Ganin 等人[25]引入的梯度反转层(GRL)。因此，模型参数的优化过程描述如下:

$$\theta_e \leftarrow \theta_e - \eta(\frac{\partial L_C}{\partial \theta_e} - \lambda\frac{\partial L_d}{\partial \theta_e}) \tag{22}$$

虚假信息识别器的参数更新：

$$\theta_c \leftarrow \theta_c - \eta\frac{\partial L_C}{\partial \theta_c} \tag{23}$$

事件分类器的参数更新：

$$\theta_d \leftarrow \theta_d - \eta\frac{\partial L_d}{\partial \theta_d} \tag{24}$$

3 实验

3.1 实验参数选取

在文本特征提取器中，从 $BERT_{base}$ 获得的文本特征的维数 d_t 为 768。对于视觉提取器，我们首先将图像的大小调整为 $224 \times 224 \times 3$，输入到预训练模型 VGG-19 中。VGG-19 的图像特征维数为 4096。通过情感特征提取器中情感特征维数为 56。文本、情感和视觉提取器中全连接层的隐藏状态 p 维数为 32。为避免过拟合，模型中对 $bert_{base}$ 和 VGG-19 的参数均进行了冻结。事件分类器中两个全连接层的维数分别设置为 64 和 32。

模型中批处理（batchsize）设置为 32，训练次数为 100 轮（epoch），学习速率为 10^{-3}，

优化算法为 Adam，每个连接层的激活函数为 Leaky RELU，dropout 概率为 0.5。

3.2 数据集

微博数据集：微博数据集在[2]中用于检测虚假信息的。在这个数据集中，真实信息是从中国的权威新闻来源收集的，比如新华社。另一方面通过微博官方辟谣系统收集了 2012 年 5 月到 2016 年 1 月的虚假信息。该系统鼓励普通用户报告可疑帖子，并由可信用户检查可疑帖子。因此这个系统也作为收集谣言新闻的权威来源。当预处理这个数据集时，我们首先删除重复和低质量的图像，以确保整个数据集的质量。然后，我们应用单程聚类方法[20]从帖子中发现新出现的事件。最后，我们将整个数据集以 7:1:2 的比例分成训练集、验证集和测试集，并确保它们不包含任何常见事件。

推特数据集：Twitter 数据集来自 Boididou 等人[21]发布的数据集，用于检测 Twitter 上的虚假内容。Twitter 数据集包含开发集和测试集，我们将 Twitter 开发集用作训练集，将测试集用作测试集。Twitter 数据集中的推文包含文本内容，附加的图像/视频和其他社交环境信息。在这项工作中，我们专注于通过结合不同模态信息来检测虚假信息。因此，我们删除了没有任何文本或图像的推文。并且训练集与测试集之间没有重叠事件。表 1 列出了这两个数据集的详细统计数据：

表 1 微博和推特数据集统计信息
Table 1 statistics of Weibo and Twitter dataset

数据集	标签	数目	总计
Twitter	Real news	7021	12995
	Fake news	5974	
Weibo	Real news	4749	9528
	Fake news	4779	

3.3 基线模型

EANN[3]模型：EANN 结合文本和视觉特征通过拼接组成多模态特征，并引入事件鉴别器消除特定于事件的依赖关系，最后进行虚假信息检测。

MAVE[16]模型：MVAE 旨在学习文本和视觉模式之间的共享表示，以检测虚假信息。利用变分自动编码器对输入数据进行重构得到共享表示，并利用二值分类器对虚假信息进行检测。

BDANN[22]模型：采用 BERT 提取文本特征与视觉特征进行拼接融合的多模态虚假信息检测最新模型。

3.4 对比实验

为了评估模型的性能，我们在微博和推特数据集进行了实验。对比实验包括了本文提出的 MFNN 和 att-MFNN 模型和三种基线模型。对比实验结果如表 2 所示。（表 2 中 P 表示精确度 Precision，R 表示召回率 Recall，Accuracy 表示准确度）

在微博数据集中，MFNN 模型已经是优于所有基线模型，比基线模型中性能最好的 BDANN，准确率还要高 2.79%，且其他指标也普遍有所提升。加入了注意力机制后，总体性能大幅度提高，准确度也提升了将近 2.33%（共提升 5.12%），准确率可以达到 89.22%，所有指标优于所有基线模型。

在 Twitter 数据集中 att-MFNN 与基线模型中性能最好的 BDANN 模型相比准确度提高了 4.51%，达到了 87.51%。除了 MFNN 模型在 Real News 召回率达到 93%，att-MFNN 模型则是在准确度、总 F1 值、总召回率等指标中均有提升，并优于所有基线模型。

表 2 微博和推特数据集上 att-MFNN 和基线模型对比实验结果

Table 2 Comparison of experimental results between att-MFNN and baseline model on Weibo and Twitter datasets

Dateset	Method	Accuracy	Real News			Fake news		
			P	R	F1	P	R	F1
Weibo	EANN	0.8163	0.82	0.82	0.82	0.81	0.80	0.81
	MVAE	0.8262	0.80	0.86	0.83	0.81	0.76	0.81
	BDANN	0.8410	0.87	0.79	0.83	0.82	0.89	0.85
	MFNN	0.8789	0.88	0.87	0.86	0.88	0.88	0.87
	att-MFNN	**0.8922**	**0.90**	**0.88**	**0.89**	**0.89**	**0.90**	**0.90**
Twitter	EANN	0.7132	0.77	0.86	0.81	0.63	0.48	0.54
	MVAE	0.7431	0.69	0.78	0.73	0.80	**0.72**	0.76
	BDANN	0.8300	0.83	0.88	0.82	0.81	0.63	0.71
	MFNN	0.8661	0.88	**0.93**	0.90	0.83	0.74	0.78
	att-MFNN	**0.8751**	**0.89**	0.92	**0.91**	**0.83**	**0.77**	**0.80**

3.5 消融实验

为了更清晰的了解模型中各个特征的作用，在微博和推特数据集中进行了消融实验。结果如表 3 所示：

表 3 微博和推特数据集上 att-MFNN 消融实验结果

Table 3 Results of att-MFNN ablation experiments on Weibo and Twitter data sets

Dateset	Method	Accuracy	Real News			Fake news		
			P	R	F1	P	R	F1
Weibo	Text only	0.8032	0.81	0.78	0.81	0.80	0.83	0.82
	Vision only	0.6393	0.63	0.54	0.55	0.65	0.72	0.65
	Emotion only	0.6205	0.63	0.52	0.57	0.61	0.72	0.66
	Text w/o	0.6526	0.63	0.68	0.66	0.68	0.62	0.65
	Vision+attention w/o	0.8464	0.86	0.82	0.84	0.84	0.87	0.85
	Vision w/o	0.8642	**0.87**	0.84	0.86	0.86	**0.89**	0.87
	Emotion w/o	0.8410	**0.87**	0.79	0.83	0.82	**0.89**	0.85
	MFNN-event	0.8567	**0.87**	0.83	0.85	0.85	0.88	0.86
	att-MFNN-event	**0.8744**	0.85	**0.89**	**0.87**	**0.90**	0.86	**0.88**
Twitter	Text only	0.7058	0.72	0.62	0.67	0.68	0.51	0.57
	Vision only	0.5961	0.53	0.70	0.60	0.69	0.52	0.59
	Emotion only	0.5262	0.45	0.68	0.55	0.65	0.41	0.50
	Text w/o	0.6675	0.57	0.82	0.67	0.80	0.56	0.66
	Vis+attention w/o	0.7151	0.68	0.77	0.69	0.75	0.51	0.62
	Vision w/o	0.7296	0.68	0.76	0.69	0.78	0.57	0.65
	Emotion w/o	0.8043	0.80	0.83	0.82	0.79	0.63	0.71
	MFNN-event	0.8555	0.86	0.88	0.84	0.83	**0.73**	0.76
	att-MFNN-event	**0.8642**	**0.88**	**0.89**	**0.87**	**0.84**	0.72	**0.77**

本次消融实验中“only”表示只含一种模态的模型，“w/o”表示从 att-MFNN 模型去除相应模态的模型，“-event”表示从模型中去除事件分类器。

在微博数据集中，单独基于文本特征的模型准确度为 80.32%，而从 MFNN 模型去除文本特征后模型准确下降到 65.26%，直观的说明了文本特征对于模型性能的影响相比较其他两个特征更大。与基于单模态的模型相比，基于两种特征的多模态模型都有了提升。视觉特征和情感特征单模态情况下，准确率为 63.93%、62.05%，而将这两个特征融合后准确率提高到了 65.26%。情感特征和视觉特征分别与文本特征融合后性能大幅度提高到 84.64%和 84.10%。我们通过注意力机制将情感特征与文本特征进行融合后与基于拼接的模型相比准确度提高 1.78%，达到 86.42%，已经优于四种基线模型。MFNN 模型和 att-MFNN 模型去除事件分类器后准确率有一定的降低为 85.67%，87.44%。说明事件分类器依然能够带来性能上的提升。

在推特数据集中，单模态中基于文本特征的模型的准确度为 70.58%高于基于情感特征（准确度：52.62%）和基于视觉特征（准确度：59.61%）的模型。文本、情感、视觉三种特征不同组合中，文本和视觉特征融合后的性能优于其他组合方式。文本和情感特征通过注意力机制融合的模型准确度高于采用拼接融合的模型 1.44%。去除事件分类器后，MFNN 模型和 att-MFNN 模型准确率分别下降 1.06%和 1.49%。与微博数据集相比较，推特数据集中单个事件关联的文本较多，不利于模型学习不同事件之间的共同特征，这也可能是在推特数据集中文本特征的准确度较低以及事件分类器对模型性能提升有限的原因。

为了更加直观了解情感特征和注意力机制在模型中发挥的作用，我们根据表 3 中的实验结果绘制出图 3。图 3 中显示了模型的准确度方面的结果，其中“w/o Emotion”表示不含情感特征，“w/ Emotion”表示含情感特征，“w/ attention”表示应用注意力机制融合文本和情感特征；“Text”表示初始状态下只含文本特征的模型，“Vision”表示初始状态下只含视觉特征模型。

(a) Weibo 数据集　　(b) Twitter 数据集

图 3 有无情感特征和注意力机制的模型性能比较

Figure.3 Comparison of model performance with or without emotional features and attention

图 3 所示，在两种数据集中，单模态的文本特征模型，视觉模型以及多模态模型加入情感特征后至少有 1%的提升，最高提升能达到 4.3%。微博数据集中情感特征带来的性能提升高于视觉特征。在论文中提到微博数据集中图片含有较多的噪声，从而影响了视觉特征的效果有关。现实生活中数据中噪声不可避免，因此单一模态或者仅靠文本和视觉特征在这种情况下，性能会遭遇瓶颈甚至下降。利用文本所含的丰富情感，结合文本和视觉特征，学习帖子更全面的特征，会带来性能的提升。

情感特征和文本特征的融合方式的不同对模型性能影响也不同。采用拼接的方式 Text (w/emotion)模型和 MFNN 模型准确度微博上分别为 84.64%、87.89%，推特上分别为 71.51%，

86.61%。Text (w/emotion)模型和 MFNN 模型在采用注意力机制融合后与采用拼接特征相比准确度分别提升 1.78%、1.33%（微博数据集），1.45%、0.90%（推特数据集）。其余指标也都有显著提升。

根据以上实验结果和分析得出结论：利用情感特征和基于注意力机制的融合方式在虚假信息检测任务中带来实质性提升。

4 结语

本文中我们主要研究多模态虚假信息检测。通过观察发现，社交媒体中的虚假内容除了图文结合，还伴随着强烈的情感煽动。因此，我们提出了文本，情感，视觉特征基于注意力机制融合的虚假信息检测模型—att-MFNN。在微博和推特两大数据集进行对比实验和消融试验表明，att-MFNN 的模型有效且优于现有的模型。未来的工作聚焦于更多特征的融合过程，例如传播特征，社交特征，评论的文本及情感特征如何融合才能使不同模态间信息互补，发挥多模态优势。

参考文献

[1]Shu K, Sliva A, Wang S, et al. Fake news detection on social media: A data mining perspective[J]. ACM SIGKDD explorations newsletter, 2017, 19(1): 22-36

[2]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. 2017: 795-816.

[3]Zaremba W, Sutskever I, Vinyals O. Recurrent neural network regularization[J]. arXiv preprint arXiv:1409.2329, 2014.

[4]Wang Y Q, Ma F L, Jin Z W, et al. Eann: Event adversarial neural networks for multi-modal fake news detection[C]//Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining. 2018: 849-857.

[5]Zhang X Y, Cao J, Li X R, et al. Mining Dual Emotion for Fake News Detection[C]//Proceedings of the Web Conference 2021. 2021: 3465-3476.

[6]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. 2020, 34(01): 549-556..

[7]Przybyla P. Capturing the Style of Fake News[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(01): 490-497.

[8]Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[9]Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.

[10]Ma J, Gao W, Mitra P, et al. Detecting rumors from microblogs with recurrent neural networks[J]. 2016:3818-3824

[11]Chen T, Li X, Yin H Z, et al. Call attention to rumors: Deep attention based recurrent neural networks for early rumor detection[C]//Pacific-Asia conference on knowledge discovery and data mining. Springer, Cham, 2018: 40-52.

[12]Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. arXiv preprint arXiv:1706.03762, 2017.

[13]Yoon Kim. 2014. Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882 (2014).

[14]Ajao O, Bhowmik D, Zargari S. Sentiment aware fake news detection on online social networks[C]//ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019:

2507-2511.
[15]Guo C, Cao J, Zhang X Y, et al. Exploiting emotions for fake news detection on social media[J]. arXiv preprint arXiv:1903.01728, 2019.
[16]Khattar D, Goud J S, Gupta M, et al. Mvae: Multimodal variational autoencoder for fake news detection[C]//The World Wide Web Conference. 2019: 2915-2921.
[17]Xu N, Mao W J, Chen G D. Multi-interactive memory network for aspect based multimodal sentiment analysis[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2019, 33(01): 371-378.
[18]Goodfellow I J, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks[J]. arXiv preprint arXiv:1406.2661, 2014.
[19]Li X, Wang C, Tan J W, et al. Adversarial Multimodal Representation Learning for Click-Through Rate Prediction[C]//Proceedings of The Web Conference 2020. 2020: 827-836.
[20]Jin Z W, Cao J, Jiang Y G, et al. News credibility evaluation on microblog with a hierarchical propagation model[C]//2014 IEEE International Conference on Data Mining. IEEE, 2014: 230-239.
[21]Boididou C, Andreadou K, Papadopoulos S, et al. Verifying Multimedia Use at MediaEval 2015[J]. MediaEval, 2015, 3(3): 7.
[22]Zhang T, Wang D, Chen H H, et al. BDANN: BERT-Based Domain Adaptation Neural Network for Multi-Modal Fake News Detection[C]//2020 International Joint Conference on Neural Networks (IJCNN). IEEE, 2020: 1-8.

Attention based Multi-Feature Fusion Neural Network for Fake News Detection

DILIXIATI •Abudureyimu[1,2,3]、MA Bo[1,2,3 *]、YANG Yating[1,2,3],WANG Lei[1,2,3]
(1.The Xinjiang Technical Institute of Physics & Chemistry,Chinese Academy of Sciences,Urumqi 830011,China; 2. University of Chinese Academy of Sciences,Beijing 100049,China; 3.Xinjiang Laboratory of Minority Speech and Language Information Processing,Urumqi 830011,China)

Abstract: In the task of identifying fake news, it is difficult to identify the false content based on the monomodal model facing the combination of graphics and texts. How to make full use of multi-modal information to identify fake news in emergencies with high accuracy and quick speed is a challenge. Therefore,this paper proposes an Attention based multi-feature fusion neural network(att-MFNN) for fake news detection.att-MFNN fristly fuses text features and emotional features based on the attention mechanism,then combines with visual features to form multi-features. finally sends multi-features to the fake news detector and event classifier.we conducted comparative experiments on Weibo and Twitter datasets about multimodal fake news detection.Experiment results show that,att-MFNN achieves 89.22% and 87.51% accuracy in the weibo and twitter datasets, and the F1, Precision and Recall indicators are better than the baseline model.

Keywords: Fake news detection;Multi-feature fusion;Attention ; Emotion extraction;