文章编号：1003-0077(2022)09-0129-10

基于双重情感感知的可解释谣言检测

葛晓义[1,2]，张明书[1,2]，魏 彬[1,2]，刘 佳[1,2]

(1. 武警工程大学 密码工程学院，陕西 西安 710086；
2.武警工程大学 网络与信息安全武警部队重点实验室，陕西 西安 710086)

摘 要：社交媒体时代给我们带来便利的同时也造成了谣言泛滥，因此通过人工智能技术进行谣言检测具有重要的研究价值。尽管基于深度学习的谣言检测取得了很好的效果，但其大多数是根据潜在特征进行谣言检测的，无法学习情感与语义之间的相关性，同时忽视了从情感角度提供解释。为解决上述问题，该文提出一种基于双重情感感知的可解释谣言检测模型，旨在利用协同注意力机制分别学习谣言语义与用户评论情感，以及谣言情感与用户评论情感的相关性进行谣言检测，并通过协同注意力权重从情感角度提供合理的解释。在公开的 Twitter15、Twitter16 和 Weibo20 数据集上的实验结果表明，该文提出的模型与对比模型相比，在准确率上分别提高了 3.9%，3.9%和 4.4%，且具有合理的可解释性。

关键词：谣言检测；协同注意力；情感特征；可解释性；语义特征

中图分类号：TP391 **文献标识码**：A

Dual Emotion-aware Method for Interpretable Rumor Detection

GE Xiaoyi[1,2], ZHANG Mingshu[1,2], WEI Bin[1,2], LIU Jia[1,2]

(1. College of Cryptographic Engineering, Engineering University of PAP, Xi'an, Shaanxi 710086, China;
2.Key Laboratory for Network and Information Security of PAP,
Engineering University of PAP, Xi'an, Shaanxi 710086, China)

Abstract: The identification of rumors is of substantial significance research value. Current deep learning-based solution brings excellent results, but fails in capturing the relationship between emotion and semantics or providing emotional explanations. This paper proposes a dual emotion-aware method for interpretable rumor detection, aiming to provide a reasonable explanation from an emotional point of view via co-attention weights. Compared with contrast model, the accuracy is increased by 3.9%, 3.3% and 4.4% on the public Twitter15, Twitter16, and Weibo20 datasets.

Keywords: rumor detection; co-attention; emotion feature; interpretable; semantics feature

0 引言

社交媒体的快捷性和便利性等优点给工作、生活和学习带来了巨大的便利，为用户发布、分享和获取各种信息提供了便捷的渠道。目前社交媒体已成为各国发布外交政策和相关评论的重要平台，也演变成网络认知战的主战场。然而不可忽视的是，社交媒体的谣言泛滥，严重影响了网络的良性发展，甚至影响着社会、经济和文化的发展。新冠肺炎疫情防控期间，博人眼球的虚假消息，对疫情防控造成了一定干扰。有效检测谣言有利于净化网络空间和维护社会稳定，具有重要的现实意义[1]，为了遏制谣言传播，消除谣言带来的影响，越来越多的学者致力于谣言检测任务[2]。

情感分析作为文本分析中确定文本表达情感极性和强度的部分，常被用于谣言检测任务中。Wu[3]等人考虑到谣言和用户评论之间存在情感关联和语义冲突，提出了自适应交互融合网络来实现特征之间的交叉交互融合，从而捕获帖子和评论之间的相似语义和冲突语义。Guo[4]等人分别提取谣言和用户评论的语义及情感特征进行谣言检测，取得较好

收稿日期：2022-03-21 定稿日期：2022-05-26
基金项目：国家社会科学基金(20BXW101，18XWW015)；武警工程大学“优秀研究生培养计划”课题

效果。Zhang[5]等人通过情感字典获取谣言和用户评论的情感表示来探究二者之间的情感差，将情感特征作为增强特征进行谣言检测。

然而上述方法仍有一定的局限性。首先，没有考虑谣言和用户评论的情感相关性，以及谣言语义和用户评论情感的相关性。用户评论往往是较短的句子，导致语义特征不够丰富，而用户评论中蕴含着对谣言明确的态度，情感更加丰富，因此用户评论的情感倾向更能反映检测内容的真假[6]。其次，没有从局部角度获取谣言和用户评论的情感特征。社交媒体中谣言和用户评论的句子往往较短，情感特征通常体现在个别情感色彩丰富的词汇上，因此获取局部情感特征更能表达情感倾向[7]。最后，在已有的可解释谣言检测模型中，仅利用谣言文本和用户评论[8]、转发用户序列和用户信息[9]等提供合理解释，忽视了从情感角度提供合理的解释。

针对现有研究的不足，本文提出一种基于双重情感感知的可解释谣言检测模型。为了从全局角度探究谣言语义和用户评论的相关性，首先，利用双向门循环单元(Bidirectional gate recurrent unit，Bi-GRU)和注意力(Attention)获取谣言语义特征和用户评论情感特征；其次，通过Co-Attention获取谣言语义特征与用户评论情感特征的相关性，以筛选与谣言语义相关的用户评论情感特征并进行融合，利用协同注意力(Co-Attention)权重提供解释。为了从局部角度探究谣言和用户评论的情感相关性，首先，通过卷积神经网络(Convolutional Neural Network，CNN)提取谣言和用户评论的情感特征，其次，通过Co-Attention学习谣言与用户评论的情感相关性，旨在获取与谣言情感相关的用户评论情感特征进行融合，并利用Co-Attention权重提供解释。本文的贡献如下：

(1) 提出一种新的可解释谣言检测模型，分别从谣言语义和用户评论情感以及谣言情感和用户评论情感出发进行谣言检测。

(2) 通过Co-attention机制学习谣言语义与评论情感的相关性，以及谣言情感与评论情感的相关性，通过Co-attention权重从情感角度产生合理的解释。

(3) 在真实数据集上的实验表明，与先进的模型相比，具有较好的检测效果和合理的解释性，实验代码开源在码云①。

1 相关工作

1.1 谣言检测

谣言检测根据特征通常分为基于谣言内容、基于社会上下文及基于混合特征的方法。谣言内容可以分为文本和视觉两个方面，文本方面指根据谣言的语言风格[10]、写作风格[11]和情感[12-14]等提取文本特征和情感特征，例如刘[13]等人提出使用卷积神经网络提取文本特征进行的谣言检测模型，Cui[14]等人通过实验表明，情感分析对系统性能的影响最大。视觉特征则是从视频或图片中提取特征[15]。基于社会上下文的检测方法一般可以分为基于用户和基于网络。前者是根据谣言发布者和转发用户的特点进行建模[11,16]，特征主要包括用户性别、粉丝数、用户配置；后者通过社交网络中的转发或关注结构的特征进行谣言检测，如：Bian[16]等人利用双向图神经网络模型学习嵌入谣言传播。基于混合特征的方法是融合多模态或者多重特征进行谣言检测，如Wu[17]等人分别学习文本和图像的表示，利用模态的上下文注意力网络融合模态内(Intra-modality)和模态间(Inter-modality)的关系进行谣言检测；Zhang[5]等人在通过情感词典获取发布者情感、用户评论情感和情感代沟，作为假新闻检测器的补充特征，取得了很好的效果。

近年来的研究趋向于可解释谣言检测[18]，主要通过提取用户信息[19]、转发序列[9]、新闻内容和用户评论[8]等来提供解释。Lu[19]等人利用两次协同注意力机制通过突出可疑的转发者以及他们关注的话语来生成解释。Jin[20]等人通过对微妙线索的细粒度建模来提高检测的，准确性和可解释性。

1.2 深度学习可解释性

虽然深度学习模型在越来越多领域得到应用，却常因不具备透明度、可信度以及不符合伦理道德标准等遭诟病，因此对深度学习可解释性的需求也越来越高[21]。近年来，深度学习可解释性模型开始在越来越多的领域应用，如网络安全[22]、推荐系统[23]、医疗[24]、社交网络[19]等。深度学习可解释性模型一般是指模型决策结果以可理解的方式呈现，能够帮助理解复杂模型的内部工作机制以及模型做

① https://gitee.com/wj_gxy/dual-emotion_aware

出特定决策的原因[25]。可解释性一般分为内在可解释性和事后可解释性[26]。内在可解释性[8]是通过构建将可解释性直接纳入其结构的自解释模型来实现的，而事后可解释性[27]需要创建第二个模型来为现有模型提供解释。在谣言检测中应当更倾向于内在可解释性，在检测结果中就存在着自解释的信息。

2　本文模型

本文提出的可解释谣言检测模型分别从全局和局部对谣言和用户评论在情感上的关系获取特征，进行谣言检测。模型如图 1 所示，主要由四部分构成：①嵌入层，利用词嵌入向量和情感嵌入向量对谣言和用户评论进行向量表示；②特征提取层，通过 Bi-GRU 与 Attention 获取谣言的语义表示和用户评论的情感表示，通过 CNN 获取谣言和用户评论的情感表示；③双重情感感知层，分别依据谣言语义特征与用户评论情感特征，以及谣言文本情感特征与用户评论情感特征，通过 Co-attention 获取谣言语义特征与用户评论情感之间的相关性和谣言情感特征与用户评论情感特征的相关性；④预测层，将 Co-attention 获取的特征进行拼接，通过 Softmax 分类。

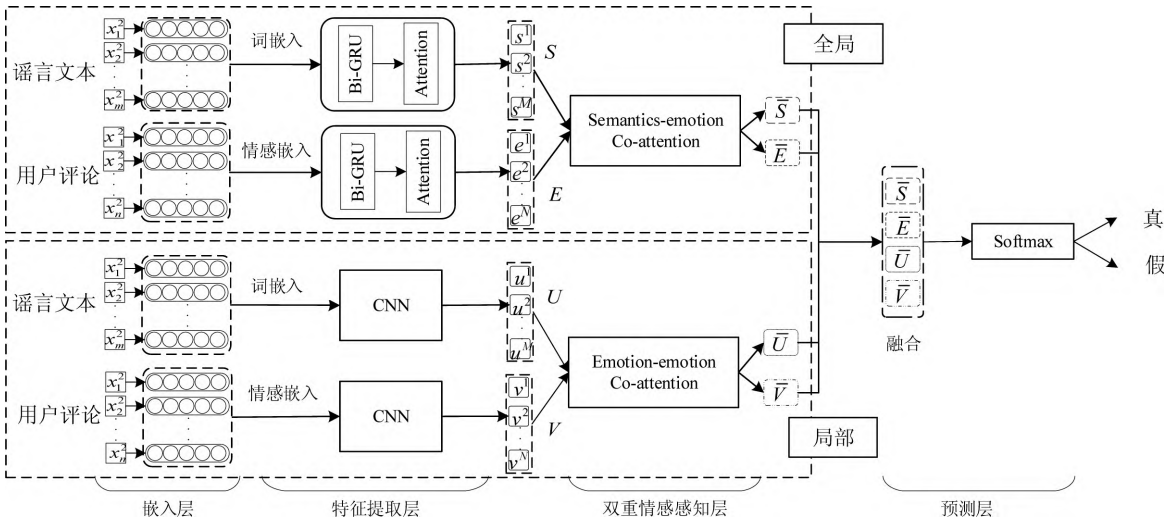

图 1　可解释谣言检测模型框架

2.1　嵌入层

在进行特征提取前，首先对每个词进行词向量嵌入和情感向量嵌入。英文词向量嵌入采用 Robyn[28]等人预先训练好的 Numberbatch 词向量，在词向量相似性上优于 Word2Vec[29]和 GloVe[30]，中文词向量采用预训练的微博词向量①。

受情感建模工作文献[31]的启发，本文将情感元素融入到预先训练好的词向量中获取情感嵌入向量。该方法基于 NRC 情感词典创建了两组约束，一组用于与情绪（例如绑架，悲伤）具有积极关系的单词，另一组用于跟踪与该情绪相反的每个单词（绑架，喜悦），喜悦是悲伤的反面。

通过增加一个新的训练阶段，使用情感词汇和基本情绪词汇将情感信息拟合到预训练的 Numberbatch 词向量中获取情感向量。在训练情感嵌入时采用的正面与反面的约束字典以及中英文向量大小统计如表 1 所示。

表 1　字典统计

统计项	中文	英文
正面字典	22 332	16 502
反面字典	22 012	16 508
向量长度	189 600	516 782

2.2　特征提取层

经过数据预处理后，一条谣言由 M 个句子组成，其中每个句子 s 由 m 个词组成 $s^i=[x_1^i, x_2^i, \cdots, x_m^i]$，一条谣言对应的用户评论由 N 个句子组成，其中每个句子 e 由 n 个词组成 $e^j=[x_1^j, x_2^j, \cdots,$

① https://github.com/Embedding/Chinese-Word-Vectors

$x_n^j]$。经过预训练的词向量和情感向量表示后用于提取语义特征和情感特征。

2.2.1 语义特征提取

理论上RNN能够捕获长期依赖，但在实践中，旧的记忆会随着序列变长而消失。为了捕获RNN的长期依赖关系，使用GRU来确保更持久的内存。虽然词中都包含上下文信息和整个句子的信息，但是句子中每个词的重要性不同。具体如图2所示。

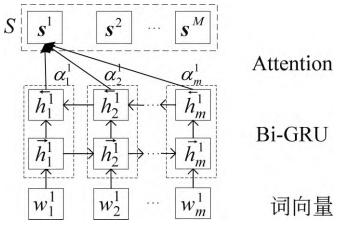

图2 语义特征提取过程

谣言中的词汇往往与上下文具有关联性，具有较强的双向语义依赖，因此逆序处理十分必要，采用Bi-GRU从词的两个方向建模获取谣言语义特征。词嵌入的向量 $s^i=[w_1^i,w_2^i,\cdots,w_m^i]$ 通过Bi-GRU可以得到，如式(1)所示。

$$\begin{aligned}\overrightarrow{h}_t^i&=\overrightarrow{\mathrm{GRU}}(w_t^i,\overrightarrow{h}_{t-1}^i),\quad t\in\{1,\cdots,m\}\\ \overleftarrow{h}_t^i&=\overleftarrow{\mathrm{GRU}}(w_t^i,\overleftarrow{h}_{t-1}^i),\quad t\in\{m,\cdots,1\}\end{aligned}\tag{1}$$

通过连接前向隐藏状态 $\overrightarrow{h}_t^i$ 和后向隐藏状态 $\overleftarrow{h}_t^i$，获取词的特征表示 $\boldsymbol{h}_t^i=[\overrightarrow{h}_t^i,\overleftarrow{h}_t^i]$。通过注意力机制学习词的重要性来获得句子向量 $\boldsymbol{s}\in\mathbb{R}^{2\mathrm{d}\times\mathrm{m}}$ 如式(2)所示。

$$\boldsymbol{s}^i=\sum_{t=1}^{m}\alpha_t^i\boldsymbol{h}_t^i\tag{2}$$

其中 α_t^i 衡量 t^{th} 单词对新闻内容 s 的重要性，α_t^i 计算如式(3)所示。

$$\begin{aligned}u_t^i&=\tanh(\boldsymbol{W}_w\boldsymbol{h}_t^i+\boldsymbol{b}_w)\\ \alpha_t^i&=\frac{\exp(u_t^iu_w^{\mathrm{T}})}{\sum_{k=1}^{m}\exp(u_k^iu_w^{\mathrm{T}})}\end{aligned}\tag{3}$$

其中，u_t^i 是通过完全嵌入层从隐藏状态 $\boldsymbol{h}_t^i$ 获得的，W_w 和 b_w 是可训练的参数，$\boldsymbol{u}_w$ 为权重矩阵。

2.2.2 情感特征提取

在模型中共提取三部分情感特征，图中一部分用户评论的情感特征与谣言语义特征提取方法相同，采用Bi-GRU与Attention的方法获得用户评论的情感特征 $\boldsymbol{E}=[e^1,e^2,\cdots,e^N]$。

其中谣言情感特征与另一部分用户评论情感特征采用CNN模型提取，CNN模型能够较好地提取局部特征，并且模型训练的效率高，因此选用一维卷积神经网络提取谣言情感特征 $U=[u^1,u^2,\cdots,u^M]$ 与用户评论情感特征 $\boldsymbol{V}=[v^1,v^2,\cdots,v^N]$，具体方法如图3所示。

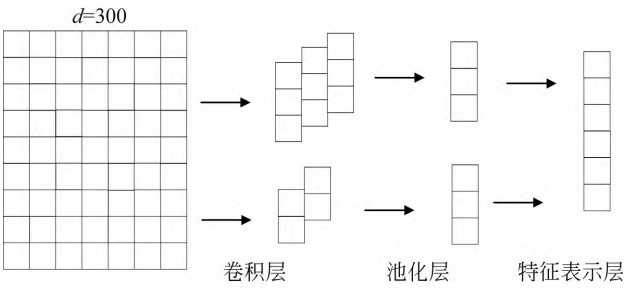

图3 基于CNN的特征表示

对用户评论中某一行评论情感嵌入后的向量 $e^j=[w_1^j,w_2^j,\cdots,w_n^j]\in\mathbb{R}^{n\times d}$ 进行卷积操作：

$$h^j=\mathrm{ReLU}(\boldsymbol{W}\cdot\boldsymbol{w}_{t:t+\lambda-1}^j+\boldsymbol{b})\tag{4}$$

其中，$W\in\mathbb{R}^{\lambda\times d}$ 是可学习的参数矩阵，b 是偏置项，ReLU是激活函数。对卷积得到的 h^j 进行最大池化可得到每一句评论的情感特征：

$$\tilde{h}^j=\max\ \mathrm{pooling}(h^j)\tag{5}$$

CNN层使用两个过滤器($\lambda\in\{2,3\}$)来获取多个特征，将不同的输出连接起来，形成 v^j 作为用户评论的单个表示。最后可以得到用户评论中每个评论的情感特征，形成用户评论的情感矩阵 $\boldsymbol{V}=[v^1,v^2,\cdots,v^N]$。

2.3 双重情感感知层

用户评论中往往包含着大量的与谣言相关的信息，但它们信息量较小，噪声较大。因此利用谣言自身进行谣言检测和解释谣言真假是薄弱的，而评论中情感丰富，与语义特征相比，情感特征更加突出，更有利于谣言检测，并能通过情感反映谣言真假的原因。通过协同注意力机制学习评论情感与谣言的相关性，利用情感的注意力权重和谣言中的词汇来进行谣言检测和谣言解释，具体过程如图4所示。

谣言表示为：$\boldsymbol{S}=[s^1,s^2,\cdots,s^M]$，评论情感特征表示为：$\boldsymbol{E}=[e^1,e^2,\cdots,e^N]$。

首先计算相似矩阵 $\boldsymbol{F}=tanh(\mathrm{EW_{se}S})$，其中，$\boldsymbol{F}\in\mathbb{R}^{\mathrm{N}\times\mathrm{M}}$，$\boldsymbol{W}_{\mathrm{se}}\in\mathbb{R}^{2\mathrm{d}\times2\mathrm{d}}$，是可学习的参数矩阵。将相似矩阵作为一个特征，则可以学习谣言语义特征和用户评论情感特征的协同表示。

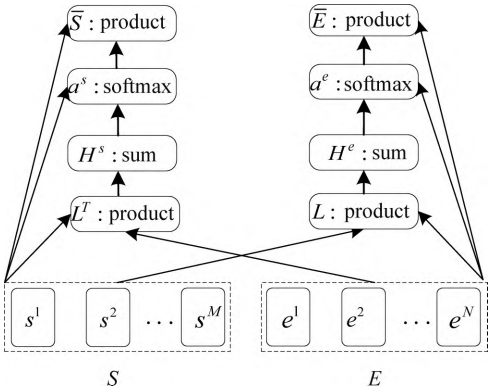

图 4　协同表示过程

$$H^s=\tanh(\boldsymbol{W}_s\boldsymbol{S}+(\boldsymbol{W}_e\boldsymbol{E})F)$$
$$H^e=\tanh(\boldsymbol{W}_e\boldsymbol{E}+(\boldsymbol{W}_s\boldsymbol{S})\boldsymbol{F}^T) \tag{6}$$

$\boldsymbol{W}_s$、$\boldsymbol{W}_e\in\mathbb{R}^{k\times 2d}$ 为可学习的参数矩阵，可以学习谣言文本和评论情感特征的注意力权重：

$$a^s=\mathrm{softmax}(\boldsymbol{W}_{hs}^{\mathrm{T}}H^s)$$
$$a^e=\mathrm{softmax}(\boldsymbol{W}_{he}^{\mathrm{T}}H^e) \tag{7}$$

其中，$a^s\in\mathbb{R}^{1\times M}$，$a^e\in\mathbb{R}^{1\times N}$ 分别是谣言中每个词和评论的情感特征中每个评论的注意权重。$\boldsymbol{W}_{\mathrm{hs}}$、$\boldsymbol{W}_{\mathrm{he}}$ 是可训练权重。最终，通过加权协同表示。

$$\bar{\boldsymbol{S}}=\sum_{i=1}^{M}a_i^s\boldsymbol{s}^i,\quad \bar{\boldsymbol{E}}=\sum_{j=1}^{N}a_i^e\boldsymbol{e}^j \tag{8}$$

利用协同注意力机制对谣言情感特征和用户评论情感特征计算相似矩阵，获取对应的权重分别生成协同表示 $\bar{\boldsymbol{U}}$、$\bar{\boldsymbol{V}}$。

2.4　预测层

将提取到的特征通过全连接层输出，最后通过 softmax 函数来获得分类的结果，如式(9)所示。

$$\dot{y}=\mathrm{softmax}([\bar{S},\bar{E},\bar{U},\bar{V}]W_f+b) \tag{9}$$

其中，$\dot{y}$ 为 softmax 函数计算的概率值，W_f 为权重矩阵，b 为偏置项。

3　实验结果与分析

3.1　实验数据

Twitter15 和 Twitter16[32] 数据集选择“真”和“假”标签数据，数据集中都包含谣言内容、用户评论和相应的转发用户序列等信息。Weibo20 由 Zhang[5] 等人在 Weibo16[33] 的基础上通过聚类算法去重，并增加了 2014 年 4 月至 2018 年 11 月被微博社区管理中心认定的虚假信息，形成新的数据集。Weibo20 数据集包含谣言内容、用户评论和标签三部分信息。数据集的统计如表 2 所示。

表 2　数据集统计

	Twitter15	Twitter16	Weibo20
“真”	372	205	3 161
“假”	370	207	3 201
总数	742	412	6 362
评论	9 659	4 122	1 983 440
评论平均条数	13	10	312

3.2　实验设置

为突出本文方法的先进性，在上述数据集进行实验，将实验结果与先进模型进行比对和分析。

- **RNN**[33]：一种基于 RNN 的方法，将社交上下文信息建模为可变长度的时间序列，用于学习谣言的连续表示。
- **text-CNN**[34]：一种基于卷积神经网络的文本分类模型，利用多个卷积滤波器来捕获不同粒度的文本特征。
- **HAN**[35]：一种基于层次注意力网络的文档分类模型，利用词级注意力和句子级注意力来学习新闻内容表示。
- **dEFEND**[8]：一种基于协同注意力的假新闻检测模型，学习新闻内容和用户评论之间的相关性，并利用 Co-attention 的权重从谣言文本和用户评论给出解释。
- **PLAN**[9]：一种关注用户交互的可解释谣言检测模型，将谣言及转发评论作为 Transformer 的输入，并利用延迟时间嵌入代替的位置嵌入进行谣言检测，通过 Attention 为帖子和标签提供解释。
- **Dual Emotion**[5]：一种基于双重情感特征的假新闻检测模型，通过学习谣言情感特征、评论情感特征及情感特征差作为假新闻检测器的补充特征。

在 Twitter15 与 Twitter16 数据集中，dEFEND 模型中谣言文本句子个数为 1，长度为 32，评论句子分别选取 12 和 9 条，长度为 32；为了对比公平，Dual emotion 模型利用 Bi-GRU 提取文本特征，分别选择 12 和 9 条评论提取情感特征；本文所提模型，谣言文本个数为 1，长度为 32，评论个数分别为 12 和 9 条，长度为 32，其他模型参照原文设置。

在 Weibo20 数据集中，dEFEND 模型中谣言文本句子个数为 1，长度为 64，评论句子选取 100 条，长度为 32；由于数据集中不存在延迟时间，因此

PLAN模型不再对比；为了对比公平，Dual emotion模型利用Bi-GRU提取文本特征，选择100条评论提取情感特征；本文所提模型的谣言文本个数为1，长度为64，评论个数为100条，长度为32，其他模型参照原文设置。

3.3 实验结果

数据集按照6:2:2的比例划分为训练集、验证集、测试集，每一个数据集中的样本比例为1∶1。实验使用Adam更新参数，学习率分别为0.001与0.005，L2正则化系数为0.001。词向量与情感词向量维度均设置为300。设置常用的评价指标为：正确率Accuracy、准确率Precision、召回率Recall以及F_1。在Twitter15、Twitter16和Weibo20上的实验结果如表3～表5所示，表中加黑数据为最优结果，下划线数据为次优结果。

表3 Twitter15上不同模型的结果对比

方法	Accuracy	Precision	Recall	F_1
RNN	0.722	0.726	0.718	0.723
Text-CNN	0.756	0.732	0.731	0.730
HAN	0.820	0.821	0.821	0.823
dEFEND	0.842	0.845	0.843	0.845
PLAN	0.845	0.869	0.863	0.845
Dual Emotion	0.852	0.848	0.851	0.851
ours	**0.891**	**0.893**	**0.892**	**0.892**

表4 Twitter16上不同模型的结果对比

方法	Accuracy	Precision	Recall	F_1
RNN	0.662	0.655	0.655	0.658
Text-CNN	0.676	0.678	0.681	0.677
HAN	0.723	0.712	0.712	0.716
dEFEND	0.774	0.772	0.774	0.774
PLAN	0.824	0.819	0.829	0.827
Dual Emotion	0.812	0.821	0.817	0.812
ours	**0.857**	**0.859**	**0.857**	**0.857**

表5 Weibo20上不同模型的结果对比

方法	Accuracy	Precision	Recall	F_1
RNN	0.809	0.809	0.810	0.809
Text-CNN	0.811	0.811	0.811	0.811
HAN	0.823	0.812	0.812	0.816
dEFEND	0.843	0.843	0.840	0.843
Dual Emotion	0.851	0.852	0.853	0.852
ours	**0.895**	**0.895**	**0.895**	**0.895**

从表3、表4与表5可以发现，在Twitter15、Twitter16与Weibo20数据集上，本文所提模型在各个指标上都显著优于其他模型，在Twitter15与Twitter16上的F_1分别提高了约4.1%与3.0%，在准确率上分别提高了约3.9%与3.3%；在Weibo20上F_1提高了约4.3%，在准确率上提高了约4.4%。实验结果证明，本文所提模型不仅优于基于单一特征的方法，更是优于基于混合特征的方法，充分体现了模型的优越性。

在RNN、text-CNN、HAN三种基于单一特征的方法中，HAN模型效果更好，说明在提取语义特征方面，HAN模型更具有优势。dEFEND、PLAN、Dual Emotion三种基于混合特征的方法明显优于基于单一特征的方法，这说明基于混合特征的模型利用不同的方法融合更多的特征往往具有更好的效果。在基于混合特征的模型中，PLAN模型将位置嵌入替换为延迟时间嵌入进行谣言检测，取得优于dEFEND模型仅考虑谣言和用户评论的效果。Dual Emotion模型在不采用Attention的情况下，仅利用谣言语义特征与情感特征融合就取得更好的结果。

该模型与基于混合特征的方法相比，也具有明显的优势。本文所提模型优于dEFEND模型和PLAN模型，说明同样仅利用谣言和用户评论，提取语义特征和情感特征取得更优结果，这表明选取更有效的特征是检测谣言的关键。本文所提模型优于Dual Emotion模型，说明同样以文本与评论情感作为谣言检测器特征，Co-Attention提取特征相关性更具有优势。不可否认的是，Dual Emotion模型提取情感特征更具有全面性，因为当前社交媒体中的用户评论更喜欢用表情或者一些图片表达自己的情感倾向，Dual Emotion模型在获取情感特征时能够针对表情对应的情感特征一并提取，但忽视了谣言文本情感与用户评论情感的内在关系。各个模块对模型性能的影响将在消融实验部分详细阐述。

3.4 消融实验

消融实验主要研究两部分内容：①评论数量对模型性能的影响；②模型各模块对模型性能的影响。

实验一分别选择不同的评论数量提取情感特征，实验结果如图5所示。

可以发现，随着评论数量的增加，模型检测性能有所提高，在推特数据集中，评论数量基数较少，随着

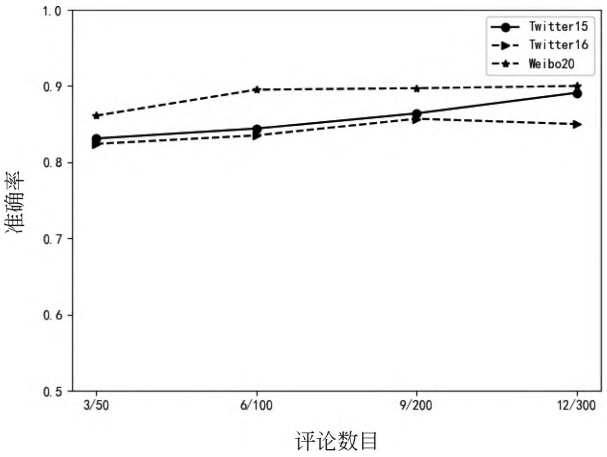

图 5　用户评论数目的影响

评论数目的增加，检测性能随之提高，但是在微博数据集中，评论基数大，随着评论数目的增加，性能提升相对较小，这是因为微博数据集中，相同的评论较多。

实验二是探究模型中各模块对模型性能的影响，“-S”：仅有谣言语义特征和用户评论情感特征的模型；“-E”：仅有谣言情感特征和用户据评论情感特征的模型；“-B”：所有特征由 Bi-GRU 与 Attention 提取；“-C”：表示所有特征均由 CNN 提取的模型；“-A”：本文所提模型；实验结果如图 6 所示。

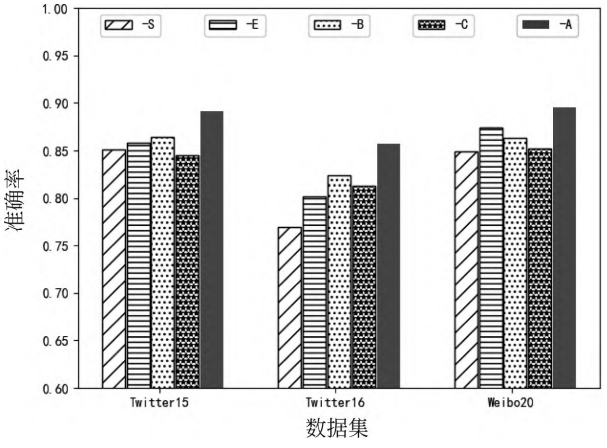

图 6　模型中各模块的影响

实验结果表明，模型中的每一个模块都具有重要的作用，同时特征的提取方式发生改变会严重影响模型的性能。“-B”与“-C”模型虽然都提取了所有特征，但是准确率依然没有本文所提模型高，这是因为“-B”模型提取的是整体特征，“-C”模型提取的是局部特征，而本文模型是整体特征与局部特征的结合，就整体与局部而言，仅提取局部特征的“-C”模型较“-B”模型具有更好的效果。尤其是当去除情感特征提取模块时，模型的准确率会显著下降，同时将“-S”模型与 dEFEND 模型对比可以发现，利用用户评论的情感特征比用户评论语义特征效果更好，这说明在模型中情感特征比语义特征具有更好的作用。

3.5　可解释性分析

在 2.3 节介绍的 Co-attention 权重使模型具有可解释性，根据权重分布的位置，可以揭示谣言检测在谣言文本和用户评论中的证据词，下面分别对模型中两个 Co-attention 权重进行案例分析。

3.5.1　CNN 提取特征的可解释性

通过 CNN 提取谣言情感特征和用户评论情感特征，Co-attention 权重可以揭示谣言文本和用户评论中关注的句子，用户评论中的权重分布如图 7 和表 6 所示。

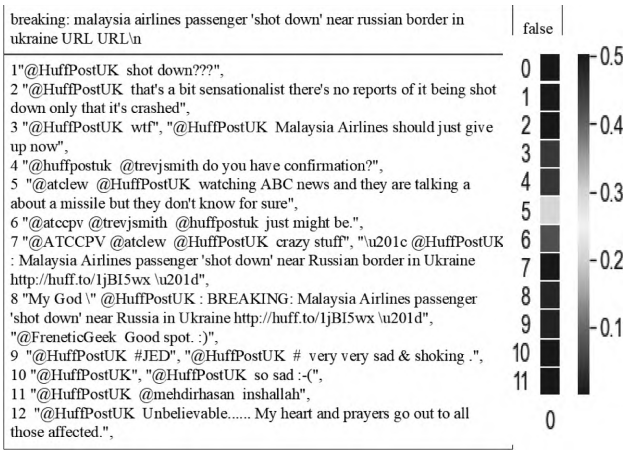

图 7　Twitter 案例分析的权重分布(一)

在图 7 中，发现第 2 句和第 6 句的权重最大，上述两句评论中均包含对谣言文本信息的否定态度，均能对此做出解释。在 NRC 的情感字典中，分别有与情绪有着正反关系的词汇，在第二句中的“sensationalist”与“sadness”对应，“crash”与“anger”对应。

表 6　Weibo 案例分析的权重分布(一)

谣言文本	用户评论
【真敢说[怒]】全国政协委员王平建向党中央建议：取消录取农村小孩上大学资格，不鼓励农村小孩上大学。其认为农村孩子身上有很多坏毛病，农村孩子上大学，城市孩子会被影响，这样城市孩子也变坏！中央必须采取行政手段禁止农村小孩上大学…农村人身上的很多坏毛病，任何一个有文化城里人都忍受不了'	真的假的？
	求辟谣，应该没人傻到这样去激起民怨。
	政协委员说的，真的假的？
	[吃惊][吃惊]假的吧
	其心可诛!!![怒]
	真的？求真相。

通过表 6 中权重选择的前 6 条用户评论可以发现，每一句都直接对谣言文本中的内容表示怀疑和否定的态度。

3.5.2 Bi-GRU 与 Attention 提取特征的可解释性

利用 Bi-GRU 与 Attention 分别提取谣言语义特征和用户评论情感特征，具体如图 8 和表 7 所示。

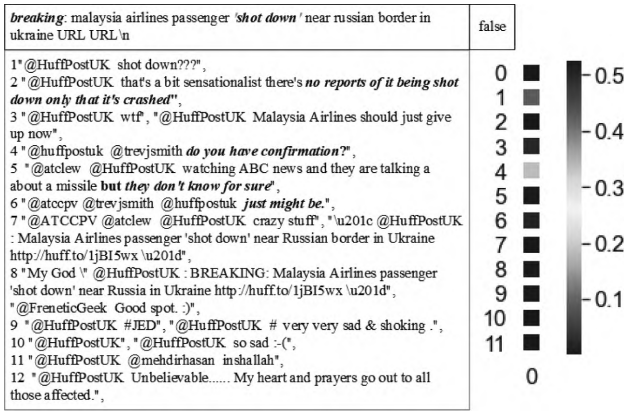

图 8 Twitter 案例分析的权重分布(二)

通过图 8 可以看出，在选择用户评论情感和谣言文本语义时，权重更倾向于与谣言文本语义匹配信息的选择，也从另一角度给出解释。

表 7 Weibo 案例分析的权重分布(二)

谣言文本	用户评论
【真敢说[怒]】全国政协委员王平建向党中央建议：取消录取农村小孩上大学资格，不鼓励农村小孩上大学。其认为农村孩子身上有很多坏毛病，农村孩子上大学，城市孩子会被影响，这样城市孩子也变坏！中央必须采取行政手段禁止农村小孩上大学…农村人身上的很多坏毛病，任何一个有文化城里人都忍受不了	wqnmlgb，你知道中国多少农民吗？煞笔也能当代表。
	难道没有其他政协委员提议把这个人宰了喂狗吗？算个什么东西？还真把自己标榜成城里人儿啊?!
	是政协委员，谁给政协抹黑呢？
	人类中总有那么几个渣滓。
	真不晓得啷个当到政协的
	n 你怎么不去死啊？天、

通过表 7 可以发现，权重所选择的前 6 条用户评论与之前完全不同，更倾向于事实的选择，其中 3 条评论与“政协”相关，其他的也都是契合谣言文本的语义信息。

3.5.3 全局与局部的可解释性

对比两种权重分布可以发现，Weibo20 数据集的用户评论基数大，两种权重选择前 6 条用户评论完全不同，Twitters 数据集的用户评论基数小，虽然前 2 条评论都是第 2 句和第 6 句，但是先后顺序、权重分布也不同。从全局看，基于 Bi-GRU 与 Attention 提取特征的可解释性是从谣言语义特征和用户评论情感特征角度出发，选择用户评论更多的是与谣言文本相关的内容，如上述案列中，用户评论中包含“政协”“农民”等词语；从局部看，基于 CNN 提取特征的可解释方法是从谣言情感特征和用户评论情感特征角度出发，选择用户评论更多的是对谣言文本所持的态度，如上述案列中，用户评论中包含“辟谣”“真的假的”“真相”等词语。

4 总结

考虑到谣言和用户评论中具有强烈的情感倾向，本文提出了一种基于双重情感感知的可解释谣言检测模型，分别提取谣言语义特征，谣言情感特征和用户评论情感特征进行谣言检测，并提供解释。实验结果表明，该模型具有较好的检测结果和较合理的解释性。同时，该模型还可以用于社交媒体上的其他任务，尤其是当前社交媒体上认知战正越演越烈，可以利用该模型进行仇恨语言检测、意识形态检测等任务。在下步工作中，我们将根据谣言的特点进行仇恨语言检测和立场检测的多任务实验，进一步研究社交媒体中认知偏移的规律。

参考文献

[1] 李奥，但志平，董方敏，等. 基于改进生成对抗网络的谣言检测方法[J]. 中文信息学报，2020，34(09)：78-88.

[2] 琚心怡. 基于深层双向 Transformer 编码器的早期谣言检测[J]. 信息通信，2020(05)：17-22.

[3] WU L W，RAO Y. Adaptive interaction fusion networks for fake news detection[C]//Proceedings of the 24th Europear Conference on Artificial Intelligence 2020，2020：2220-2227.

[4] Guo C，Cao J，Zhang X，et al. Exploiting emotions for fake news detection on social media[J]. arXiv：1903.01728，2019.

[5] Zhang X，Cao J，Li X，et al. Mining dual emotion for fake news detection[C]//Proceedings of the Web Conference 2021，2021：3465-3476.

[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(03)：198-204.

[7] Yin W，Kann K，Yu M，et al. Comparative study of CNN and RNN for natural language processing[J]. ArXiv：1702.01923，2017.

[8] Shu K，Cui L，Wang S，et al. dEFEND：Explainable

fake news detection[C]//Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Anchorage AK USA: ACM, 2019: 395-405.

[9] Khoo L M S, Chieu H L, Qian Z, et al. Interpretable rumor detection in microblogs by attending to user interactions[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2020: 8783-8790.

[10] Horne B, Adali S. This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news[C]//Proceedings of the international AAAI Conference on Web and Social Media, 2017: 759-766.

[11] Yang F, Liu Y, Yu X, et al. Automatic detection of rumor on Sina Weibo[C]//Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics -MDS '12. Beijing, China: ACM Press, 2012: 1-7.

[12] **沈瑞琳，潘伟民，彭成，等. 基于多任务学习的微博谣言检测方法[J]. 计算机工程与应用**, 2021, 57(24): 192-197.

[13] **刘政，卫志华，张韧弦. 基于卷积神经网络的谣言检测[J]. 计算机应用**, 2017, 37(11): 3053-3056.

[14] Cui L, Wang S, Lee D. SAME: Sentiment-aware multi-modal embedding for detecting fake news [C]//Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2019: 41-48.

[15] Khattar D, Goud J S, Gupta M, et al. MVAE: Multimodal variational autoencoder for fake news detection [C]//Proceedings of the World Wide Web Conference. San Francisco CA USA: ACM, 2019: 2915-2921.

[16] Bian T, Xiao X, Xu T, et al. Rumor detection on social media with bi-directional graph convolutional networks[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2020: 549-556.

[17] Wu Y, Zhan P, Zhang Y, et al. Multimodal fusion with coattention networks for fake news detection [C]//Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP. Online: Association for Computational Linguistics, 2021: 2560-2569.

[18] Kou Z, Zhang D, Shang L, et al. What and why towards duo explainable fauxtography detection under constrained supervision[J]. IEEE Transactions on Big Data, 2021(01): 1-14.

[19] Lu Y J, Li C T. GCAN: Graph-aware co-attention networks for explainable fake news detection on social media[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online 2020: 505-514.

[20] Jin Y, Wang X, Yang R, et al. Towards fine-grained reasoning for fake news detection[G]. AAAI Technical Track on Knowledge RepresenlntIon and Reasoning, 36(5): 5746-5754.

[21] **成科扬，王宁，师文喜，等. 深度学习可解释性研究进展[J]. 计算机研究与发展**, 2020, 57(06): 1208-1217.

[22] Xu X, Zheng Q, Yan Z, et al. Interpretation-enabled software reuse detection based on a multi-level birthmark model[C]//Proceedings of the IEEE/ACM 43rd International Conference on Software Engineering. IEEE, 2021: 873-884.

[23] Lv Y E, Yang Y, Zeng J X. An interpretable mechanism for personalized recommendation based on cross feature[J]. Journal of Intelligent and Fuzzy Systems, 2021, 40(2): 1-12.

[24] Pintelas E G, Liaskos M, Livieris I E, et al. A novel explainable image classification framework: case study on Skin cancer and Plant disease prediction[J]. Neural Computing and Applications, 2021, 33(22): 15171-15189.

[25] **刘军民，李凌敏，侯梦然，等. 深度学习的可解释性研究综述[J/OL]. 计算机应用**. https://kns.cnki.net/kems/detail/51.1307.TP.20220408.1318.006.html, 2022-04-11.

[26] Du M, Liu N, Hu X. Techniques for interpretable machine learning[J]. Communications of the ACM, 2019, 63(1): 68-77.

[27] Wu L, Rao Y, Zhao Y, et al. DTCA: Decision tree-based co-attention networks for explainable claim verification[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online 2020: 1024-1035.

[28] Speer R, Chin J, Havasi C. ConceptNet 5.5: An open multilingual graph of general knowledge [C]//Proceedings of the 31st AAAI Conference on Artificial Intelligence, 2017: 4444-4451.

[29] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. Arxiv preprint arXiv: 1301.3781 vs., 2013.

[30] Pennington J, Socher R, Manning C. GloVe: Global vectors for word representation[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Doha, 2014: 1532-1543.

[31] Seyeditabari A, TABARI N, Gholizade S, et al. Emotional embeddings: Refining word embeddings to capture emotional content of words[J].Arxiv preprint arXiv:1906.00112, 2019.

[32] 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,

2017：708-717.
[33] Ma J，Gao W，Mitra P，et al. Detecting rumors from microblogs with recurrent neural networks [C]// Proceedings of the 25th International Joint Conference on Artificial Intelligence. 2016：3818-3824.
[34] Kim Y. Convolutional neural networks for sentence classification[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing，2014：1746-1751.
[35] Yang Z，Yang D，Dyer C，et al. Hierarchical attention networks for document classification[C]//Proceedings of the NAACL，2016：1480-1489.

葛晓义(1995—)，**硕士研究生，主要研究领域为自然语言处理、舆情分析。**
E-mail：wj_gxy@163.com

魏彬(1982—)，**博士，讲师，主要研究领域为网络安全、舆情分析。**
E-mail：weibin82@126.com

张明书(1977—)，**通信作者，博士，副教授，主要研究领域为网络安全、舆情分析。**
E-mail：zms2099@163.com

(上接第128页)

[25] Tianqi Ch，Carlos G. XGBoost：A scalable tree boosting system. [C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Association for Computing Machinery，New York，USA，2016：785-794.
[26] Guolin K，Qi M，Thomas F，et al. LightGBM：A highly efficient gradient boosting decisiontree[J]. Advances in Neural Information Processing Systems 30. December 2017：3149-3157.
[27] **周志华. 机器学习[M]. 北京：清华大学出版社，**2016：98-101.
[28] Menze B H，Kelm B M，Masuch R，et al. A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data[J]. BMC Bioinformatics. 2009，10(1)：213.
[29] Kipf T N，Welling M. Semi-supervised classification with graph convolutional networks [C]//Proceedings of the 5th International Conference on Learning Representations，2017：1-14.

傅湘玲(1975—)，**博士，副教授，主要研究领域为智慧金融、文本挖掘等。**
E-mail：fuxiangling@bupt.edu.cn

赵朋亚(1995—)，**硕士研究生，主要研究领域为欺诈检测、网络表征学习等。**
E-mail：zpy1101936864@bupt.edu.cn

闫晨巍(1995—)，**博士研究生，主要研究领域为知识图谱、表示学习等。**
E-mail：chenwei.yan@bupt.edu.cn