专 题

# 融合语义增强的社交媒体虚假信息检测方法研究*

王 昊 龚丽娟 周泽聿 范 涛 王永生
（南京大学信息管理学院 南京 210023）
（江苏省数据工程与知识服务重点实验室 南京 210023）

**摘要**：【目的】通过构建自动化检测模型有效识别社交媒体中的虚假信息，探讨如何解决人工识别、单特征机器学习等现存方法难以兼顾海量数据处理的速度与准确性的问题。【方法】本文以新浪微博社交平台为研究对象，以单一文本特征BFID模型作为实验基准模型，提出两种融合语义增强的虚假信息检测方法。【结果】以单一文本特征BFID模型的结果为基线，本文提出的融合情感特征的BFID-SEN模型在虚假信息识别的部分准确率上提升约1.59个百分点；融合图片特征的BFID-IMG模型通过结合深度残差网络ResNet，在虚假信息识别的部分准确率上稳定提升约0.78个百分点。【局限】由于融合情感特征的语料数量、情感类别与多模态虚假信息数据集有限，模型训练不充分，因此语义增强的融合效果有限。【结论】本文提出的两种融合语义增强方法均能在一定程度上更好地识别虚假信息。

**关键词**：虚假信息 语义增强 多模态 新浪微博 情感分析
**分类号**：TP391
**DOI**：10.11925/infotech.2096-3467.2022.0923
**引用本文**：王昊，龚丽娟，周泽聿等. 融合语义增强的社交媒体虚假信息检测方法研究[J]. 数据分析与知识发现，2023，7(2)：48-60.(Wang Hao, Gong Lijuan, Zhou Zeyu,et al. Detecting Mis/Dis-information from Social Media with Semantic Enhancement[J]. Data Analysis and Knowledge Discovery, 2023, 7(2): 48-60.)

## 1 引 言

随着互联网技术和社交平台的蓬勃发展，用户能够以匿名的形式自由发表观点。但平台的互动性与匿名性特点[1]，让部分用户为达个人目的而不负责任地发布一些被扭曲的事实或凭空捏造的虚假信息，如在新冠疫情期间，网络虚假信息由于内容偏激、情感强烈，刺激了网友们的负面情绪，进而在互联网中飞速传播并广泛蔓延，给现实生活带来了极大的负面影响。

目前为止，“谣言”还没有一个公认的确切定义。广义上一般将基于不存在事实而凭空捏造的话语与未经公认的传说等都统称为“谣言”[2]，而狭义上一般指流传的未经证实或凭空捏造的言论。随着社交平台用户数量激增，用户产生的数据呈指数型增长，面对海量的数据，若仅通过现有的人工方式或单特征机器学习方法进行甄别，虽然能较为准确地判别出虚假信息，但是时间和人力成本高昂，因此有必要实现网络虚假信息的自动检测。

在对虚假信息自动检测的研究中，大多数学者将其转化为分类问题解决，尽管目前已有基于单一特征的机器学习算法以维持网络安全环境，但多数研究停留在文本深层次语义特征或浅层情感提取中，对于深层次语义特征增强方面的研究较为缺乏。近年来，深度学习及多模态等技术在文本分类领域

通讯作者（Corresponding author）：周泽聿（Zhou Zeyu），ORCID：0000-0003-2757-2992，E-mail：mf20140111@smail.nju.edu.cn。

*本文系国家自然科学基金项目（项目编号：72074108）和中央高校基本科研业务费专项资金资助项目（项目编号：010814370113）的研究成果之一。

The work is supported by the National Natural Science Foundation of China (Grant No. 72074108), the Fundamental Research Funds for the Central Universities (Grant No. 010814370113).

取得了一定的效果[3]。本文以新浪微博社交平台为研究对象，提出两种融合语义增强的虚假信息检测方法，以提高虚假信息自动化检测的效果。

## 2　研究现状

### 2.1　虚假信息检测

由于“谣言”一词没有一个公认的确切定义，因此根据检测目标的不同，网络谣言检测问题可划分为两类：一是检测真实性尚未确定的信息[4]，即认为“谣言”的关键特征在于“未经证实”；二是识别最终被验证为虚假的谣言，不少专家学者认为被广为传播且实际为虚假的谣言危害性更强[5-7]，因此“谣言”等同于错误的、虚假的信息[8-10]，同时认为与这类信息相对应的是来自官方渠道的新闻信息[11]，或者认为与官方信息不相同的便是谣言，也可称为虚假信息[12]。本文属于第二类，将“谣言”与错误或虚假的信息等同看待，因此本文的研究目标是识别检测社交媒体信息中的虚假信息。

在虚假信息检测研究中，根据检测对象粒度可将现有研究分为三类：基于单文本的细粒度虚假信息检测、基于多文本的粗粒度虚假信息检测、基于信息源的虚假信息检测[13]。

（1）基于单文本的细粒度虚假信息检测，即判断单条文本信息是否是虚假信息，类似于垃圾信息的识别，可以通过将虚假信息检测问题转化为分类问题来解决，如祖坤琳等[14]基于微博评论情感倾向构建情感特征，利用SVM模型将虚假信息检测问题转化为分类问题。

（2）基于多文本的粗粒度虚假信息检测主要是针对事件或主题层面的虚假信息检测[15]，将多条文本组成的事件作为研究对象，从传播特征、用户特征、文本特征等多个维度构建特征向量，典型代表是针对突发事件的虚假信息检测，其实质是将该突发事件看作一个整体[16]，对这一事件中各个类型的虚假信息进行识别，这种方式只能针对特定主题的虚假信息进行有效识别，并不具有普适性。

（3）基于信息源的虚假信息检测将网络虚假信息传播模型转化为某种信息传播模型（如传染病传播模型SI、SIS、SIR等），识别虚假信息传播的原始节点。目前的研究主要是针对单个信息源进行检测，如陈一新等[17]添加“辟谣者”角色并设定对应的节点状态转化规则以构建基于社交网络的网络虚假信息传播模型；刘彻等[18]提出一种改进的IMPA算法用于检测虚假信息源，在相同的任务下新算法有更高的效率。

由于现有研究大多将虚假信息自动检测转化为分类问题，因此识别效果很大程度上取决于特征的选择与质量。其中在特征选择上分为用户行为特征与文本特征。对于用户行为特征，由于虚假信息发布者与正常用户在行为上或多或少存在差异，因此用户行为特征主要关注信息在产生、传播、接收过程中各参与者的用户信息与行为[19-20]，如参与者近期的发文量、转发量、粉丝量、关注数、交互行为[6]等。在英文社区与中文社区中，Chang等[19]以Twitter平台中的用户行为数据为基础，将识别极端用户的方法用于检测网络政治虚假信息；Wu等[6]以新浪微博作为数据源，提出利用虚假信息传播结构在语义特征的基础上添加传播特征；Cai等[21]从人群反应中提取转发与评论特征，与文本特征结合用于微博网络谣言识别；Liang等[22]提出5个新特征并通过实验验证了新方法与新特征的有效性。这类方法一般人工干预较多且对于不同的场景规则可能差异较大，不具有较好的普适性[21]。对于文本特征，最初多采用基于显性特征的方式构建特征，如词性或标点符号类型与数量等句法特征、字符串长度或大小写等形式的词语特征[23]、符号特征、模因特征[24]（如是否包含链接）等，这种方式仅适用于数据量小且领域针对性强的情况，对于大数据量的场景不具有普适性，因此对于文本隐性特征的挖掘逐渐被大众所重视[25]。目前大多数研究主要基于递归神经网络[26-27]、BP神经网络[28]、卷积神经网络[29]等深度学习方法挖掘文本深层次语义，如Ma等[26]提出基于递归神经网络的新浪微博虚假信息检测模型，通过学习隐藏上下文信息检测网络虚假信息；Chen等[30]为解决上下文信息随时间变化的关系，以递归神经网络（RNN）为基础提出深度注意模型（RNNS）；程亮等[31]基于BP神经网络模型改进其激活函数，对微博虚假信息话题进行检测。另外，也有不少学者从文本与文本之间的语义相关性角度着手，如Zhang等[32]将内外一致性、信息匹配度、受欢迎度等特征作为隐式特征，与

其他浅层显式特征进行结合。由此看出，相较于人工方法，基于神经网络的方法进行特征工程往往会更为高效与准确。

### 2.2 多模态应用

除了包含深层次语义特征，文本特征还蕴含丰富的情感特征，许多学者基于这一特点，对文本中蕴含的情感特征进行挖掘并结合文本语义特征用于解决相关问题，如情感词的统计性特征[33]、情感词典、多任务学习[34]、深度学习等。除了文本本身的情感，也有研究者将评论等传播性特征的情感因素考虑在其中[35]，如沈瑞琳等[36]提出将虚假信息检测与情感分析结合起来的多任务学习方法，将情感分析作为虚假信息检测的辅助方法；陈帆[37]以医药疾病类领域的虚假信息数据为基础，提出微博文本来源可信度这一指标，并加入微博评论情感得分以及传播结构树的相似度作为补充特征；李巍胤[38]将微博评论情感倾向作为补充特征，结合微博传播树结构提出一种基于情感极性的微博虚假信息识别模型等。

此外，用户在新浪微博社交平台还能够通过添加图片或视频等多媒体信息的形式来达到强化表达的目的。相较于文字信息而言，图片能够更加生动快速地传递信息，更好地吸引人们的注意力，相关研究表明，带有图片的微博平均转发量是不带有图片的微博平均转发量的11倍[39]，从中不难看出，图片特征在信息传播过程中是十分重要的。图片特征的提取与构建主要表现在以下三个方面：

（1）统计特征：对图片的附属特征等统计性特征进行提取，如是否包含图片或视频链接等。Gupta等[40]将用户是否上传头像作为评估用户可信度的指标；Wu等[6]通过单条微博是否包含多媒体信息构建特征；Sun等[41]提出利用搜索引擎与计算图片相似度的方法获取原始图片的发布时间构建特征，并用于识别文本与图片不相匹配的虚假信息事件中。

（2）视觉语义特征：相关研究表明，通常真实新闻所包含的图片之间更具有差异性，而虚假新闻所包含的图片多样性较差，因此可以通过图片相似度、清晰度等方面构建特征[42-43]。

（3）复杂语义特征：随着深度神经网络算法在计算机视觉领域的不断发展，利用深度学习的方法提取图片深度特征比人工构建特征有更好的效果与效率[44]，也更能适应海量数据的处理。

在图文多模态特征融合上，如何利用图片或视频等视觉特征进行多模态虚假信息识别依然处于探索阶段，如张少钦等[45]利用预训练的Faster RCNN（Region-CNN）网络进行图片特征提取，并结合注意力机制对文本与图片特征进行融合。谢豪等[46]提出一种基于多层语义融合的社交媒体图文信息情感分类模型以充分挖掘图文之间的关联性和互补性。范涛等[47]构建一种DNNs-SVM的多模态融合网民情感识别模型。张国标等[48]通过对比分析不同特征组合方式和不同分类方法，认为融合文本特征和图像特征的多模态模型可以有效提升虚假新闻检测效果。基于新浪微博平台本身的特性以及当前互联网环境的不断发展，图片或视频等形式的多媒体信息在信息传播过程中表现出愈发重要的作用，如何将多媒体信息应用于辅助谣言检测也是当前的研究热点之一。

综上：大多数研究将虚假信息检测问题转化为分类问题处理，除了利用神经网络的方法挖掘文本深层次语义特征之外，情感特征作为文本特征的重要内容之一，许多学者将其作为文本特征的重要补充特征进行研究，但大多停留在浅层情感的提取中，缺乏对深层次情感特征的提取。多媒体信息在信息传播过程中表现出十分重要的作用，如何利用图片或视频等视觉特征进行多模态谣言识别依然处于探索阶段。因此，本文以深度学习为基础，针对社交平台中的谣言信息，提出一种新的融合情感特征与图片特征的虚假信息检测方法，提高当前网络环境下虚假信息自动化检测的性能和效率。

## 3 方法与数据

### 3.1 研究框架

本文以单一文本特征的BFID（BERT False-Information-Detection）模型为基准模型，同时构建了融合情感特征的BFID-SEN（BFID-Sentiment）模型与融合图片特征的BFID-IMG（BFID-Image）模型，研究框架如图1所示。

（1）基于单一文本特征的BFID虚假信息检测模型的主要内容是利用BERT模型进行文本表示，再将文本矩阵输入BiLSTM+Attention模块中，最后输

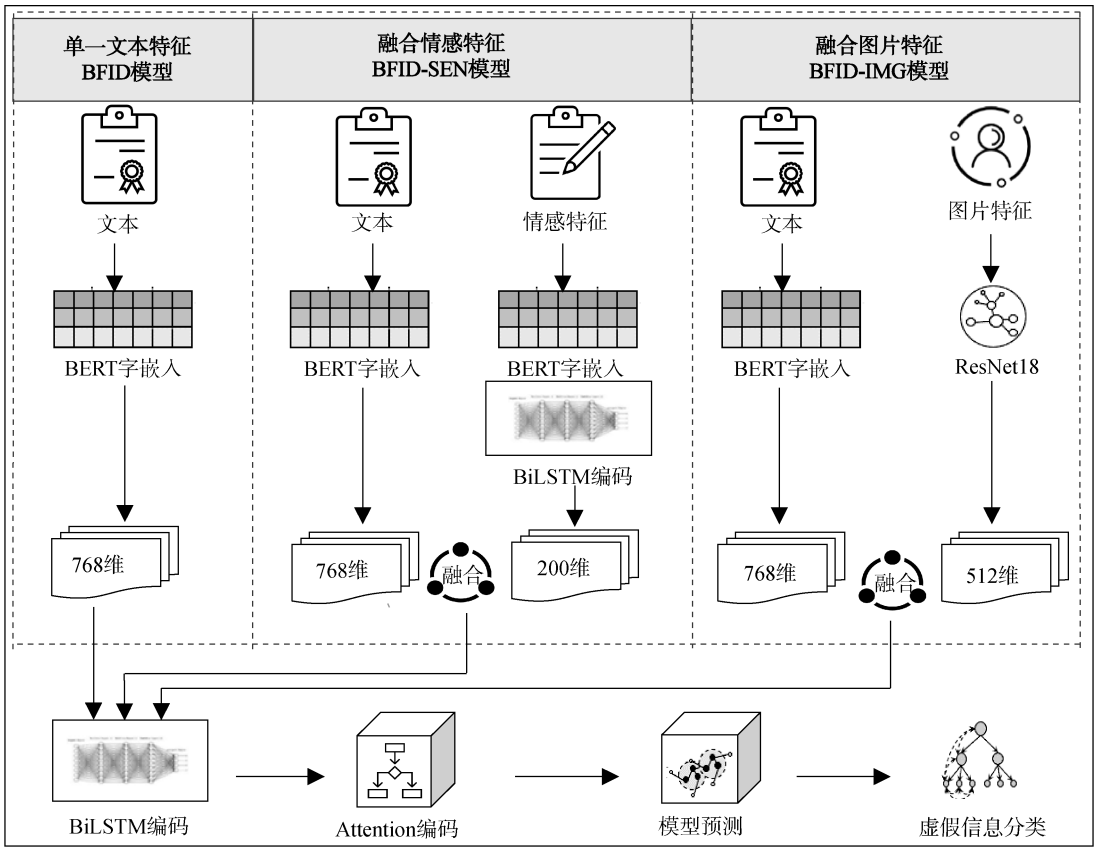



图 1　研究框架

Fig.1　Research Framework

出虚假信息的检测效果。

（2）融合情感特征的 BFID-SEN 模型借助公开的新浪微博情感分析语料，将其作为训练集，通过 BERT+BiLSTM 模型进行训练得到本文所使用的情感分析模型，再将本文实验数据作为测试集，抽取情感分析模型的中间层向量输出作为融合情感因素的补充特征，并将其与基于单一文本特征中的文本向量矩阵进行拼接，得到新的文本向量矩阵用于构建模型并对虚假信息进行检测。

（3）融合图片特征的 BFID-IMG 模型与 BFID-SEN 模型类似，首先利用 ResNet 提取图片特征，考虑到句子向量维度为 768 维，为与文本向量维度保持一致，选择 ResNet18 模型的全连接层的输出向量与单一文本特征向量矩阵进行拼接，得到新的特征矩阵构建模型用于虚假信息检测。其中，在多模态数据集中，文本与图片存在一对一和一对多的关系，因此在图片特征的处理上，对于一条文本记录对应一张图片的情况，直接获取 ResNet18 模型的全连接层输出作为图片特征向量；对于一条文本记录对应多张图片的情况，利用 ResNet18 模型获取各个有效图片的全连接层输出后进行纵向相加求平均，得到的结果即为该条文本记录所对应的图片特征。

### 3.2　基于单一文本特征的 BFID 模型构建

目前基于单一特征的机器学习分类算法大多停留在文本深层次语义特征或浅层情感特征的提取上，同时人工干预较多，且对于不同的场景规则可能差异较大[19-26]，不具有较好的普适性。近年深度学习及多模态等技术在文本分类任务上展现出了优越性能，因此，本文以基于单一文本特征的 BFID 虚假信息检测模型作为基准模型，其主要内容是利用 BERT 模型进行文本向量化，再将文本向量矩阵输入 BiLSTM+Attention 模块中，最后输出虚假信息的检测结果。

BERT（Bidirectional Encoder Representations

from Transformers）模型是由谷歌于2018年提出的一个无监督预训练模型[49]，使用双向Transformer模块，主要包括预训练和微调两个部分。BERT模型的主要创新点在于预训练阶段分别从词语和句子层面捕获语义表示以便更好地训练模型，从而提高模型的使用效果以及普适性。

双向长短期记忆网络（Bi-directional Long Short-Term Memory，BiLSTM）[50]是由前向LSTM与后向LSTM组合而成，与传统的LSTM模型相比，BiLSTM的优势在于既能够捕获之前时间尺度出现的信息，也能够获得当前时刻之后出现的信息。门控循环单元（Gate Recurrent Unit，GRU）是由Cho等[51]提出的LSTM模型的一个变体，相较于LSTM模型具有输入门、遗忘门和输出门三个函数用于控制数据流动，GRU只有两个门，分别为更新门和重置门，因此模型训练所需参数更少，训练速度也较LSTM更快。

Attention机制由Treisman等[52]提出，其本质是实现信息处理资源的高效分配，能够对重要信息赋予更高的权重，对不重要或不相关的信息赋予更低的权重，同时能够不断动态地调整权重。这使得这一方法即使在不同的情况下也能提取重要特征，具有较高的鲁棒性和可扩展性[53]。目前，Attention机制已广泛应用于语音识别、图像检测、自然语言处理等多个领域[54]，并取得了相当不错的效果。

### 3.3 融合情感特征的BFID-SEN模型构建

由于基于情感词典与情感统计特征等方法均具有难以捕获深层次语义的局限性，同时本文在数据预处理阶段并未删除以“[表情文本]”形式出现的微博表情，即认为这样的文本形式也属于浅层情感特征的范畴，因此以构建情感分析模型的方式提取深层情感特征，并通过融合语义增强的方法，将文本特征与深层情感特征结合，这种简单特征融合的方法具有相对简单且能产生相当高的精度的优点[55]。

BFID-SEN模型主要分为三部分：

（1）将公开的已标注好的新浪微博情感分类语料作为训练集，以本文采集的新浪微博虚假信息数据集作为测试集，将训练集数据输入BERT模型中得到文本向量矩阵$M \in R^{k\times d}$，其中，$k$是句子$S$的词汇量，$d$表示单词嵌入的维数；再将得到的文本向量矩阵输入BiLSTM神经网络中得到用于情感分类的BERT+BiLSTM模型，可以得到带有上下文信息的文本特征表示$S=\{f_1,f_2,\cdots,f_k\}$，其中，$f_i \in R^{k\times df}$表示第$i$个单词的文本特征，$df$是BiLSTM模型的隐藏状态向量的维数。将测试集向量矩阵输入已保存的模型中，获取神经网络中隐藏层的输出，该输出向量即代表在情感分类问题中虚假新闻文本对应的向量表示，可认为从情感分类的角度引入了文本的情感特征，得到的结果为一个200维度的隐藏层向量矩阵。

（2）以本文采集的虚假新闻语料为基础，利用BERT模型获取每条文本对应的句子向量，得到[虚假信息语料长度×968]格式的文本向量矩阵。

（3）将前两个部分的文本向量矩阵进行横向拼接，将得到的句子矩阵输入BiLSTM+Attention模块中，得到最后的分类结果。

### 3.4 融合图片特征的BFID-IMG模型构建

迄今为止，深度学习方法已广泛应用于图像领域，尤其是图像分类方面[56]，目前用于提取图片特征的方法中较为主流的方法有卷积神经网络、深度残差网络、密集连接网络、自动编码器等。其中，深度残差网络（Residual Neural Network，ResNet）有助于神经网络的快速训练[57]，同时提升模型准确率[58]。ResNet核心是建立前面层和后面层的“短路连接”，从而有助于在训练过程中实现梯度的反向传播，同时采用跳跃连接的方式，不仅能够提取到多层特征信息，还有助于提升模型性能。目前常用的方法为ResNet18、ResNet 34、ResNet50、ResNet101和ResNet152，其主要区别在于网络深度及Kernel的选择不同，各模型结构上的具体区别如表1所示。

考虑到句子向量的维度为768维，采用ResNet18模型用于图片特征提取，具体方法是将全连接层的输出向量（512维）与文本句子向量（768维）进行拼接，并作为输入矩阵用于分类。BFID-IMG模型主要包括三个部分：

（1）数据预处理，对GIF格式的动图、ResNet模型无法成功提取特征的无效图片进行删除处理，仅保留JPG、PNG与JPEG等普通格式的有效图片文件。

（2）由于所使用的数据集中，文本记录与图片存在一对一和一对多的关系，因此在特征的处理上，对

表 1　不同 ResNet 模型各层结构

Table 1　The Layer Structure of Different ResNet Models

| Layer Name | Output Size | 18-Layer | 34-Layer | 50-Layer | 101-Layer |
|---|---|---|---|---|---|
| Conv1 | 112 × 112 | 7 × 7, 64, stride 2 | | | |
| | | 3 × 3 max *pool*, stride 2 | | | |
| Conv2_x | 56 × 56 | $\begin{bmatrix}3\times3,64\\3\times3,64\end{bmatrix}\times2$ | $\begin{bmatrix}3\times3,64\\3\times3,64\end{bmatrix}\times3$ | $\begin{bmatrix}1\times1,64\\3\times3,64\\1\times1,256\end{bmatrix}\times3$ | $\begin{bmatrix}1\times1,64\\3\times3,64\\1\times1,256\end{bmatrix}\times3$ |
| Conv3_x | 28 × 28 | $\begin{bmatrix}3\times3,128\\3\times3,128\end{bmatrix}\times2$ | $\begin{bmatrix}3\times3,128\\3\times3,128\end{bmatrix}\times4$ | $\begin{bmatrix}1\times1,128\\3\times3,128\\1\times1,512\end{bmatrix}\times4$ | $\begin{bmatrix}1\times1,128\\3\times3,128\\1\times1,512\end{bmatrix}\times4$ |
| Conv4_x | 14 × 14 | $\begin{bmatrix}3\times3,256\\3\times3,256\end{bmatrix}\times2$ | $\begin{bmatrix}3\times3,256\\3\times3,256\end{bmatrix}\times6$ | $\begin{bmatrix}1\times1,256\\3\times3,256\\1\times1,1024\end{bmatrix}\times6$ | $\begin{bmatrix}1\times1,256\\3\times3,256\\1\times1,1024\end{bmatrix}\times23$ |
| Conv5_x | 7 × 7 | $\begin{bmatrix}3\times3,512\\3\times3,512\end{bmatrix}\times2$ | $\begin{bmatrix}3\times3,512\\3\times3,512\end{bmatrix}\times3$ | $\begin{bmatrix}1\times1,512\\3\times3,512\\1\times1,2048\end{bmatrix}\times3$ | $\begin{bmatrix}1\times1,512\\3\times3,512\\1\times1,2048\end{bmatrix}\times3$ |
| | 1 × 1 | Average pool, 1000-d fc, softmax | | | |
| FLOPs | | $1.8\times10^9$ | $3.6\times10^9$ | $3.8\times10^9$ | $7.6\times10^9$ |

于一条文本对应一张图片的情况，直接利用 ResNet 模型提取全连接层的输出向量作为最终图片特征，而对于一条文本对应多张图片的情况，对 ResNet 模型全连接层的向量进行求和平均，得到特征向量作为该条文本记录对应的最终图片特征向量，向量维度为 512 维。

（3）结合以 BERT 为基础构建的句子向量矩阵，将上一步形成的图片补充特征矩阵进行拼接，得到新的特征矩阵（矩阵大小为［记录数目×1280］）后，利用与前文相同的方法进行实验，得到各文本记录对应的分类结果。

## 4　实验结果与分析

### 4.1　数据来源及预处理

（1）数据来源

本文以新浪微博为数据来源，采集时间跨度为 2012 年至 2020 年，数据主要分为虚假信息与真实信息两个部分。实际场景中，虚假信息往往混杂在海量的真实信息中，两类信息的数量较为不平衡；在神经网络中，当某个类别的样本数量非常庞大时，损失函数的值会被样本数量较大的类别所影响，使得模型分类倾向于样本量较大的类别。可以从模型层面控制损失函数或从数据层面通过人为控制正负样本的比例以解决样本不均衡的问题。本文从数据层面将数据集中的正负样本比例控制在 1∶1 左右，形成实验初始数据集。

数据集整体情况如表 2 所示，数据形式大多为一条文本记录对应一张图片，也存在一部分记录表现为一条文本对应多张图片的形式，图片类型主要有 JPG、JPEG、PNG 等。

表 2　去重后数据情况

Table 2　Data After De-duplication

| 数据类型 | 文本 | 文本+图片 |
|---|---|---|
| 虚假信息 | 20 778 | 12 261 |
| 真实信息 | 20 683 | 12 561 |
| 总计 | 41 461 | 24 822 |

（2）数据预处理

①仅包含文本的数据集

仅包含文本的数据集预处理包括三个方面：第一，对明显无效的记录进行删除，如显示已被删除、无效内容等；第二，对包含 HTML 标签的文本进行数据清洗，即利用正则表达式的方法，对无用的 HTML 标签进行替换删除，同时提取其中包含的图片与视频网址链接，以便后续根据链接下载图片与视频文件；第三，按照 8∶1∶1 的比例划分训练集、测试集与验证集，数据划分结果如表 3 所示。

表3 仅含文本数据集划分结果

Table 3 Text-Only Data Segmentation Result

| 训练集 | | 验证集 | | 测试集 | |
|---|---|---|---|---|---|
| 虚假信息 | 真实信息 | 虚假信息 | 真实信息 | 虚假信息 | 真实信息 |
| 16 678 | 16 491 | 2 072 | 2 074 | 2 028 | 2 118 |

新浪微博平台允许用户在发布文字观点时附加表情以强化其观点表达，在获取得到的文本中，这些表情以文本的形式包含在英文中括号中，如表达愤怒情绪的表情呈现为“[生气]”、“[愤怒]”等文本表示形式，表达开心情绪的表情呈现为“[开心]”、“[祝贺]”等文本表示形式，表达难过情绪的表情呈现为“[难过]”、“[大哭]”等文本表示形式。考虑到这些表情文本在实验数据中普遍存在，且表情带有一定的情感信息，属于情感浅层特征，对于分类是有用的，因此本文对于表情文本不作删除处理。同样地，对于文本中出现的标点符号也不作删除处理。

②同时包含文本和图片的数据集

以仅包含文本的数据集作为基础，从中筛选出同时包含文本与图片或视频链接的记录作为多模态虚假信息数据集。多模态虚假信息数据集的预处理包括三个方面：第一，由于新浪微博平台中，视频信息的有效期普遍较短，数据集中绝大部分的视频链接对应的是无法正常播放的无效视频，因此根据网址链接的特点，将包含视频链接的记录进行删除，仅保留包含文件链接的记录；第二，数据集中包含的图片类型大多是JPG、JPEG、PNG、GIF形式，其中GIF格式的图片文件是一种特殊的图片形式，考虑到这类文件兼具静态图片与动态视频的特殊性，且在数据集中占比极小，仅占多模态数据集总数的0.017%，因此对GIF格式的文件作删除处理，即仅考虑静态图片特征；第三，按照8∶1∶1的比例划分训练集、测试集与验证集，最终得到的多模态虚假信息数据集如表4所示。

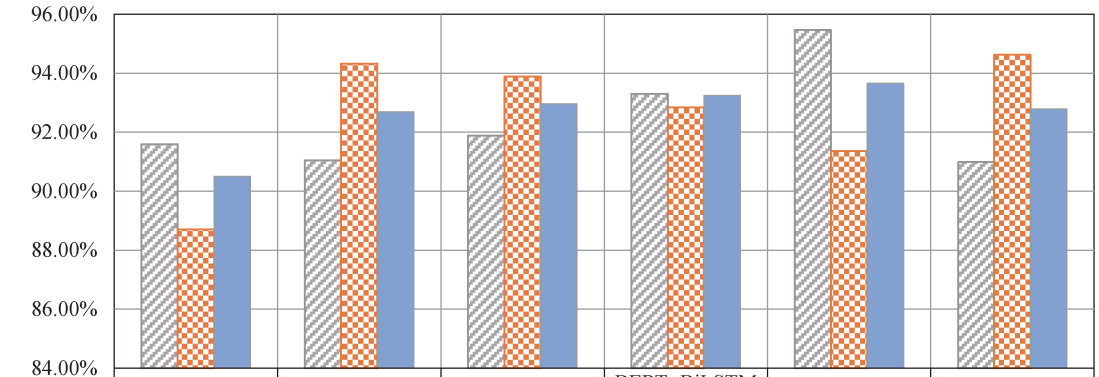
表4 多模态虚假信息数据集划分结果

Table 4 Multi-Modal False Information Data Segmentation Result

| 训练集 | | 验证集 | | 测试集 | |
|---|---|---|---|---|---|
| 虚假信息 | 真实信息 | 虚假信息 | 真实信息 | 虚假信息 | 真实信息 |
| 9 741 | 9 772 | 1 199 | 1 240 | 1 222 | 1 216 |

### 4.2 基于单一文本特征的BFID模型

利用BERT模型对虚假信息进行文本表示，并以BERT模型为基础，对比不同中间层模型，分析不同模型在虚假信息识别问题上的效果，实验结果如图2所示。其中，Precision表示虚假信息部分的识别准确率，Recall表示虚假信息部分的召回率，Accuracy表示整体准确率。




| | BERT | BERT+LSTM | BERT+BiLSTM | BERT+BiLSTM+Att | BERT+GRU | BERT+BiGRU |
|---|---|---|---|---|---|---|
| Precision | 91.60% | 91.05% | 91.89% | 93.31% | 95.47% | 90.99% |
| Recall | 88.71% | 94.33% | 93.89% | 92.85% | 91.37% | 94.63% |
| Accuracy | 90.50% | 92.69% | 92.96% | 93.25% | 93.66% | 92.79% |

图2 以BERT为基础的不同模型实验效果

Fig.2 Experimental Results of Different Models Based on BERT

就整体准确率而言，BERT+GRU模型的整体效果最优，其次为BERT+BiLSTM+Attention，两者整体准确率均达到93%以上，并且两者相差约为0.41个百分点，但是考虑到实际应用场景中，虚假信息的召回率也十分重要，尽管BERT+GRU模型的整体准确率略高于BERT+BiLSTM+Att，但其负样本召回率在各模型中处于较低的水平，相比之下，BERT+BiLSTM+Att模型既有较高的整体准确率，在应用层面上也保持了负样本较高的召回率。综上，在基于单一文本特征的BFID模型中选择BERT+BiLSTM+Att模型作为实验基准。

### 4.3　融合情感特征的BFID-SEN模型

目前已标注好且公开的情感分析数据较少，本文选择新浪微博的语料（情感分为正负两类，每一类数据均为6万条，共计12万条）作为训练集。BERT+BiLSTM模型被广泛应用于情感分类任务中，并取得了较好的效果，因此本文利用该模型提取中间层带有情感特征的输出向量，结果如表5所示，情感分类的整体准确率达到86.88%，说明该语料的质量较高，可以作为训练数据应用于后续实验中。

根据BFID-SEN模型，得到不同模型的分类结果如图3所示。为探索在移除模型或算法部分特征情况下模型的性能变化，通过消融实验探究BFID-SEN模型识别方法的优越性。从图3可以看出：

表5　情感分析语料的分类结果

Table 5　Classification Results of Emotional Corpus

| 数据类型/准确率 | Precision | Recall | F1-Score |
|---|---|---|---|
| 真实信息 | 87.65% | 85.61% | 86.62% |
| 虚假信息 | 86.15% | 88.12% | 87.12% |
| Accuracy | **86.88%** | **86.88%** | **86.88%** |

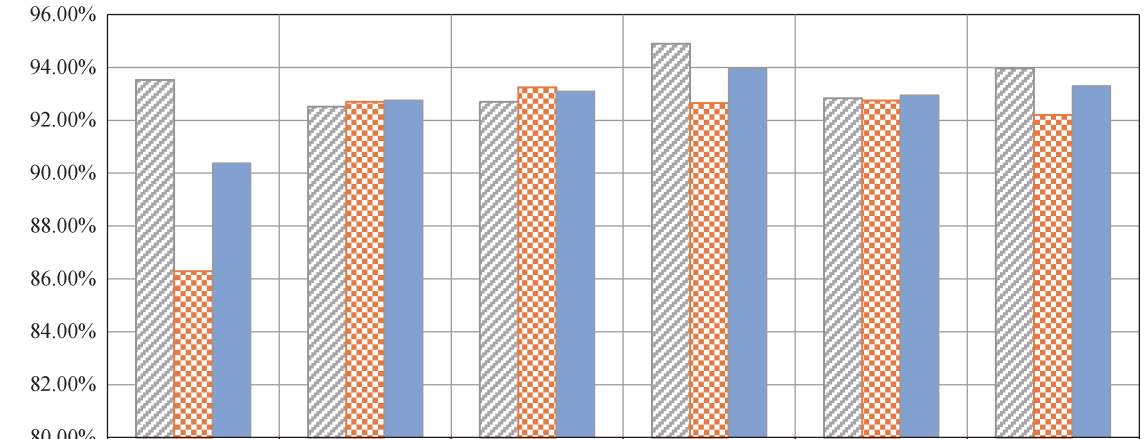
（1）就虚假信息部分的识别准确率而言，实验中所有模型的分类效果都较好，均在90%以上，且BERT+BiLSTM+Att模型的准确率最高，为94.90%。

（2）就虚假信息部分的召回率而言，除了BERT模型表现较差之外，其他模型的负样本召回率均在90%以上，且BERT+BiLSTM模型的负样本召回率最高，为93.25%。

（3）就整体准确率而言，同样是BERT+BiLSTM+Att模型效果最优，为93.97%；而BERT模型效果最差，为90.38%。

综合来看，利用本文所提出的融合语义增强的方法得到的各个实验结果中，BERT+BiLSTM+Att模型表现最优，在保证虚假信息部分较高的准确率与召回率的同时，整体效果也最优。

将BFID-SEN模型的最优实验结果与基于单一




| | BERT | BERT+LSTM | BERT+BiLSTM | BERT+BiLSTM+Att | BERT+GRU | BERT+BiGRU |
|---|---|---|---|---|---|---|
| Precision | 93.53% | 92.52% | 92.70% | 94.90% | 92.84% | 93.97% |
| Recall | 86.29% | 92.70% | 93.25% | 92.65% | 92.75% | 92.21% |
| Accuracy | 90.38% | 92.76% | 93.10% | 93.97% | 92.96% | 93.30% |

图3　添加情感特征后不同模型的实验结果

Fig.3　Experimental Results of Different Models After Adding Emotional Features

文本特征的BFID模型结果进行对比，如表6所示。

表6 BFID-SEN模型与BFID模型结果对比

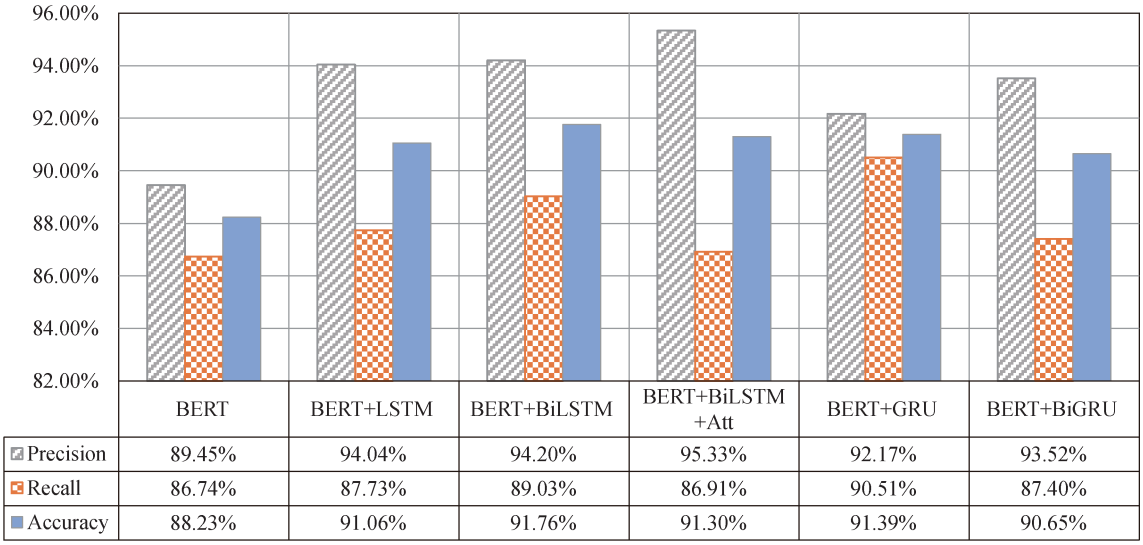
Table 6 Comparison Between BFID-SEN and BFID

| 模型名称 | 虚假信息部分的识别准确率 | 模型的整体准确率 |
|---|---|---|
| BFID模型 | 93.31% | 93.25% |
| BFID-SEN模型 | 94.90% | 93.97% |

BFID-SEN模型对提高虚假信息检测的效果具有一定的促进作用，可以有效抽取情感分析模型的中间层向量输出作为融合情感因素的补充特征，并将其与基于单一文本特征中的文本向量矩阵进行拼接，使新得到的文本向量矩阵能够有效检测虚假信息，整体准确率为93.97%。同时，虚假信息部分的准确率达到94.90%，相比于单一文本特征效果略微提升，提升效果不太明显的原因一方面在于目前公开情感分析语料的数据量有限；二是情感语料类别较为单一，导致模型训练不充分，或者捕获的情感特征较为单一，从而在提升效果上表现得并不显著。融合情感特征的方法很大程度上依赖于情感分析语料的质量以及数据体量，以保证模型能够充分学习到文本中的情感特征，进而提升虚假信息识别的准确率。

### 4.4 融合图片特征的BFID-IMG模型

根据BFID-IMG模型，得到实验结果如图4所示。整体而言，各个模型在负样本识别准确率上得到一定的提升，在实际应用场景中，负样本如果能够更准确地被识别到，那么也能大大降低人力成本，具有较为理想的实际应用价值。




| | BERT | BERT+LSTM | BERT+BiLSTM | BERT+BiLSTM+Att | BERT+GRU | BERT+BiGRU |
|---|---|---|---|---|---|---|
| Precision | 89.45% | 94.04% | 94.20% | 95.33% | 92.17% | 93.52% |
| Recall | 86.74% | 87.73% | 89.03% | 86.91% | 90.51% | 87.40% |
| Accuracy | 88.23% | 91.06% | 91.76% | 91.30% | 91.39% | 90.65% |

图4 添加图片特征后不同模型的实验结果

Fig.4 Experimental Results of Different Models After Adding Image Features

（1）就整体准确率而言，添加图片特征之后，即使数据量较前两个模型减少了将近一半，但各个模型依然有较为稳定的实验效果，即除了BERT模型之外，其他各模型整体准确率均达到90%以上，且BERT+BiLSTM模型的整体准确率最高，为91.76%。

（2）就虚假信息部分的识别准确率而言，BERT+BiLSTM+Att模型的识别准确率最高，达到95.33%，BERT模型依然效果最差。

（3）综合考虑各评价指标以及实际应用场景，不难看出，通过消融实验，BERT+BiLSTM+Att模型在综合表现上效果最佳，在保持较高整体准确率的同时，对于虚假信息这部分样本的识别效果也有较为明显的提升。

将BFID-IMG模型的最优实验结果与基于单一文本特征的BFID模型结果进行对比，如表7所示。

模型在虚假信息部分的识别准确率大多得到显著提升，虚假信息部分的准确率达到95%以上，相较

表 7　BFID-IMG 模型与 BFID 模型结果对比

Table 7　Results Between BFID-IMG and BFID

| 模型名称 | 虚假信息部分的识别准确率 | 模型的整体准确率 |
|---|---|---|
| BFID 模型 | 94.55% | 92.08% |
| BFID-IMG 模型 | 95.33% | 91.30% |

于 BFID 模型提升 0.78 个百分点，这样的提升效果可以看出 BFID-IMG 模型能有效利用 ResNet 提取到图片特征，同时与单一文本特征向量矩阵进行拼接，得到新的特征矩阵构建模型用于虚假信息检测。实验结果表明引入图片特征的 BFID-IMG 模型对于提升虚假信息识别效果具有一定积极作用。与不加入图片特征的实验结果相比，模型的整体准确率略有下降，其原因是为了与文本形成的一维向量保持一致，选取 ResNet18 模型全连接层的一维向量进行拼接，这个过程会造成一部分有用信息的丢失。由于实验数据少了一半左右，导致模型在多模态虚假信息数据上训练不充分，提升效果并不显著。

## 5　结　语

本文提出融合情感特征的 BFID-SEN 模型以及融合图片特征的 BFID-IMG 模型用于提升虚假信息检测的准确性，其中 BFID-SEN 模型的构建借助公开情感分析语料用于训练情感分析模型，进而将该模型隐藏层向量与 BERT 文本向量结合，从而完成模型构建；同理，BFID-IMG 模型将 BERT 模型与 ResNet 模型进行结合，两种模型均侧重于特征层面的融合。实验结果表明，在实验训练语料文本特征、数量级别的固有缺陷下，BFID-SEN、BFID-IMG 模型虚假信息部分识别准确率均优于基准模型。

综上，本文提出的两种模型在提升虚假信息检测的效率上均有一定的提升效果，但在文本与图片融合语义增强部分，未考虑文本与图片两种模态的数据在中间层与决策层方面的融合，也未考虑实际应用场景中的不平衡数据集，因此未来可以在公开数据集上进一步实验，以确保模型面向不同数据集的鲁棒性。此外，已有研究对视频信息的利用较少，同时也存在包含文字信息的图片，因此未来虚假信息检测领域的研究可考虑充分利用视频信息以及图片中的文本。目前已标注好的相关数据集十分有限，未来也可从半监督或无监督学习等方法着手展开研究。

## 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] 高玉君, 梁刚, 蒋方婷, 等 . 社会网络谣言检测综述[J]. 电子学报 , 2020, 48(7): 1421-1435. (Gao Yujun, Liang Gang, Jiang Fangting, et al. Social Network Rumor Detection: A Survey[J]. Acta Electronica Sinica, 2020, 48(7): 1421-1435.)

[3] 范涛, 王昊, 郝琳娜, 等 . 基于视频上下文和高维融合的突发事件中网民情感分析研究[J]. 情报科学, 2021, 39(5): 176-183. (Fan Tao, Wang Hao, Hao Linna, et al. Sentiment Analysis of Online Users in the Emergency Based on Video Context and High-Dimensional Fusion[J]. Information Science, 2021, 39(5): 176-183.)

[4] Bondielli A, Marcelloni F. A Survey on Fake News and Rumour Detection Techniques[J]. Information Sciences, 2019, 497: 38-55.

[5] Chen W L, Yeo C K, Lau C T, et al. Behavior Deviation: An Anomaly Detection View of Rumor Preemption[C]//Proceedings of the 7th Annual Information Technology, Electronics and Mobile Communication Conference. IEEE, 2016: 1-7.

[6] Wu K, Yang S, Zhu K Q. False Rumors Detection on Sina Weibo by Propagation Structures[C]//Proceedings of the 31st International Conference on Data Engineering. IEEE, 2015: 651-662.

[7] Okazaki N, Nabeshima K, Watanabe K, et al. Extracting and Aggregating False Information from Microblogs[C]//Proceedings of the 2013 Workshop on Language Processing and Crisis Information. 2013: 36-43.

[8] Yang F, Liu Y, Yu X H, et al. Automatic Detection of Rumor on Sina Weibo[C]//Proceedings of the 2012 ACM SIGKDD Workshop on Mining Data Semantics. 2012: 13.

[9] Mendoza M, Poblete B, Castillo C. Twitter Under Crisis: Can We Trust What We RT? [C]//Proceedings of the 1st Workshop on Social Media Analytics. 2010: 71-79.

[10] Yang Y K, Niu K, He Z Q. Exploiting the Topology Property of Social Network for Rumor Detection[C]//Proceedings of the 12th International Joint Conference on Computer Science and Software Engineering. 2015: 41-46.

[11] Wang S H, Terano T. Detecting Rumor Patterns in Streaming Social Media[C]//Proceedings of the 2015 IEEE International Conference on Big Data. IEEE, 2015: 2709-2715.

[12] Jain S, Sharma V, Kaushal R. Towards Automated Real-Time Detection of Misinformation on Twitter[C]//Proceedings of the

2016 International Conference on Advances in Computing, Communications and Informatics. 2016: 2015-2020.

[13] 陈燕方, 李志宇, 梁循, 等. 在线社会网络谣言检测综述[J]. 计算机学报, 2018, 41(7): 1648-1677. (Chen Yanfang, Li Zhiyu, Liang Xun, et al. Review on Rumor Detection of Online Social Networks[J]. Chinese Journal of Computers, 2018, 41(7): 1648-1677.)

[14] 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.)

[15] Kwon S, Cha M, Jung K, et al. Prominent Features of Rumor Propagation in Online Social Media[C]//Proceedings of the 13th International Conference on Data Mining. IEEE, 2013: 1103-1108.

[16] 杨文太, 梁刚, 谢凯, 等. 基于突发话题和领域专家的微博谣言检测方法[J]. 计算机应用, 2017, 37(10): 2799-2805. (Yang Wentai, Liang Gang, Xie Kai, et al. Rumor Detection Method Based on Burst Topic Detection and Domain Expert Discovery [J]. Journal of Computer Applications, 2017, 37(10): 2799-2805.)

[17] 陈一新, 陈馨悦, 刘奕, 等. 基于SIDR模型的谣言传播与源头检测研究[J]. 数据分析与知识发现, 2021, 5(1): 78-89. (Chen Yixin, Chen Xinyue, Liu Yi, et al. Detecting Rumor Dissemination and Sources with SIDR Model[J]. Data Analysis and Knowledge Discovery, 2021, 5(1): 78-89.)

[18] 刘彻, 刘祖根. 基于信息传递的谣言源检测新算法[J]. 计算机与现代化, 2020(3): 54-59. (Liu Che, Liu Zugen. A New Algorithm for Rumor Source Detection Based on Information Transmission[J]. Computer and Modernization, 2020(3): 54-59.)

[19] Chang C, Zhang Y H, Szabo C, et al. Extreme User and Political Rumor Detection on Twitter[C]//Proceedings of the 12th International Conference on Advanced Data Mining and Applications. 2016: 751-763.

[20] Zubiaga A, Aker A, Bontcheva K, et al. Detection and Resolution of Rumours in Social Media: A Survey[J]. ACM Computing Surveys (CSUR), 2018, 51(2): 1-36.

[21] Cai G Y, Wu H, Lv R. Rumors Detection in Chinese via Crowd Responses[C]//Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. IEEE, 2014: 912-917.

[22] Liang G, He W B, Xu C, et al. Rumor Identification in Microblogging Systems Based on Users' Behavior[J]. IEEE Transactions on Computational Social Systems, 2015, 2(3): 99-108.

[23] Castillo C, Mendoza M, Poblete B. Information Credibility on Twitter[C]//Proceedings of the 20th International Conference on World Wide Web. 2011: 675-684.

[24] Takahashi T, Igata N. Rumor Detection on Twitter[C]//Proceedings of the 6th International Conference on Soft Computing and Intelligent Systems, and the 13th International Symposium on Advanced Intelligence Systems. IEEE, 2012: 452-457.

[25] Ratkiewicz J, Conover M, Meiss M, et al. Detecting and Tracking the Spread of Astroturf Memes in Microblog Streams[OL]. arXiv Preprint, arXiv: 1011.3768.

[26] 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.

[27] 王鑫芸, 王昊, 邓三鸿, 等. 面向期刊选择的学术论文内容分类研究[J]. 数据分析与知识发现, 2020, 4(7): 96-109. (Wang Xinyun, Wang Hao, Deng Sanhong, et al. Classification of Academic Papers for Periodical Selection[J]. Data Analysis and Knowledge Discovery, 2020, 4(7): 96-109.)

[28] 黄亚驹, 陈福集, 游丹丹. 基于混合算法和BP神经网络的网络舆情预测研究[J]. 情报科学, 2018, 36(2): 24-29. (Huang Yaju, Chen Fuji, You Dandan. Research on the Prediction of Network Public Opinion Based on Hybrid Algorithm and BP Neural Network[J]. Information Science, 2018, 36(2): 24-29.)

[29] 徐绪堪, 周泽聿. 基于多尺度BiLSTM-CNN的微信推文的情感分类模型及应用研究[J]. 情报科学, 2021, 39(5): 130-137. (Xu Xukan, Zhou Zeyu. A Multi-Scale BiLSTM-CNN Based Emotion Classification Model for WeChat Tweets and Its Application[J]. Information Science, 2021, 39(5): 130-137.)

[30] Chen T, Li X, Yin H Z, et al. Call Attention to Rumors: Deep Attention Based Recurrent Neural Networks for Early Rumor Detection[C]//Proceedings of the 22nd Pacific-Asia Conference on Knowledge Discovery and Data Mining. 2018: 40-52.

[31] 程亮, 邱云飞, 孙鲁. 微博谣言检测方法研究[J]. 计算机应用与软件, 2013, 30(2): 226-228. (Cheng Liang, Qiu Yunfei, Sun Lu. Research on Detecting Microblogging Rumours[J]. Computer Applications and Software, 2013, 30(2): 226-228.)

[32] Zhang Q, Zhang S Y, Dong J, et al. Automatic Detection of Rumor on Social Network[C]//Proceedings of the 4th Natural Language Processing and Chinese Computing. 2015: 113-122.

[33] Andreevskaia A, Bergler S. Mining WordNet for Fuzzy Sentiment: Sentiment Tag Extraction from WordNet Glosses[C]//Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics. 2006: 209-216.

[34] 杨晗迅, 周德群, 马静, 等. 基于不确定性损失函数和任务层级注意力机制的多任务谣言检测研究[J]. 数据分析与知识发现, 2021, 5(7): 101-110. (Yang Hanxun, Zhou Dequn, Ma Jing, et al. Detecting Rumors with Uncertain Loss and Task-Level Attention Mechanism[J]. Data Analysis and Knowledge Discovery, 2021, 5 (7): 101-110.)

[35] 张柳, 王晰巍, 黄博, 等. 基于字词向量的多尺度卷积神经网络

微博评论的情感分类模型及实验研究[J]. 图书情报工作, 2019, 63(18): 99-108. (Zhang Liu, Wang Xiwei, Huang Bo, et al. A Sentiment Classification Model and Experimental Study of Microblog Commentary Based on Multivariate Convolutional Neural Networks Based on Word Vector[J]. Library and Information Service, 2019, 63(18): 99-108.)

[36] 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.(Shen Ruilin, Pan Weimin, Peng Cheng, et al. Microblog Rumor Detection Method Based on Multi-Task Learning[J]. Computer Engineering and Applications, 2021, 57(24): 192-197.)

[37] 陈帆. 基于LSTM情感分析模型的微博谣言识别方法研究[D]. 武汉: 华中师范大学, 2018. (Chen Fan. Microblog Rumor Detection Research Based on LSTM Sentiment Analysis Model [D]. Wuhan: Central China Normal University, 2018.)

[38] 李巍胤. 基于情感分析的微博谣言识别模式研究[D]. 重庆: 重庆大学, 2016. (Li Weiyin. Research on Microblog Rumors Detection Pattern Based on Sentiment Analysis[D]. Chongqing: Chongqing University, 2016.)

[39] Jin Z W, Cao J, Zhang Y D, et al. Novel Visual and Statistical Image Features for Microblogs News Verification[J]. IEEE Transactions on Multimedia, 2017, 19(3): 598-608.

[40] Gupta M, Zhao P X, Han J W. Evaluating Event Credibility on Twitter[C]//Proceedings of the 2012 SIAM International Conference on Data Mining. 2012: 153-164.

[41] Sun S Y, Liu H Y, He J, et al. Detecting Event Rumors on Sina Weibo Automatically[C]//Proceedings of the 15th Asia-Pacific Web Conference. 2013: 120-131.

[42] 王雨竹, 谢珺, 陈波, 等. 基于跨模态上下文感知注意力的多模态情感分析[J]. 数据分析与知识发现, 2021, 5(4): 49-59.(Wang Yuzhu, Xie Jun, Chen Bo, et al. Multi-Modal Sentiment Analysis Based on Cross-Modal Context-Aware Attention[J]. Data Analysis and Knowledge Discovery, 2021, 5(4): 49-59.)

[43] 张国标, 李洁. 融合多模态内容语义一致性的社交媒体虚假新闻检测[J]. 数据分析与知识发现, 2021, 5(5): 21-29.(Zhang Guobiao, Li Jie. Detecting Social Media Fake News with Semantic Consistency Between Multi-Model Contents[J]. Data Analysis and Knowledge Discovery, 2021, 5(5): 21-29.)

[44] 王仁武, 孟现茹. 图片情感分析研究综述[J]. 图书情报知识, 2020(3): 119-127.(Wang Renwu, Meng Xianru. Review of Image Sentiment Analysis[J]. Documentation, Information & Knowledge, 2020(3): 119-127.)

[45] 张少钦, 杜圣东, 张晓博, 等. 融合多模态信息的社交网络谣言检测方法[J]. 计算机科学, 2021, 48(5): 117-123. (Zhang Shaoqin, Du Shengdong, Zhang Xiaobo, et al. Social Rumor Detection Method Based on Multimodal Fusion[J]. Computer Science, 2021, 48(5): 117-123.)

[46] 谢豪, 毛进, 李纲. 基于多层语义融合的图文信息情感分类研究[J]. 数据分析与知识发现, 2021, 5(6): 103-114.(Xie Hao, Mao Jin, Li Gang. Sentiment Classification of Image-Text Information with Multi-Layer Semantic Fusion[J]. Data Analysis and Knowledge Discovery, 2021, 5(6): 103-114.)

[47] 范涛, 吴鹏, 曹琪. 基于深度学习的多模态融合网民情感识别研究[J]. 信息资源管理学报, 2020, 10(1): 39-48.(Fan Tao, Wu Peng, Cao Qi. The Research of Sentiment Recognition of Online Users Based on DNNS Multimodal Fusion[J]. Journal of Information Resources Management, 2020, 10(1): 39-48.)

[48] 张国标, 李洁, 胡潇戈. 基于多模态特征融合的社交媒体虚假新闻检测[J]. 情报科学, 2021, 39(10): 126-132.(Zhang Guobiao, Li Jie, Hu Xiaoge. Fake News Detection Based on Multimodal Feature Fusion on Social Media[J]. Information Science, 2021, 39 (10): 126-132.)

[49] 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. 2019: 4171-4186.

[50] 陈德鑫, 占袁圆, 杨兵, 等. 基于CNN-BiLSTM模型的在线医疗实体抽取研究[J]. 图书情报工作, 2019, 63(12): 105-113.(Chen Dexin, Zhan Yuanyuan, Yang Bing, et al. Research on Extraction of Online Medical Entities Based on Mixed Deep Learning Model [J]. Library and Information Service, 2019, 63(12): 105-113.)

[51] Cho K, Van Merriënboer B, Gulcehre C, et al. Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation[OL]. arXiv Preprint, arXiv:1406.1078.

[52] Treisman A M, Gelade G. A Feature-Integration Theory of Attention[J]. Cognitive Psychology, 1980, 12(1):97-136.

[53] 祁瑞华, 简悦, 郭旭, 等. 融合特征与注意力的跨领域产品评论情感分析[J]. 数据分析与知识发现, 2020, 4(12): 85-94. (Qi Ruihua, Jian Yue, Guo Xu, et al. Sentiment Analysis of Cross-Domain Product Reviews Based on Feature Fusion and Attention Mechanism[J]. Data Analysis and Knowledge Discovery, 2020, 4 (12): 85-94.)

[54] 周瑛, 刘越, 蔡俊. 基于注意力机制的微博情感分析[J]. 情报理论与实践, 2018, 41(3): 89-94. (Zhou Ying, Liu Yue, Cai Jun. Sentiment Analysis of Micro-Blogs Based on Attention Mechanism[J]. Information Studies: Theory & Application, 2018, 41(3): 89-94.)

[55] Poria S, Cambria E, Howard N, et al. Fusing Audio, Visual and Textual Clues for Sentiment Analysis from Multimodal Content [J]. Neurocomputing, 2016, 174: 50-59.

[56] 王树义, 刘赛, 马峥. 基于深度迁移学习的微博图像隐私分类研究[J]. 数据分析与知识发现, 2020, 4(10): 80-92.(Wang Shuyi, Liu Sai, Ma Zheng. Microblog Image Privacy Classification with Deep Transfer Learning[J]. Data Analysis and Knowledge Discovery, 2020, 4(10): 80-92.)

[57] Targ S, Almeida D, Lyman K. Resnet in Resnet: Generalizing Residual Architectures[OL]. arXiv Preprint, arXiv: 1603.08029.
[58] 郝旭政, 柴争义. 一种改进的深度残差网络行人检测方法[J]. 计算机应用研究, 2019, 36(5): 1569-1572.(Hao Xuzheng, Chai Zhengyi. Improved Pedestrian Detection Method Based on Depth Residual Network[J]. Application Research of Computers, 2019, 36(5): 1569-1572.)

## 作者贡献声明:

王昊:提出研究思路,设计研究方案;
龚丽娟:采集数据和进行实验,撰写与修订论文;
周泽聿:参与提出研究思路,分析数据和进行实验;
范涛:对论文提出修订意见;
王永生:论文修订。

## 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] 周泽聿.融合语义增强的社交媒体虚假信息检测方法研究. DOI: 10.57760/sciencedb.05190.

收稿日期:2022-08-31
收修改稿日期:2022-10-25

# Detecting Mis/Dis-information from Social Media with Semantic Enhancement

Wang Hao Gong Lijuan Zhou Zeyu Fan Tao Wang Yongsheng
(School of Information Management, Nanjing University, Nanjing 210023, China)
(Jiangsu Key Laboratory of Data Engineering and Knowledge Service, Nanjing 210233, China)

**Abstract:** [**Objective**] This paper builds an automated detection model to effectively identify mis/dis-information from social media, aiming to balance the speed and accuracy of processing massive data. [**Methods**] The classification model is the mainstream processing technique to detect for mis/dis-information. However, most of them could not extract deep semantic features from the texts. Therefore, we used the single text feature BFID model (BERT False-Information-Detection) as the benchmark model, and proposed two new methods with fused semantic enhancement to detect the mis/dis-information. [**Results**] We examined the new models with data from Sina Weibo. The accuracy of the model based on fused sentiment feature BFID-SEN (BFID-Sentiment) increased about 1.59 percentage point, while the accuracy of model with fused image feature BFID-IMG (BFID-Image) model improved by 0.78 percentage point. [**Limitations**] The ability to fuse semantic enhancement is limited due to the small corpus size, sentiment categories and multimodal disinformation training datasets. [**Conclusions**] The proposed methods are able to more effectively identify false information from social media.

**Keywords:** False Information Semantic Enhancement Multi-Modal Sina Weibo Sentiment Analysis