

基于改进 Transformer 的社交媒体谣言检测

郑洪浩, 郝一诺, 于洪涛, 李邵梅, 吴翼腾

(信息工程大学, 河南 郑州 450001)

摘要: 随着互联网的快速发展, 社交媒体日益广泛而深刻地融入人们日常生活的各个方面。社交媒体逐渐成为人们彼此之间用来分享意见、见解、经验和观点的工具和平台, 是人们获取分享信息、表达交流观点的主要途径。社交媒体在互联网的沃土上蓬勃发展, 爆发出令人眩目的能量。由于社交媒体的开放性, 用户规模庞大且来源复杂众多, 容易产生各种各样的谣言虚假信息。社交媒体谣言左右着网民对事件的认知、动摇着社会的稳定。因此, 如何准确高效地检测谣言成为当下亟待解决的问题。现有基于 Transformer 的社交媒体谣言检测模型忽略了文本位置信息。为有效提取文本位置信息, 充分利用文本潜在信息, 提出了一种基于改进 Transformer 的社交媒体谣言检测模型。该模型从相对位置和绝对位置两方面对传统 Transformer 进行改进: 一方面采用可学习的相对位置编码捕捉文本的方向信息和距离信息; 另一方面采用绝对位置编码将不同位置词语映射到不同特征空间。实验结果表明, 与其他基准模型相比, 所提模型在 Twitter15、Twitter16 和 Weibo 3 种数据集上的准确率分别提高了 0.9%、0.6%和 1.4%。实验结果验证了所提的位置编码改进有效, 基于位置编码改进的 Transformer 模型可显著提升社交媒体谣言检测效果。

关键词: 社交媒体谣言检测; 改进 Transformer; 位置信息

中图分类号: TP391

文献标志码: A

DOI: 10.11959/j.issn.2096-109x.2022042

Rumor detection in social media based on enhanced Transformer

ZHENG Honghao, HAO Yinuo, YU Hongtao, LI Shaomei, WU Yiteng

Information Engineering University, Zhengzhou 450001, China

Abstract: With the rapid development of the Internet, social media is increasingly integrated into all aspects of people's daily life. Social media has gradually become a tool and even a platform for people to share opinions, insights, experiences and viewpoints. It is the main method for people to obtain and share information as well as express and exchange opinions. Currently, social media mainly includes social networking sites, Weibo, Twitter, blogs, forums, podcasts and so on. Due to the openness of social media, the user scale is large and the sources are

收稿日期: 2021-08-31; 修回日期: 2022-02-15

通信作者: 于洪涛, 15937101921@139.com

基金项目: 国家自然科学基金(61601513); 郑州市协同创新重大专项(162/32410218)

Foundation Items: The National Natural Science Foundation of China (61601513), Major Collaborative Innovation Projects of Zhengzhou (162/32410218)

引用格式: 郑洪浩, 郝一诺, 于洪涛, 等. 基于改进 Transformer 的社交媒体谣言检测[J]. 网络与信息安全学报, 2022, 8(4): 168-174.

Citation Formats: ZHENG H H, HAO Y N, YU H T, et al. Rumor detection in social media based on enhanced Transformer[J]. Chinese Journal of Network and Information Security, 2022, 8(4): 168-174.

complex and numerous, then all kinds of rumors and false information may be generated easily. Rumors on social media influence netizens' understanding of events and shake the stability of society. Therefore, how to accurately and efficiently detect rumors has become an urgent problem to be solved. Existing Transformer based social media rumor detection models ignored the text location information. To effectively extract text location information and make full use of text potential information, a rumor detection model in social media was proposed and it was based on the enhanced Transformer. This model enhanced the traditional Transformer from two aspects of relative position and absolute position. It captured the direction information and distance information of the text using learnable relative position coding and mapped words from different positions to different feature spaces using absolute position coding. Experimental results show that, compared with the best benchmark model, the accuracy of the proposed model on Twitter15, Twitter16 and Weibo datasets is enhanced by 0.9%, 0.6% and 1.4%, respectively. Experimental results verify the effectiveness of the proposed location coding. And the enhanced Transformer based on location coding can significantly improve the effects of social media rumor detection.

Keywords: rumor detection in social media, enhanced Transformer, position information

0 引言

谣言是指在人和人之间传播的, 真实性不能很快得到证明或得不到证明的, 含有公众关心信息的一种特殊陈述^[1]。近年来, 随着 Twitter、微博(Weibo)等大型社交媒体平台的迅速发展, 广泛传播的谣言已成为危害人们生活的“痼疾”。谣言检测应运而生, 其对于改善网络信息生态环境质量、维护社会稳定具有重要意义。

自 Mikolov 等^[2]起, 研究者陆续在谣言检测中引入深度学习方法以避免人工特征。基于深度学习的谣言检测的基础方法论将其看作一个分类问题, 通过微博文本信息、发布者资料和传播特征对谣言真伪进行辨别^[3]。现有基于深度学习的谣言检测模型的主要差异也体现在对3种信息的处理之中。

Transformer^[4]因其并行计算结构和长文本获取能力, 在微博文本信息的获取上取得了显著突破。Yuan 等^[5]首先提出使用 Transformer 处理微博文本信息。进一步, 琚心怡等^[6]提出使用深层双向 Transformer 用于语义特征提取。

然而, 对于基于自注意力机制的 Transformer 来说, 其原生的正弦位置编码不能完整地表达位置信息。位置信息尤为重要, 失去了位置信息, 具有语义信息的句子就会变成一个词袋。位置信息包含两种: 相对位置信息和绝对位置信息。相对位置是指将位置之间的差异性映射到向量空间, 绝对位置是指不同位置的序列映射到不同特征空间。

针对上述问题, 本文首先对 Transformer 的正

弦位置编码进行深入分析, 然后提出一种基于改进 Transformer 的社交媒体谣言检测方法, 并将改进的 Transformer 命名为 TPE (transformer with positional encoding)。该方法对现有的 Transformer 进行两方面的改进: 一是使用可学习的相对位置编码以捕捉谣言文本的方向和距离信息; 二是使用绝对位置编码, 将谣言文本中不同位置的词映射到不同的特征空间。此外, 该方法使用图注意力网络将发布者资料与传播特征融入特征表示之中, 这样有利于提升模型的谣言检测性能。实验表明, 与现有的微博文本获取模型相比, TPE 可以更加有效地获取文本信息。与基准方法相比, 本文方法在 Twitter 15^[7]、Weibo16^[7]、Weibo^[8]这3种数据集上的准确率均有提高。

1 相关工作

Transformer 是 Vaswani 等^[4]提出的一种基于自注意力机制的模型, 其编码层由自注意力层和全连接层两个子层构成。对于既不使用卷积也不使用递归的 Transformer 是无法从模型结构上获取位置信息的。所以, Vaswani 等^[4]提出使用正弦位置向量与词向量相加的方式嵌入位置信息。Shaw 等^[9]在机器翻译任务上使用可学习的相对位置编码替换正弦位置编码。Devlin 等^[10]和 Liu 等^[11]提出将可学习的绝对位置编码应用于预训练模型。Yang 等^[12]将固定的相对位置编码应用于 XLnet 模型, 显著地提高了预训练模型的效果。

本文与上述研究的不同之处在于以下两点:

对 Transformer 原生的正弦位置编码进行深入分析；提出一种可以同时获取相对位置信息和绝对位置信息的新位置编码范式。接下来，对原生 Transformer 的工作原理进行介绍。

对于每头注意力，Transformer 的计算过程如下所示。

$$\mathbf{H} = \mathbf{E} + \mathbf{U} \quad (1)$$

$$\mathbf{Q} = \mathbf{H}\mathbf{W}_q, \mathbf{K} = \mathbf{H}\mathbf{W}_k, \mathbf{V} = \mathbf{H}\mathbf{W}_v \quad (2)$$

$$\mathbf{A} = \mathbf{Q}\mathbf{K}^T \quad (3)$$

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{A}}{\sqrt{d}}\right)\mathbf{V} \quad (4)$$

其中， $\mathbf{E} \in \mathbb{R}^{N \times d}$ 表示词向量， $\mathbf{U} \in \mathbb{R}^{N \times d}$ 表示位置向量， $\mathbf{H} \in \mathbb{R}^{N \times d}$ 表示模型的输入向量， $\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v \in \mathbb{R}^{d \times d}$ 表示映射矩阵， $\mathbf{A} \in \mathbb{R}^{d \times d}$ 表示注意力得分矩阵， $\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) \in \mathbb{R}^{N \times d}$ 表示输出向量， N 表示输入向量的长度， d 表示输入向量的维度。

2 谣言检测模型

2.1 微博文本信息编码

相对于现有的 RNN 类和 CNN 类模型，Transformer 具有获取长文本信息和并行计算结构的优势。然而，Transformer 的正弦位置编码在相对位置获取上仍存在局限和缺失，影响微博文本信息的获取。正弦编码的位置嵌入存在相对位置信息缺失严重的问题，具体表现在距离性信息表达模糊和方向性信息损失严重。本节先对 Transformer 的正弦位置编码进行分析，而后提出具体的改进方法。

2.1.1 正弦位置编码的性质

在 Transformer 中，只在计算自注意力得分的过程中发生位置交互。为了深入地对正弦位置编码进行研究，本文将自注意力得分的计算展开为“(a)文本-文本”“(b)文本-位置”“(c)位置-文本”“(d)位置-位置”4项。

$$\begin{aligned} a_{ij} = & \underbrace{\mathbf{e}_i \mathbf{W}^Q (\mathbf{W}^K)^T \mathbf{e}_j}_{(a)} + \underbrace{\mathbf{e}_i \mathbf{W}^Q (\mathbf{W}^K)^T \mathbf{u}_j}_{(b)} + \\ & \underbrace{\mathbf{u}_i \mathbf{W}^Q (\mathbf{W}^K)^T \mathbf{e}_j}_{(c)} + \underbrace{\mathbf{u}_i \mathbf{W}^Q (\mathbf{W}^K)^T \mathbf{u}_j}_{(d)} \end{aligned} \quad (5)$$

其中， $\mathbf{e}_i$ 表示 $\mathbf{E}$ 的行向量， $\mathbf{u}_i$ 表示 $\mathbf{U}$ 的行向量。

在自注意力得分的计算中，发生位置交互的

只有 (d) 项，本文进一步通过研究 (d) 项中 $\mathbf{u}_i$ 和 $\mathbf{u}_j$ 的点积，检验正弦位置编码是否可以表达相对位置信息。

$$w_m = \left(\frac{1}{10\,000}\right)^{\frac{2m}{d}} \quad (6)$$

$$\mathbf{u}_i \mathbf{u}_j^T = \text{sum} \left[\begin{bmatrix} \sin(w_1 i) \\ \cos(w_1 i) \\ \vdots \\ \sin(w_{\frac{d}{2}} i) \\ \sin(w_{\frac{d}{2}} i) \end{bmatrix} \odot \begin{bmatrix} \sin(w_1 j) \\ \cos(w_1 j) \\ \vdots \\ \sin(w_{\frac{d}{2}} j) \\ \sin(w_{\frac{d}{2}} j) \end{bmatrix} \right] = \sum_{m=0}^{\frac{d}{2}} \cos(w_m (i - j)) \quad (7)$$

其中， $m \in [0, \frac{d}{2}]$ 。

从式(7)可以看出， $\mathbf{u}_i$ 和 $\mathbf{u}_j$ 的点积只与 $i-j$ 相关，进一步将 $i-j$ 定义为 k ，将式(7)重新定义为式(8)。

$$f(k) = \mathbf{u}_i \mathbf{u}_{j+k}^T = \sum_{m=0}^{\frac{d}{2}} \cos(w_m k) \quad (8)$$

经分析，可以得出相对位置编码具有以下两个性质。

性质 1 相对位置编码不能准确地反映距离信息。距离信息反映在 $f(k)$ 中，主要体现在函数单调性表达上。当输入向量维度为 50 时， $f(k)$ 两侧的单调区间仅为 3。在非单调区间内，对于不同的 k ， $f(k)$ 会有相同的取值，因此 $f(k)$ 不能准确地反映距离关系。

为了深入研究 $f(k)$ 单调区间的大小，进一步求出 $f(k)$ 的一阶导数 $f'(k)$ ，具体如式(9)所示。

$$f'(k) = \sum_{m=0}^{\frac{d}{2}} -w_m \sin(w_m k) \quad (9)$$

$f(k)$ 的正负区间主要取决 w_m 的大小。在正弦位置编码中， $w_k = (1/10\,000)^{\frac{2m}{d}}$ 。因此， $f(k)$ 的单调区间最终取决于 d 的大小。微博文本在 Weibo、Twitter15、Twitter16 数据集集中的平均长度分别为 34.88、14.52 和 14.05，若要正弦位置编码精确地表

达微博文本的距离信息, 至少需要词向量的维度达到 2 743。这样高维度的向量会带来大量的计算开销且难以获取。综上, 在使用常规词向量的情况下, 正弦位置编码无法反映距离信息。

性质 2 相对位置编码不能反映方向信息。如式(10)所示, $f(k)$ 具有对称性, 不能表征方向信息。

$$f(k) = \sum_{m=0}^{\frac{d}{2}} \cos(w_m k) = \sum_{m=0}^{\frac{d}{2}} \cos(w_m (-k)) = f(-k) \quad (10)$$

2.1.2 方法的具体改进

针对上述问题, 本文重新定义现有的注意力得分计算范式。① 使用可学习的相对位置编码以捕捉谣言文本的方向和距离信息。② 使用绝对位置编码, 将谣言文中不同位置的词映射到不同的特征空间。重新定义之后的注意力得分计算如下:

$$a_{ij} = \underbrace{e_j W_c^Q (W_c^K)^T e_l^T}_{(e)} + \underbrace{e_j W_c^Q (W_r^K)^T r_k}_{(f)} + \underbrace{e_l W_c^K (W_r^Q)^T r_k}_{(g)} + \underbrace{e_j W_c^Q b}_{(h)} \quad (11)$$

$$r_k = [\dots \sin(w_m' k), \cos(w_m' k) \dots]^T \quad (12)$$

方法具体的改进有以下 3 点。

改进 1 构建可以反映相对位置的(f)项和(g)项, 其分别揭示第 i 个输入与其下文之间的相对位置和第 i 个输入与其上文之间的相对位置。(f)项和(g)项利用相对位置编码 r_k , 直接以相对位置 k 进行建模。因为 r_k 中 $\sin(w_m' k)$ 具有 $\sin(-w_m' k) = -\sin(w_m' k)$ 的性质, 故注意力得分 a_{ij} 可以反映方向信息。因为 $\sin(w_m' k)$ 具有局部单调性, 故注意力得分 a_{ij} 可以模糊地反映距离信息。此外, 相对位置信息和文本信息是异质的信息, 本文将词向量 e 与相对位置编码 r_k 分别映射到不同的查询向量 W_c^Q 、 W_r^Q 和不同的键向量 W_c^K 、 W_r^K 。

改进 2 使用可学习 w_m' , 以一种数据驱动的方式自适应地调整单调区间, 从而更精确地获取距离信息。

改进 3 构建可以反映绝对位置的(h)项, (h)使用可学习的向量 b 与第 i 个输入的查询向量点积的方式表征输入之间的不同, 反映各个输入的绝对位置信息。

上述的改进是在文献[5, 13-17]的研究基础上进行的。

2.2 发布者资料与传播特征

为了获取发布者资料与传播特征, Liu 等^[18]提出将传播路径建模为多元的时间序列, 并应用递归网络和卷积网络的组合来捕捉用户特征传播路径。然而, 该方式不能直接融合不同微博和用户之间的全局结构信息。社交媒体是异构的网络, 具有用户、帖子、地理位置和标签等实体, 以及好友、转发和空间邻域等关系。因此全局结构信息可以丰富模型信息以提升谣言检测的性能。Yuan 等^[5]提出用图注意力网络融合不同微博和用户之间的全局结构信息, 并在主流数据集上取得了较好的效果。本文采用与文献[5]同样的结构融合发布者资料与传播特征。

3 改进 Transformer 的有效性验证实验

3.1 实验设置

(1) 数据

本文实验采用主流社交媒体平台的 3 类数据集: Twitter15、Twitter16 和 Weibo。数据集的统计如表 1 所示, 其中真实谣言是指微博文本中已注明该文本是谣言。此外, 原始的数据集中并不包含用户特征, 实验采用 Yuan 等^[5]爬取的用户信息。

表 1 数据集的统计
Table 1 Statistics of datasets

统计	Weibo	Twitter15	Twitter16
谣言文本	4 664	1 490	818
非谣言 (NR)	2 351	374	205
虚假谣言 (FR)	2 313	370	205
未经核实的谣言 (UR)	0	374	203
真实谣言 (TR)	0	372	205
用户	2 746 818	276 663	173 487

所有数据集均使用 10% 的数据作为验证集, 其余数据采用 3:1 的比例划分训练集和测试集。实验采用由预训练模型 Word2vec^[19]训练出的 300 维向量作为词表示。

(2) 评价指标和参数设置

为了评估本文改进的有效性并与其他方法进行公平的比较, 实验采用与其他研究者相同的评价指标 (准确率、精确率、召回率和 $F1$ 值), 采用反向传播的方法进行训练。在优化模型方面, 实验采用随机梯度下降和动量联合的方式, 学习

率更新采用三角法^[20]。

3.2 与主流模型的对比实验

3.2.1 对比模型

将本文模型与所选取的基准模型在相同的数据集上开展实验, 本文选取了以下几个基准模型。

(1) DTC 模型^[21], 该模型通过决策树融合了多种新闻特征进行分类。

(2) SVM-TS 模型^[22], 该模型利用时间序列特征模拟微博事件特征并通过 SVM 分类器进行分类。

(3) DT-Rank 模型^[23], 该模型基于决策树, 为每个微博簇添加短语信息以对假新闻排序。

(4) GRU-2 模型^[8], 该模型利用 GRU 网络从用户注释中学习微博事件的深层信息并完成分类。

(5) RvNN 模型^[24], 该模型利用递归神经网络构建自下而上和自上而下的树结构模型并完成分类。

(6) PPC 模型^[18], 该模型结合递归神经网络和卷积神经网络建模传播路径信息以完成分类。

(7) GLAN 模型^[5], 该模型通过 Transformer 和图注意力网络获取全局信息以完成分类。

GLAN 模型在实验采用的 3 种数据集上效果最好。

3.2.2 实验结果与分析

为了验证本文方法的有效性, 将本文模型与其他基准模型进行比较。在 Twitter15、Twitter16 和 Weibo 这 3 种数据集上的实验结果分别如表 2、表 3、表 4 所示, 本文模型记为 TPE-GAT, 未使用位置改进的模型记为 Transformer-GAT。为了保证比较的公平性, 实验引用了文献[18,24]的实验结果。

表 2 Twitter15 数据集上的实验结果
Table 2 Results on Twitter15 datasets

模型	准确率	F1 值			
		NR	FR	UR	TR
DTC	45.4%	73.3%	35.5%	41.5%	31.7%
SVM-TS	54.4%	79.6%	47.2%	48.3%	40.4%
DT-Rank	40.9%	50.1%	31.1%	47.3%	36.4%
GRU-2	64.6%	79.2%	57.4%	59.2%	60.8%
RvNN	72.3%	68.2%	75.8%	65.4%	82.1%
PPC	84.2%	81.1%	87.5%	79.0%	81.8%
GLAN	90.5%	92.4%	91.7%	92.7%	85.2%
Transformer-GAT	89.1%	93.8%	90.9%	82.8%	88.6%
TPE-GAT	91.4%	94.2%	91.2%	90.8%	89.2%

表 3 Twitter16 数据集上的实验结果
Table 3 Results on Twitter16 datasets

模型	准确率	F1 值			
		NR	FR	UR	TR
DTC	46.5%	64.3%	39.3%	40.3%	41.9%
SVM-TS	57.4%	75.5%	42.0%	52.6%	57.1%
DT-Rank	41.4%	39.4%	27.3%	34.4%	63.0%
GRU-2	63.3%	77.2%	48.9%	59.3%	68.6%
RvNN	73.7%	66.2%	74.3%	70.8%	83.5%
PPC	86.3%	82.0%	89.8%	83.7%	84.3%
GLAN	90.2%	92.1%	86.9%	96.8%	84.7%
Transformer-GAT	89.2%	90.5%	86.7%	85.7%	93.5%
TPE-GAT	90.8%	90.4%	88.3%	92.2%	92.1%

表 4 Weibo 数据集上的实验结果
Table 4 Results on Weibo datasets

模型	类别	准确率	精准率	召回率	F1 值
DTC	FR	83.1%	84.7%	81.5%	83.5%
	NR		81.5%	84.7%	83.0%
SVM-TS	FR	85.7%	83.9%	88.5%	86.1%
	NR		87.8%	83.0%	85.7%
DT-Rank	FR	73.2%	73.8%	71.5%	72.6%
	NR		72.6%	74.9%	73.7%
GRU-2	FR	91.0%	87.6%	95.6%	91.4%
	NR		95.2%	86.4%	90.6%
PPC	FR	92.1%	89.6%	96.2%	92.3%
	NR		94.9%	88.9%	91.8%
GLAN	FR	94.6%	94.3%	94.8%	94.5%
	NR		94.9%	94.3%	94.6%
Transformer-GAT	FR	93.4%	94.0%	92.8%	93.4%
	NR		92.8%	94.0%	93.4%
TPE-GAT	FR	96.0%	96.5%	95.5%	96.0%
	NR		95.2%	97.0%	96.1%

(1) 从表 2~表 4 可以看出, 在基准模型中, SVM-TS 模型因能捕获更多的时间和结构特征, 在 Twitter15、Twitter16 和 Weibo 数据集上分别达到了 54.4%、57.4%和 85.7%的准确率, 在基于传统机器学习的模型中取得了较好的效果。相对于基于深度学习的模型, 基于传统机器学习的模型性能明显不足, GLAN 模型同时获取局部语义信息和全局结构信息, 在性能上取得了明显的提升。这进一步证明了基于深度学习的模型可以自动学习到潜在特征, 提升谣言检测的效果。而本文模型在前人研究的基础上, 改进了现有的 Transformer, 在 Twitter15、Twitter16 和 Weibo 数据集上分别达到了 91.4%、90.8%和 96.0%的准确

率, 比 SVM-TS 模型分别提升了 37.0%、33.4% 和 10.3%, 比 GLAN 模型分别提升了 0.9%、0.6% 和 1.4%。此外, 本文模型的 F1 值均高于基准模型的最优值。实验结果证明了本文模型的有效性。

(2) 从表 2~表 4 可以看出, 进行位置改进后的模型 TPE-GAT 相比 Transformer-GAT, 在 Twitter15、Twitter16 和 Weibo 数据集上分别提高了 2.3%、1.6% 和 2.6% 的准确率。实验结果证明了位置改进的有效性。

(3) Transformer 由于参数量较大, 在小型数据集上的表现稍差。与 GLAN 相比, 本文模型在 Weibo、Twitter15、Twitter16 数据集上的提升递减, 其原因可能是 Transformer 的数据依赖性。

综上所述, 本文模型能在社交媒体谣言检测问题上表现出更好的效果, 本文对 Transformer 的位置改进具有有效性。

3.3 消融实验

为了进一步验证本文位置改进的贡献, 设计了 5 组模型变体进行消融实验。

(1) TPE₀-GAT 使用式(5)作为注意力得分计算范式。

(2) TPE₁-GAT 使用式(11)中(e)+(f)+(g), 不使用可学习 w'_m 作为注意力得分计算范式。

(4) TPE₂-GAT 使用式(11)中(e)+(f)+(g), 使用可学习 w'_m 作为注意力得分计算范式。

(4) TPE-GAT 使用式(11)作为注意力得分计算范式。

(5) Only TPE 只使用 TPE 获取文本信息而不使用结构信息进行谣言检测。

从表 5 可以看出, 对于谣言检测, 文本信息是有效信息, 仅使用文本信息就可以达到较高的准确率。因此, 改进 Transformer 以更好地获取文本信息对于社交媒体谣言检测问题具有重要意义。本文提出的 3 种改进都具有有效性。在 Twitter15、Twitter16 和 Weibo 数据集上, TPE₁-GAT 模型分别高于 TPE₀-GAT 模型 3.1%、2.2% 和 1.9% 的准确率, 这说明相对位置的改进能较大幅度提升 Transformer 对于文本信息的获取能力。TPE₂-GAT 模型分别高于 TPE₁-GAT 模型 -0.3%、0.4% 和 1.1% 的准确率, 这说明精确的距离信息可以小幅度提升 Transformer 对于文本信

息的获取能力, 同时可学习的参数由于初始化的不稳定性, 有进一步改进的空间。TPE-GAT 模型分别高于 TPE₂-GAT 模型 0.5%、0.6% 和 0.5% 的准确率, 这说明绝对位置信息也可以小幅度提升 Transformer 对于文本信息的获取能力。

表 5 Weibo、Twitter15 和 Twitter16 数据集上的消融实验结果

Table 5 The ablation study results on the Weibo, Twitter15, Twitter16 datasets

模型	准确率		
	Twitter15	Twitter16	Weibo
TPE ₀ -GAT	88.1%	87.6%	92.5%
TPE ₁ -GAT	91.2%	89.8%	94.4%
TPE ₂ -GAT	90.9%	90.2%	95.5%
TPE-GAT	91.4%	90.8%	96.0%
Only TPE	75.9%	72.1%	82.0%

综上, 本文提出的 3 点具体改进可以提升 Transformer 对于文本序列位置信息的获取能力, 且可以获取更准确的语义信息从而提升 Transformer 在社交媒体谣言检测问题上的性能。

4 结束语

本文提出了一种基于改进 Transformer 的社交媒体谣言检测方法, 通过对 Transformer 进行相对位置和绝对位置的改进, 提升 Transformer 获取微博文本信息的能力。此外, 该方法利用图注意力网络获取发布者资料与传播特征, 并将信息融入微博的文本表示之中。在 Twitter15、Twitter16 和 Weibo 这 3 个公开的数据集中, 与基准方法相比, 本文方法取得了更高的正确率和 F1 值, 验证了本文方法在社交媒体谣言检测问题上的有效性。通过消融实验, 也进一步验证了本文对 Transformer 位置改进能明显提升 Transformer 获取微博文本信息的能力。

根据实验发现, 由于 Transformer 参数量大、对数据的依赖性强, 在小数据集上效果一般。因此, 下一步将研究如何在保证模型效果的前提下, 减少模型的参数。

参考文献:

- [1] LIU Z Y, ZHANG L, TU C C, et al. Statistical and semantic analysis of rumors in Chinese social media[J]. Scientia Sinica, 2015. 45(12): 1536-1546.

- [2] MIKOLOV T, SUTSKEVER L, CHEN K, et al. Distributed representations of words and phrases and their compositionality[J]. Advances in Neural Information Processing Systems, 2013, (26): 3111-3119.
- [3] RUCHANSKY N, SEO S, LIU Y. CSI: a hybrid deep model for fake news detection[J]. 2017: arXiv: 1703.06959.
- [4] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017: 6000-6010.
- [5] YUAN C Y, MA Q W, ZHOU W, et al. Jointly embedding the local and global relations of heterogeneous graph for rumor detection[C]//Proceedings of 2019 IEEE International Conference on Data Mining (ICDM). 2019: 796-805.
- [6] 据心怡. 基于深层双向 Transformer 编码器的早期谣言检测[J]. 信息通信, 2020, 33(5): 17-22.
- QU X Y. Early rumor detection based on deep two-way Transformer encoder[J]. Information & Communications, 2020, 33(5): 17-22.
- [7] MA J, GAO W, WONG K F. Detect rumors in microblog posts using propagation structure via kernel learning[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. 2017: 708-717.
- [8] MA J, GAO W, MITRA P, et al. Detecting rumors from microblogs with recurrent neural networks[C]// Proceedings of International Joint Conference on Artificial Intelligence. 2016.
- [9] SHAW P, USZKOREIT J, VASWANI A. Self-attention with relative position representations[J]. arXiv preprint arXiv:1803.02155, 2018.
- [10] 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.
- [11] LIU Y H, OTT M, GOYAL N, et al. RoBERTa: a robustly optimized BERT pretraining approach[J]. arXiv preprint arXiv:1907.11692, 2019.
- [12] YANG Z L, DAI Z H, YANG Y M, et al. XLNet: generalized autoregressive pretraining for language understanding[J]. CoRR, 2019, abs/1906.08237.
- [13] YAN H, DENG B C, LI X N, et al. TENER: adapting transformer encoder for name entity recognition[J]. arXiv preprint arXiv:1911.04474, 2019.
- [14] DAI Z H, YANG Z L, YANG Y M, et al. Transformer-XL: attentive language models beyond a fixed-length context[J]. arXiv preprint arXiv:1901.02860, 2019.
- [15] HE P C, LIU X D, GAO J F, et al. DeBERTa: decoding-enhanced BERT with disentangled attention[J]. arXiv preprint arXiv:2006.03654, 2020.
- [16] KE G L, HE D, LIU T Y. Rethinking the positional encoding in language pre-training[J]. arXiv preprint arXiv:2006.15595, 2020.
- [17] WANG B Y, ZHAO D H, LIOMA C, et al. Encoding word order in complex embeddings[J]. arXiv preprint arXiv:2006.15595, 2020.
- [18] LIU Y, WU Y F. Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks[C]//Proceedings of Thirty-Second AAAI Conference on Artificial Intelligence. 2018.
- [19] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv:1301.3781, 2013.
- [20] SMITH L N. Cyclical learning rates for training neural networks[C]//Proceedings of 2017 IEEE Winter Conference on Applications of Computer Vision (WACV). 2017: 464-472.
- [21] CASTILLO C, MENDOZA M, POBLETE B. Information credibility on twitter[C]//Proceedings of the 20th International Conference on World Wide Web-WWW '11. 2011: 675-684.
- [22] MA J, GAO W, WEI Z Y, et al. Detect rumors using time series of social context information on microblogging websites[C]// Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. 2015: 1751-1754.
- [23] ZHAO Z, RESNICK P, MEI Q. Enquiring minds: early detection of rumors in social media from enquiry posts[C]//Proceedings of International World Wide Web Conferences Steering Committee. 2015.
- [24] MA J, GAO W, WONG K F. Rumor detection on twitter with tree-structured recursive neural networks[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. 2018: 1980-1989.

[作者简介]



郑洪浩（1992-），男，山东济宁人，信息工程大学硕士生，主要研究方向为自然语言处理。



郝一诺（1997-），女，江苏徐州人，信息工程大学硕士生，主要研究方向为无线物理层安全、自然语言处理。



于洪涛（1970-），男，辽宁丹东人，博士，信息工程大学研究员、博士生导师，主要研究方向为网络大数据分析与处理。



李邵梅（1982-），女，湖北钟祥人，博士，信息工程大学副研究员，主要研究方向为计算机视觉。



吴翼腾（1992-），男，山东乐陵人，信息工程大学博士生，主要研究方向为人工智能安全。