基于社会背景信息的多任务谣言检测方法

王梦园 [1,2]
WANG Mengyuan

摘 要 随着互联网的迅猛发展，社交媒体如微信、Twitter 等呈现出飞速增长的趋势。一些居心叵测的人利用社交媒体平台制造并传播谣言，对个人、国家和社会产生了负面影响。因此，研究自动化的谣言检测方法具有重要价值。现有研究发现，在谣言检测过程中，不仅可以利用谣言及其相关推文的文本内容，还可将谣言传播过程中产生的一系列社会背景信息（如用户信息、转发次数和点赞数等）作为识别谣言的依据。与此同时，为克服数据量不足对模型性能的影响，多任务学习应运而生。多任务学习旨在通过对共享参数进行训练，实现任务间的信息交互，从而提高各任务的性能。在多任务学习和相关社会背景信息研究的启示下，将基于文本内容的谣言检测任务和基于社会背景信息的谣言检测任务相结合，构建了一种基于社会背景信息的多任务谣言检测模型，以期提升谣言检测效果。此外，还探讨了如何利用多任务学习技术优化谣言检测模型，以及如何从社会背景信息中提取更有价值的特征来提高模型的准确性。

关键词 谣言检测；社会背景信息；多任务学习

doi：10.3969/j.issn.1672-9528.2024.05.040

0 引言

互联网的发展使得社交媒体迅猛发展，极大方便了人们的日常生活，尤其是在信息获取与传递方面。然而，这一领域的开放性也为一些不怀好意的人提供了散布虚假信息的途径，以达到不可告人的目的。互联网的快速传播使这些虚假信息在短时间内影响广泛，给个人和社会带来严重的负面影响。例如，2011 年日本福岛核电站泄漏事件引发的国内抢盐风波，不仅浪费了公众的财产，还扰乱了市场秩序，对社会产生了深远的影响。因此，研究互联网上的谣言自动检测具有重要的理论价值和实际意义。

现有的利用社会背景信息的谣言检测研究一般采取的是传统机器学习方法，例如文献 [1] 使用用户信用信息及其对账单的评论信息完成谣言检测任务，文献 [2] 使用推文转发者序列的个人用户信息作为用户特征等。

随着研究的不断深入，多任务谣言检测方法逐渐流行。例如文献 [3] 中将谣言检测任务和立场检测任务通过共享参数的形式组合成一个多任务谣言检测模型来加强本身的检测效果。而文献 [4] 则是在文献 [3] 的基础之上使用了注意力机制和视觉特征来进行谣言检测。但这些方法存在任务之间关联性较弱的缺点。本文模型在训练过程中实现了任务间更深入的信息交互，并采用注意力机制关注关键信息、降低噪声干扰，从而实现更好的检测效果。后续在 Pheme 数据集上的实验结果也证明了本文模型的有效性。

1 基于社会背景信息的多任务谣言检测模型

1.1 基于社会背景信息的多任务谣言检测问题定义

本文将社会背景信息定义为：与谣言的文本内容特征相关的信息特征，其中包含谣言发布者以及评论转发者的个人信息，如个人描述、是不是认证用户等。同时，还包括推文本身有关的信息，如点赞数、转发数等。

在多任务谣言检测方法的启发下，本文模型将基于文本内容的谣言检测任务和基于社会背景信息的谣言检测任务通过多任务的形式组合在一起，形成了基于社会背景信息的多任务谣言检测方法。同时，使用层级注意力机制来减少噪声干扰，从而加快模型收敛速度，提高检测精度。本文将整个数据集划分为 $\{C_1, C_2, ..., C_{m-1}, C_m\}$，其中 $C_i=\{(X_0, X_1, ..., X_t), (D_0, D_1, ..., D_t), (I_0, I_1, ..., I_t)\}$，$X_0$ 代表谣言初始文字的特征向量表示，而 X_1-X_t 为对其的评论或者转发时的推文的特征向量表示，D_i 是其对应的发帖用户的个人描述的特征向量表示，I_i 是其对应除了用户个人描述外的其他离散的较为重要的社会背景信息特征，C_i 表示本文模型的一个检测对象，本文模型的目的就是判断一个 C_i 是不是谣言。

1. 三峡大学湖北省水电工程智能视觉检测重点实验室 湖北宜昌 443002
2. 三峡大学计算机与信息学院 湖北宜昌 443002
[基金项目] NSFC- 新疆联合基金重点项目（U1703261）

1.2 基于社会背景信息的多任务谣言检测模型结构

相较于基于内容的单任务学习模型来说，本文所提出的基于社会背景信息的多任务谣言检测模型可以利用相关任务，学习到在基于内容的检测任务或基于社会背景信息的检测任务单独训练和学习时所忽略的信息，因为其是从社会背景信息和文本内容两个角度对谣言进行分析，从而形成了两个特征空间。在这两个特征空间中，差异越大的两个推文之间的距离也就越大。本文的两个任务通过共享参数的方式在训练的过程中进行信息交互，从而在本模型架构中通过任务之间的信息共享来提升检测效果。例如，基于文本内容的检测任务可以学习到谣言的社会背景信息并将其应用在本身的检测任务中，反之亦然。与此同时，在共享空间中，那些具有强烈虚假意义的内容特征很大概率会被投影到具有虚假意义的社会背景信息特征附近，真实信息也是如此。

本文所提出的基于社会背景信息的多任务谣言检测模型与现有的多任务谣言检测模型的不同主要在于：（1）构成模型的两个任务都是基本的谣言检测任务，一个是基于社会背景信息的谣言检测任务，一个是基于文本内容的谣言检测任务，而目前的多任务谣言检测模型大多是将立场检测任务和谣言检测任务组合在一起，共同构成了多任务谣言检测模型；（2）本文模型两个任务所使用的数据集是同一个数据集，分别从社会背景信息和文本内容两个角度去分析谣言的特征；（3）为了能够重点关注那些对谣言检测更加重要的单词和句子，从而起到加速模型收敛、减少噪声干扰以及提升检测精度的作用，本文还使用了层级注意力机制 [5]。

本文所提出的模型总体结构如图 1 所示。其中，基于社会背景信息的谣言检测任务是利用从数据中学习谣言及其回复的社会背景信息特征来捕捉谣言，而基于文本内容的谣言检测任务则是通过文本内容来识别推文是不是谣言，两个任务通过共享参数的方式进行信息交互和共享。

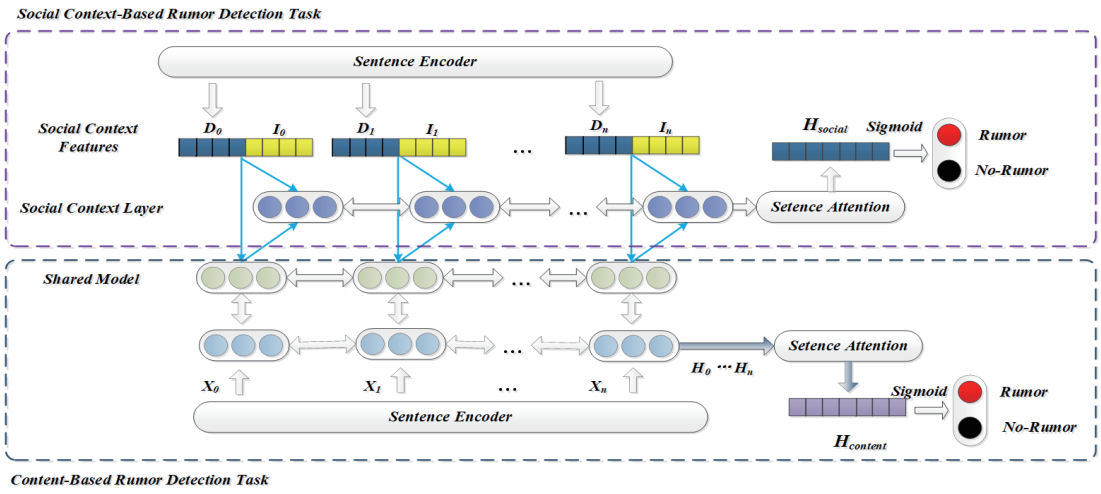

图 1 基于社会背景信息的多任务谣言检测模型结构

其中，Sentence Encoder 通过 Word Attention 对推文或者个人用户描述经过 Word2Vec 编码的词向量矩阵得到的特征向量表示，具体结构如图 2 所示。其中，Word1-Wordn 为一个句子中各个词对应的词向量。

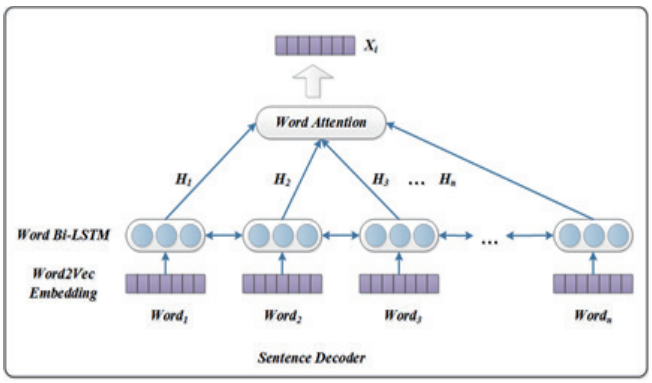

图 2 Sentence Encoder

2 相关实验及分析

2.1 训练过程

在训练时，两个基本的谣言检测任务同时进行训练，并且均使用交叉熵损失函数，因此需要同时计算两个任务各自的损失函数。该模型的损失函数为：

$$Loss_s = -\left(Y_s \cdot log\left(Y_{ts}\right) + \left(1 - Y_s\right) \cdot log\left(1 - Y_{ts}\right)\right)$$

$$Loss_c = -\left(Y_c \cdot log\left(Y_{tc}\right) + \left(1 - Y_c\right) \cdot log\left(1 - Y_{tc}\right)\right) \quad (1)$$

$$Loss = \alpha_s \cdot Loss_s + \alpha_c \cdot Loss_c$$

式中：$Loss_s$ 为基于社会背景信息的检测任务的损失函数，α_s、α_c 分别为社会背景信息任务和文本内容任务损失函数的系数，$Loss_c$ 代表了基于文本内容检测任务的损失函数，$Loss$ 代表整个模型总体的损失函数，Y_s 为基于社会背景信息的检测任务得到的预测结果，Y_{ts} 是其对应的真实的结果，Y_c 和 Y_{tc} 分别为基于文本内容检测任务的预测结果和真实结果。

2.2 实验及结果

本文所使用的数据集为 Pheme 数据集 [6]。为了和其他模型进行比较，来证明本文模型的有效性，采用准确率（accuracy）、精确率（precision）、召回率（recall）以及 F_1 分数来对模型性能进行评价，其计算

公式分别为（2）～（5）。其中，TP 为真阳率，TN 为真阴率，FP 为假阳率，FN 是假阴率。

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \tag{2}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{3}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \tag{4}$$

$$F_1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

选择 9 个谣言检测模型进行对比，其中前 5 个模型为经典模型，而后 4 个是目前在 Pheme 数据集上表现较好的多任务谣言检测模型，其中各个算法模型的参数设置与原论文中的设置一致。对比算法如下。

DT-Rank[7]：采用决策树和基于统计特征的谣言检测模型。

DTC[8]：使用信息的可靠性作为检测依据的决策树分类器。

SVM-TS[9]：使用谣言的用户、内容、对社会的影响和传播模式作为检测依据的基于时间序列结果的 SVM 分类器。

CNN[10]：使用 CNN 网络对句子进行分类的一种检测方法。

LSTM-MTL[11]：通过硬参数共享的方式，将谣言检测任务和立场检测任务结合起来形成多任务学习模型。

GRU-MTL：通过较低的层来提取公共特征，并使用较高层次的层来提取各自任务特定的特征（见文献 [3]）。

Bayesian-DL[12]：首先使用贝叶斯分类器提取谣言的不确定性，然后使用 LSTM 对回复进行编码。

Trans-MTL[13]：通过 Transformer 模型的方法将立场分类和谣言检测两个任务组合在一起，形成一个多任务谣言检测模型。其通过选定的共享层来学习两个任务的共享信息，同时采用门限机制和注意力机制来选择任务之间的共享特征。

MM-MTL：将使用文本信息和视觉信息相结合，并使用 GRU-MTL 中的方式将立场分类任务和谣言检测任务结合起来，采取注意力机制将立场分类中的信息有选择地使用在谣言检测中，见文献 [4]。

实验结果如表 1 所示。可以看出，本文的模型（CS-HA-MTL）在 F_1 分数、准确率、精确率以及召回率上均高于目前较为先进的多任务谣言检测的结果。同时表明，本文的模型在 Pheme 数据集上具有较高的有效性。

表 1 Pheme 数据集实验结果

	F_1/%	Pre/%	Rec/%	Acc/%
DT-Rank	61.70	54.90	70.40	56.20
DTC	58.40	57.90	58.80	58.20
SVM-TS	66.30	64.20	68.60	65.10
CNN	73.09	72.95	73.52	73.65
LSTM-MTL	77.15	68.77	87.87	74.94
GRU-MTL	78.73	76.30	81.31	79.14
Bayesian-DL	78.78	78.29	79.29	80.33
Trans-MTL	80.09	73.41	88.10	81.27
MM-MTL	82.02	78.84	85.45	82.21
CS-HA-MTL	86.78	87.03	86.53	83.08

表 1 中展示了本文模型在 Pheme 数据集上的实验结果。从结果可以看到，本文所提出的基于社会背景信息的多任务谣言检测模型在四个指标上均已超过目前已有较为先进的谣言检测模型。并且从表 1 中可以得到以下观察。Pheme 数据集上的结果表明，绝大多数的基于深度学习的检测方法取得的检测效果要优于基于传统机器学习算法的检测方法。因此，传统机器学习的特征工程所提取到的特征往往是数据的浅层特征，而基于深度学习的检测算法可以自动提取到更加深层次的特征。绝大多数的基于深度学习的模型取得的结果优于基于传统机器学习算法的模型。大部分的多任务谣言检测模型的实验结果好于单任务学习的检测模型，因为多任务的检测模型在训练的过程中可以进行信息交互，从而可以利用其他任务的信息辅助本身进行检测，取得更好的检测效果。而本文模型中两个任务均为谣言检测任务，并从社会背景信息和文本内容两个角度对谣言进行分析，充分利用数据的信息，同时在训练过程中进行信息交互，实现信息共享，从而取得更好的模型表现力。

从具体的结果来看，本文基于社会背景的多任务谣言检测模型比现有表现最好的多任务谣言检测模型（MM-MTL）在准确率上高出 0.87%，并且在 F_1 分数、精确率和召回率上分别高出 4.76%、8.19%、1.08%。本文的多任务检测模型的两个任务都是基本的谣言检测任务，其相关性要高于其他多任务检测模型的任务之间的相关性，因此在模型训练过程中能够实现更加深入的信息交互。再者，本文的模型还使用了层级注意力机制来重点关注那些重点的单词和句子，在减少噪声干扰的同时也能加速模型的收敛速度，还能够提高模型的检测精度。

综上所述，再结合表 1 的实验结果可以得出，本文模型表现较好的原因有以下几方面。

（1）本文多任务模型的两个任务均是谣言检测任务，

其相关性要高于立场检测和谣言检测之间的相关性。

（2）在模型训练的过程中，基于社会背景信息的谣言检测任务和基于文本内容的谣言检测任务的信息得到了更加深入的交互，起到了更好的促进作用，因此能够提高模型的检测效果。

（3）本文模型同时使用文本内容和社会背景信息两种特征，从不同的角度对谣言进行分析，能够更加精准地捕捉谣言异常的文本和社会背景信息。

（4）层级注意力机制的使用使得模型能够重点关注那些对检测更为有用的信息而忽略不重要的信息，从而起到减少噪声干扰、加速模型收敛的作用，最终促使模型取得更好的实验结果。

3 结语

在本研究中，提出了一种创新的基于社会背景信息的多任务谣言检测模型（CS-HA-MTL）。该模型有效地利用图 1 所示的社会背景信息特征作为输入，并与基于文本内容的谣言检测任务共享参数，构建了一个高效的多任务框架。为了提高模型在复杂场景下的性能，本文采用了层级注意力机制，确保关键信息得到充分关注。在 Pheme 数据集上的对比实验结果表明，本文的模型在谣言检测领域具有显著优势，取得了优异的检测效果。

参考文献：

[1]LI Q, ZHANG Q, SI L. Rumor detection by exploiting user credibility information, attention and multi-task learning[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.Stroudsburg, PA:Association for Computational Linguistics,2019:1173-1179.

[2]LU Y, LI C. GCAN: graph-aware co-attention networks for explainable fake news detection on social media[EB/OL].(2020-04-24)[2024-01-27].https://arxiv.org/abs/2004.11648v1.

[3]MA J, GAO W, WONG K. Detect rumor and stance jointly by neural multi-task learning[C]//Companion Proceedings of the Web Conference 2018.New York:ACM,2018:585-593.

[4]ZHANG H, QIAN S, FANG Q, et al. Multi-modal meta multi-task learning for social media rumor detection[J]. IEEE transactions on multimedia, 2021,24:1449-1459.

[5]YANG Z, YANG D, DYER C, et al. Hierarchical attention networks for document classification[C]//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.Stroudsburg, PA:Association for Computational Linguistics,2016: 1480-1489.

[6]ZUBIAGA A, LIAKATA M, PROCTER R. Learning reporting dynamics during breaking news for rumour detection in social media[EB/OL].(2016-10-24)[2024-01-28].https://arxiv.org/abs/1610.07363v1.

[7]ZHAO Z, RESNICK P, MEI Q. Enquiring minds: early detection of rumors in social media from enquiry posts[C]// Proceedings of the 24th International Conference on World Wide Web.New York:ACM,2015:1395-1405.

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

[9]MA J, GAO W, WEI Z, 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. New York:ACM,2015:1751-1754.

[10]CHEN Y. Convolutional neural network for sentence classification[D]. Waterloo:University of Waterloo, 2015.

[11]KOCHKINA E, LIAKATA M, ZUBIAGA A. All-in-one: multi-task learning for rumour verification[EB/OL].(2018-06-10)[2024-01-30].https://arxiv.org/abs/1806.03713.

[12]ZHANG Q, LIPANI A, LIANG S, et al. Reply-aided detection of misinformation via bayesian deep learning[C]//The World Wide Web Conference 2019.New York:ACM,2019: 2333-2343.

[13]WU L, RAO Y, JIN H, et al. Different absorption from the same sharing: sifted multi-task learning for fake news detection[C]// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Stroudsburg, PA:Association for Computational Linguistics,2019: 4644-4653.

【作者简介】

王梦园（1999—），女，湖北随州人，硕士，研究方向：自然语言处理。

（收稿日期：2024-03-01）