中国人民大学

《Web 信息处理与应用》课程大作业

题目名称：知识赋能的健康虚假信息检测模型

学生姓名：刘　妃

学生学号：2019000170

专业年级：信息学院 2019 级

分　　数：

2020 年 6 月

目 录

一、 研究背景 2

二、 问题描述 4

三、 相关工作 5

3.1 虚假信息检测任务数据集 5

3.2 虚假信息检测方法 5

3.3 知识使能的算法模型 7

四、 数据集 9

五、 模型构建 10

六、 实验 12

6.1 数据集划分 12

6.2 baseline 选择 12

6.3 实验环境、超参数设置 12

6.4 实验结果 12

七、 总结 15

参考文献 16

一、研究背景

随着新媒体的大流行，网络环境积累了丰富的信息、数据、知识等，人们可以通过互联网随时随地吸收现代文明信息，缩小知识差距，然而，良莠不齐的信息实质上却在无形中扩大了这种距离。据调查，美国公民在社交媒体上获得的 65%的新闻均为不实信息。

通常，不实信息在社交网络中具有更快的传播速度，更深、更广的渗透能力，如流行病般疯狂蔓延在社交网络中。2003 年大卫·罗斯科夫（David J. Rothkopf）创造了新词“Infodemic”，英文全称为 Information Epidemic，译为“信息疫情”、“信息流行病”等，含义指“一些事实，混入恐惧、猜测和流言，被现代信息技术在世界范围内迅速放大和传递，以完全扭曲根本现实的方式在国家和国际范围内影响经济、政治甚至安全”。 麻省理工科技评论（MIT Technology Review）指出，2019 新冠病毒带来的是史上第一场社交媒体信息疫情。近年来，虚假信息检测（MID）已经成为一个严峻、迫切的研究课题，各大综合、专业顶尖期刊、会议都设立了 MID 研究专栏。

健康虚假信息作为网络不实信息的重要组成部分，有着最广泛的受众，81. 5%的美国公民会通过互联网搜索健康或医疗相关信息，同时 68. 9%的美国成人将互联网视为搜寻健康信息的第一渠道。老年群体作为最具潜力的网民，在网络使用时长、活跃度上均不断提升，是健康虚假信息的重灾区。健康虚假信息的渗透对公众健康构成重大威胁，同时增加了医疗系统的压力，如曾在 Facebook 引起全球轰动的

“生姜治癌”推文事件。

健康虚假信息比事件类的不实信息更难识别，需要大量的专业知识，同时健康虚假信息检测（HMID）对于模型可解释性的要求也更高，有助于公众对检测结果的接受，以及支持专家对信息的治理。当前对健康虚假信息的内容、传播特征已经有少量的研究，但 HMID 的研究，尤其是可解释的 HMID 还存在较大的空缺。基于此，本文旨在探索知识赋能的 HMID 模型构建。

二、问题描述

本文的主要任务为健康虚假信息的检测，即健康信息的真假判定，为二分类问题。

首先对健康虚假信息进行定义。现在广泛认可的错误信息分为：rumor、fake news、hoax、click-bait、disinformation 和 misinformation，各类型之间具有内容、产生机制、传播机制等的细微差别。本文将健康虚假信息定义为：网络流传的关于健康、医药、食品安全等，被查证为错误或不正确的信息。本文的目标即对网络上关于健康、医药、食品安全等相关信息真、假的自动化判定，如图 1 所示。

图 1 健康信息真、假判定

三、相关工作

3.1 虚假信息检测任务数据集

尽管虚假信息检测（MID）任务近几年已经得到了不断的重视，但是由于其数据获取的难度，仍旧未形成统一、标准、大规模的评测数据集。

MID 任务大多关注于具体事件性信息，如娱乐绯闻、政治行动、突发公共事件等，也有日常普通的全领域推文信息。研究人员根据课题需要主要通过 Twitter、Facebook、微博或者辟谣平台进行事件关键词检索获取数据。公开的，较具有代表性的数据集有 BuzzFeedNews（BuzzFeed journalist，美国大选）、LIAR（POLITIFACT. COM，政治言论、行为）、CREDBANK（众包，新闻事件）、BuzzFace（BuzzFeed journalist，新闻事件）、FacebookHoax 等。

健康虚假信息检测（HMID）任务还处于探索起步阶段，现有的相关研究较少，数据集有 Twitter、在线健康社区等获取的具体医疗事件、突发公共卫生事件、药物副作用信息等。由于健康信息的难以辨认，健康虚假信息数据集的构建具有更高的门槛，健康防护、医药治疗、食品安全等相关数据集尚未形成。同时由于我国中药养生的特色，中文 HMID 任务数据可能具有独特的特征。

3.2 虚假信息检测方法

虚假信息检测算法按照特征不同，主要可以分为四类：

A. 基于内容的检测，包括信息文本、链接、图片、音频视频等，多源异构数据融合的检测是新兴的研究内容；

B. 基于环境的检测，包括信息传播、交互特征，用户网络，话题、事件、立场等信息，是现在主流的检测方法；

C. 基于知识的检测，包括专家、众包、知识库、知识图谱的知识应用，还处于萌芽探索阶段；

D. 特征融合的检测，即应用上述三类特征进行综合检测，是实际生产的主要方法。

虚假信息检测算法按照模型结构不同，主要分为五类：

A. 传统机器学习，主要依赖于具体数据集的特性，通过特征工程，构建有效的特征和分类算法，过分依赖于数据集本身及特征构建；

B. 深度学习，端到端的深度学习算法在众多的分类任务中刷新了传统机器学习算法的记录，在 MID 任务中同样取得很好的效果，是现在主流的研究内容；

C. 图算法，包括传统的随机游走、链接预测等图论算法，以及最新流行的图神经网络算法，在知识图谱嵌入的 MID 任务中可进行应用，尚处于探索阶段；

D. 人工参与，包括专家决策、众包、规则筛选等，在实际应用中起到关键作用；

E. 方法融合，即上述四类方法的融合检测。

分析虚假健康信息的特点：较少具有个人观点、立场；非事件性

信息，通常不具有短时的爆发；目标群体广泛；主要关注于信息的核心知识，而不是全文的信息，同时信息为专业领域知识。由此我们认为，基于社交环境的检测方法可能不可行，基于知识的检测方法辅助信息内容或许可以获得较好效果。

3.3 知识使能的算法模型

知识与信息内容融合的方法，我们称之为知识使能的方法，在推荐、问答领域已有较多的研究，主要为知识库、知识图谱的应用，按照应用方式不同可分为三类：

A. 基于知识图谱嵌入：知识图谱向量化研究已取得较多的成果，现在主流的如 transE、transD 等 trans 系列模型，能够对实体、关系进行较好的表示建模。基于此，在推荐、问答领域研究人员们对 word embedding 和 KG embedding 的融合进行了大量的探索，包括普通的拼接、CNN 多通道建模、attention 融合、语言模型预训练嵌入等；

B. 基于知识图谱路径：基于知识图谱路径学习在推荐任务中研究较多，主要是因为具体的推荐任务，实体间关系较为简单、清晰，利于元路径的构建。研究人员们对路径建模的 CNN、RNN 等方法均进行了探索实践，对模型可解释性研究具有重要意义；

C. 基于知识图谱的文本丰富：基于知识图谱的文本内容丰富讨论较少，主要思想即为通过知识图谱的实体、关系链接，

对原始文本进行扩增，可视为一种信息增强方法，如何有效得抽取知识、加入知识，在丰富原始文本的基础上，避免信息混淆和信息干扰，是研究的重点与难点，主要的代表模型有 Liu 等提出的 K-BERT 模型。

本文主要进行了基于知识图谱的文本丰富方法探索应用。

四、数据集

虚假信息检测数据集的构建方法主要分为两种：自顶向下和自底向上。自底向上主要为按关键词检索，专家、众包标注的方法；自顶向下主要为按标签进行获取的方法。健康信息的识别需要更多的专家知识，自底向上的方法对于构建较为丰富的数据集有较大的限制与挑战，因此本文采用自顶向下的构建方法，为了保证数据集的质量和可信度，对健康信息网络发布平台进行了广泛的检索，发现专业辟谣平台的数据可以较好的支持数据集构建和后续研究。

通过对国内辟谣平台的筛选，本文选取腾讯较真为最终的数据收集平台。腾讯较真辟谣平台是一个多形式运维的辟谣平台，具有较高的流量，信息发布及时，包含食品安全和医疗健康两个栏目，与本文的研究问题一致。通过 fiddler 对腾讯较真小程序进行抓包，解析其请求和响应报文，对两个栏目的所有数据进行了采集，经去重等清洗，最终获得 4283 条判定为“假”信息和 1326 条判定为“真”信息。

为保证较为充分的数据集，通过采集腾讯医典小程序的概述、日常、预防栏目内容补充真实信息，最终得到 4167 条“假信息”，数据集总数为 8450 条。数据示例如表 1 所示。

表 1 数据示例

Text	Label	来源
蘑菇汤竟然是高血压、高血糖、高血脂的天敌	0	腾讯较真
喝苏打水会增加钠的摄入	1	腾讯较真
高血压避免或尽量减少饮酒和含酒精饮品	1	腾讯医典

五、模型构建

模型为基于 Bert 模型的改进，核心模块包括：knowledge layer、embedding layer、seeing layer 和 mask-transformer encoder，如图 2 所示。

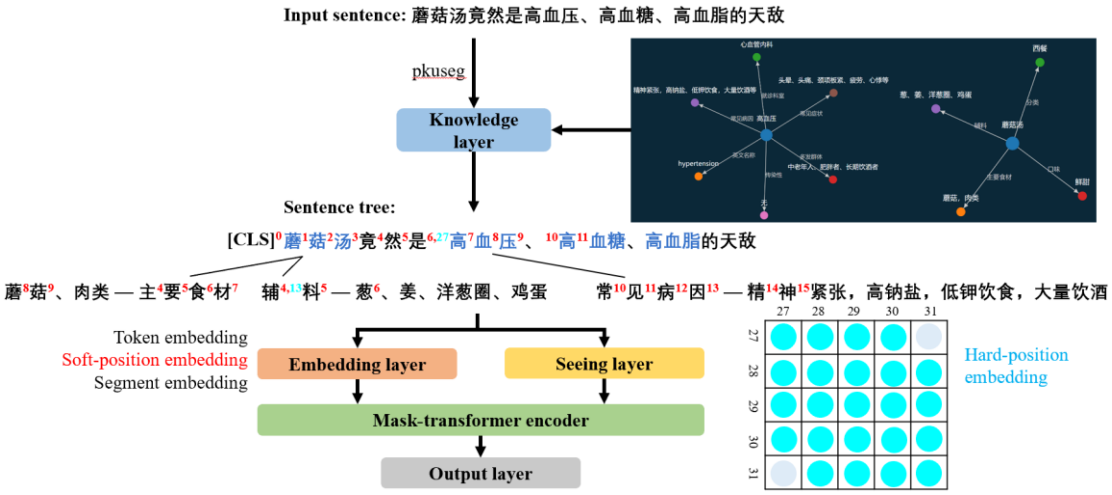

图 2 模型架构

Knowledge layer 用于将知识图谱注入到句子中并完成句子树的转换，具体分为两个步骤：knowledge query 和 knowledge injection。其中 knowledge query 分为两个步骤：对句子使用 pkuseg 进行分词，都分词后的每个 word 去查询知识图谱中相关的三元组，得到句子相关三元组集合。Knowledge injection 把对应的三元组注入到句子中形成句子树，每个 word 可能对应有多条三元组。

Embedding layer 继承 Bert 的结构，分为 token embedding、position embedding 和 segment embedding。其中 token embedding 和 segment embedding 与 Bert 一致，position embedding 采用基于树的深度优先编码，称为 soft-position embedding，通过基于树的深度优先编码保

留句子树的结构信息。

Seeing layer 主要服务于 mask-transformer attention，通过一个矩阵来存储 word 的分支关系，在同一分支上的为可见信息，不同分支上的为不可见信息，矩阵索引为句子的硬编码，即顺序编码。

Mask-transformer layer 为原始 Bert 的 self-attention transformer 的改进，不可见信息在计算隐层状态时没有贡献。借助 seeing layer 构造的矩阵，重构 softmax 的计算，如下公式所示。

$$M_{ij} = \begin{cases} 0 & w_i \ominus w_j \\ -\infty & w_i \oslash w_j \end{cases}$$

$$Q^{i+1}, K^{i+1}, V^{i+1} = h^i W_q, h^i W_k, h^i W_v$$

$$S^{i+1} = softmax(\frac{Q^{i+1} {K^{i+1}}^\top + M}{\sqrt{d_k}})$$

$$h^{i+1} = S^{i+1} V^{i+1}$$

总的来说，模型巧妙利用了 Bert 结构的特征，将知识注入原始文本，起到信息增强的效果，实现知识赋能，通过利用 soft-position 和 seeing layer 有效规避了信息混淆，模型可以嫁接各类 Bert 模型和知识图谱，同时可扩展应用于各类下游任务。

六、实验

6.1 数据集划分

数据集正负样例基本平衡，按照数据集大小，划分 6:2:2 的训练集、验证集和测试集。

6.2 baseline 选择

为充分验证本文提出模型的有效性，选取不同结构的经典模型作为 baseline，包括 TextCNN、BiLSTM、FastText、Transformer 和 Bert 模型。

6.3 实验环境、超参数设置

实验采用 Pytorch 框架，在一块 TITAN V 显卡下运行。

Bert 模型采用 Google Bert 的中文预训练模型，知识图谱采用一个医学知识图谱（13864 条三元组）和 CnDbpedia（517 万条三元组）。

Baseline 除 Bert 外，batch size 为 128，Bert 和本文的模型 batch size 设为 16，均训练 20 个轮次，采用交叉熵作为损失函数，embedding length、learning rate 和 dropout 根据具体模型设定。本文模型的 sequence length 为 256，learning rate 为 2e−5，dropout 为 0.5。

6.4 实验结果

采用 precision、recall、f1 分别对正负标签进行评价，使用 accuracy

作为整体模型效果的评价，得到模型结果如表 2 所示。可见本文模型在正负样例和各项评价指标上均具有突出优势，accuracy 相较 Bert 模型提升了 3 个百分点。

表 2 实验结果

Models	T			F			Acc.
	P.	R.	F1	P.	R.	F1	
TextCNN	0.8399	**0.9242**	0.8800	**0.9131**	0.8189	0.8635	0.8723
BiLSTM	0.8699	0.8973	0.8834	0.8910	0.8621	0.8763	0.8800
FastText	0.8649	0.8891	0.8769	0.8827	0.8573	0.8698	0.8734
Transformer	0.8335	0.8763	0.8544	0.8658	0.8201	0.8424	0.8486
Bert	0.882	0.853	0.867	0.861	0.889	0.875	0.8711
K-Bert(Medical)	0.924	0.888	**0.906**	0.895	0.929	**0.912**	**0.9089**
K-Bert(CnDbpedia)	**0.925**	0.885	0.904	0.892	**0.930**	0.911	0.9077

同时本文提出的模型与 Bert 模型相比在模型复杂度上基本一致，模型大小无明显差别，对训练过程 loss 变化和 accuracy 变化进行分析（图 3、4），可以得到，二者的收敛速度基本一致。医学知识图谱和 CnDbpedia 效果差异不大，主要是医学知识图谱与本任务具有更高的相关性，但是图谱大小和关系类型都还不够丰富，可以考虑对医学知识图谱和通用知识图谱的融合丰富。

为验证 seeing layer 的有效性，对不使用 seeing layer 的模型进行实验，相同参数设置下，模型收敛效果和准确率都有所降低。

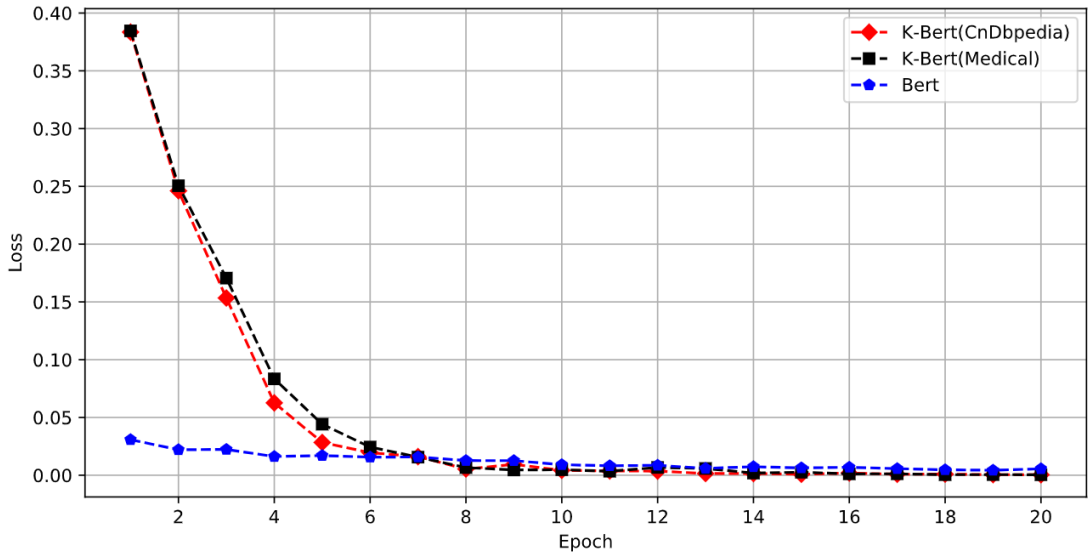

图 3 训练过程 loss 变化

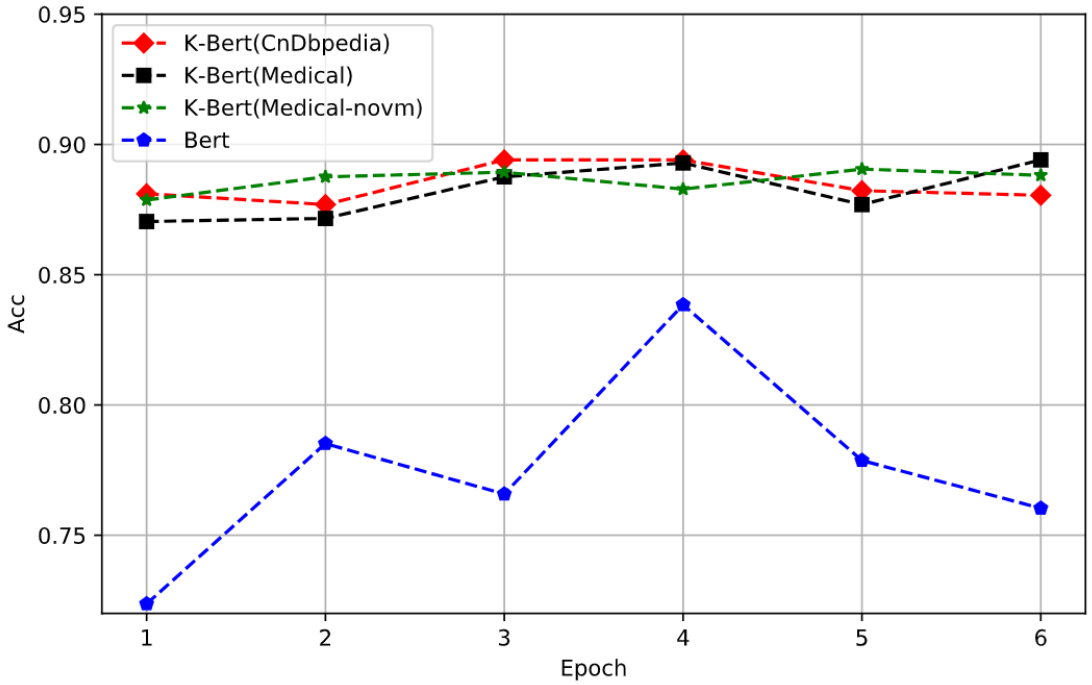

图 4 训练过程 accuracy 变化

七、总结

本文对健康虚假信息检测任务进行了特征分析和文献梳理，构建了一个可信的、高质量的数据集，可用于后续的研究。

本文对知识赋能的检测模型进行了探索实验，验证了基于知识图谱对文本内容进行知识增强的检测方法在本任务中能够取得较为突出的效果，同时实验也一定程度上验证了本文构建数据集的质量。

同时本文模型也存在一些不足，比如 knowledge query 时没有对实体进行选择和对齐，缺少对内容中关系信息的增强，仅对一跳关系进行了嵌入，没有考虑远距离的知识；对比实验中未对 kg embedding 的方法进行比对。在未来的工作中将对这些问题进行进一步尝试解决，提出创新性的检测算法。

对于 HMID 任务的数据集，也可以通过日常积累，进一步丰富，搜集具有多特征的信息，以丰富数据集。

参考文献

[1] M. Verma and D. Ganguly, “LiRME: Locally interpretable ranking model explanation,” *SIGIR 2019 - Proc. 42nd Int. ACM SIGIR Conf. Res. Dev. Inf. Retr.*, vol. 1, no. 1, pp. 1281–1284, 2019, doi: 10.1145/nnnnnnn.nnnnnnn.

[2] G. Gorrell *et al.*, “SemEval-2019 Task 7: RumourEval, Determining Rumour Veracity and Support for Rumours,” pp. 845–854, 2019, doi: 10.18653/v1/s19-2147.

[3] Q. Li, Q. Zhang, L. Si, and Y. Liu, “Rumor Detection on Social Media: Datasets, Methods and Opportunities,” pp. 66–75, 2019, doi: 10.18653/v1/d19-5008.

[4] S. Hamidian and M. T. Diab, “Rumor Detection and Classification for Twitter Data,” 2019, [Online]. Available: http://arxiv.org/abs/1912.08926.

[5] S. Mukherjee, G. Weikum, and C. Danescu-Niculescu-Mizil, “People on drugs: Credibility of user statements in health communities,” *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, pp. 65–74, 2014, doi: 10.1145/2623330.2623714.

[6] A. Kinsora, K. Barron, Q. Mei, and V. G. V. Vydiswaran, “Creating a Labeled Dataset for Medical Misinformation in Health Forums,” *Proc. - 2017 IEEE Int. Conf. Healthc. Informatics, ICHI 2017*, pp. 456–461, 2017, doi: 10.1109/ICHI.2017.93.

[7] Y. Wang and Q. Chen, “Knowledge Base Question Answering System Based on Knowledge Graph Representation Learning,” pp. 170–179.

[8] X. Wang, D. Wang, C. Xu, X. He, Y. Cao, and T.-S. Chua, “Explainable Reasoning over Knowledge Graphs for Recommendation,” *Proc. AAAI Conf. Artif. Intell.*, vol. 33, no. November, pp. 5329–5336, 2019, doi: 10.1609/aaai.v33i01.33015329.

[9] E. Dai, Y. Sun, and S. Wang, “Ginger Cannot Cure Cancer: Battling Fake Health News with a Comprehensive Data Repository,” 2020, [Online]. Available: http://arxiv.org/abs/2002.00837.

[10] N. Hassan, F. Arslan, C. Li, and M. Tremayne, “Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster,” *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, vol. Part F129685, pp. 1803–1812, 2017, doi: 10.1145/3097983.3098131.

[11] Y. Hao *et al.*, “An end-to-end model for question answering over knowledge base with cross-attention combining global knowledge,” *ACL 2017 - 55th Annu. Meet. Assoc. Comput. Linguist. Proc. Conf. (Long Pap.*, vol. 1, pp. 221–231, 2017, doi: 10.18653/v1/P17-1021.

[12] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu, “FakeNewsNet: A Data Repository with News Content, Social Context and Spatialtemporal Information for Studying Fake News on Social Media,” 2018, doi: 10.1089/big.2020.0062.

[13] T. Mitra and E. Gilbert, “CREDBANK: A large-scale social media corpus with associated credibility annotations,” *Proc. 9th Int. Conf. Web Soc. Media, ICWSM 2015*, pp. 258–267, 2015.

[14] 魏海斌，苏菁涵，吴一波，郑智源, “大数据背景下健康辟谣现状研究述评,” 卫生软科学, pp. 6–10, 2019.

[15] C. Journal and O. F. Computers, “‹¡{— g : [fb,” 2017.

[16] J. Zhou, F. Liu, and H. Zhou, “Understanding health food messages on Twitter for health literacy promotion,” *Perspect. Public Health*, vol. 138, no. 3, pp. 173–179, 2018, doi: 10.1177/1757913918760359.

[17] W. Y. Wang, “‘Liar, liar pants on fire’: A new benchmark dataset for fake news detection,”

ACL 2017 - 55th Annu. Meet. Assoc. Comput. Linguist. Proc. Conf. (Long Pap., vol. 2, pp. 422–426, 2017, doi: 10.18653/v1/P17-2067.

[18] V. Qazvinian, E. Rosengren, D. R. Radev, and Q. Mei, "Qazvinian et al. - 2011 - Rumor has it Identifying Misinformation in Microblogs(2)," *Conf. Empir. Methods Nat. Lang. Process.*, pp. 1589–1599, 2011.

[19] L. Shi, S. Li, X. Yang, J. Qi, G. Pan, and B. Zhou, "Semantic Integration of Heterogeneous Medical Knowledge and Services," *Res. Artic. Semant. Heal. Knowl. Graph*, vol. 2017, pp. 8–10, 2017, doi: 10.1155/2017/2858423.

[20] Z. Zhang, X. Han, Z. Liu, X. Jiang, M. Sun, and Q. Liu, "ERNIE: Enhanced Language Representation with Informative Entities," pp. 1441–1451, 2019, doi: 10.18653/v1/p19-1139.

[21] R. Sicilia, S. Lo Giudice, Y. Pei, M. Pechenizkiy, and P. Soda, "Twitter rumour detection in the health domain," *Expert Syst. Appl.*, vol. 110, pp. 33–40, 2018, doi: 10.1016/j.eswa.2018.05.019.

[22] W. Liu *et al.*, "K-BERT: Enabling Language Representation with Knowledge Graph," 2019, [Online]. Available: http://arxiv.org/abs/1909.07606.

[23] Y. Jiang, Y. Liu, and Y. Yalin, "LanguageTool based University rumor detection on Sina Weibo," *2017 IEEE Int. Conf. Big Data Smart Comput. BigComp 2017*, pp. 453–454, 2017, doi: 10.1109/BIGCOMP.2017.7881755.

[24] L. Chen, X. Wang, and T. Q. Peng, "Nature and diffusion of gynecologic cancer-related misinformation on social media: Analysis of tweets," *J. Med. Internet Res.*, vol. 20, no. 10, 2018, doi: 10.2196/11515.

[25] Z. Jin, J. Cao, Y. Zhang, and J. Luo, "News verification by exploiting conflicting social viewpoints in microblogs," *30th AAAI Conf. Artif. Intell. AAAI 2016*, pp. 2972–2978, 2016.

[26] B. Shi and T. Weninger, "Fact Checking in Heterogeneous Information Networks," pp. 101–102, 2016, doi: 10.1145/2872518.2889354.

[27] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake News Detection on Social Media," *ACM SIGKDD Explor. Newsl.*, vol. 19, no. 1, pp. 22–36, 2017, doi: 10.1145/3137597.3137600.

[28] Y. Li, X. Zhang, and S. Wang, "Fake vs. real health information in social media in China," *Proc. Assoc. Inf. Sci. Technol.*, vol. 54, no. 1, pp. 742–743, 2017, doi: 10.1002/pra2.2017.14505401139.

[29] S. Ahmed, K. Hinkelmann, and F. Corradini, "Combining machine learning with knowledge engineering to detect fake news in social networks - A survey," *CEUR Workshop Proc.*, vol. 2350, 2019.

[30] Y.-J. Li, X.-L. Shen, C. M. K. Cheung, and M. K. O. Lee, "Literature review on health misinformation on social media Health Misinformation on Social Media: A Literature Review Completed Research Paper," 2019, [Online]. Available: http://pacis2019.org/wd/Submissions/PACIS2019_paper_263.pdf.

[31] G. L. Ciampaglia, P. Shiralkar, L. M. Rocha, J. Bollen, F. Menczer, and A. Flammini, "Computational fact checking from knowledge networks," *PLoS One*, vol. 10, no. 6, pp. 1–13, 2015, doi: 10.1371/journal.pone.0128193.

[32] J. Z. Pan, S. Pavlova, C. Li, N. Li, Y. Li, and J. Liu, "Content based fake news detection using knowledge graphs," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 11136 LNCS, pp. 669–683, 2018, doi: 10.1007/978-3-030-00671-6_39.

[33] J. Ma *et al.*, "Detecting rumors from microblogs with recurrent neural networks," *IJCAI Int. Jt.*

Conf. Artif. Intell., vol. 2016-Janua, pp. 3818–3824, 2016.

[34] P. Qi, J. Cao, T. Yang, J. Guo, and J. Li, "Exploiting multi-domain visual information for fake news detection," *Proc. - IEEE Int. Conf. Data Mining, ICDM*, vol. 2019-Novem, pp. 518–527, 2019, doi: 10.1109/ICDM.2019.00062.

[35] F. Yang, X. Yu, Y. Liu, and M. Yang, "Automatic detection of rumor on Sina Weibo," *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, vol. 2, 2012, doi: 10.1145/2350190.2350203.

[36] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," no. Mlm, 2018, [Online]. Available: http://arxiv.org/abs/1810.04805.