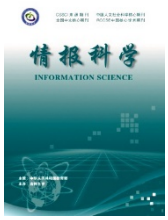

情报科学
Information Science
ISSN 1007-7634,CN 22-1264/G2

《情报科学》网络首发论文

题目：基于 Bert-BiLSTM 混合模型的社交媒体虚假信息识别研究
作者：冯由玲，康鑫，周金娉，李军
网络首发日期：2024-01-29
引用格式：冯由玲，康鑫，周金娉，李军．基于 Bert-BiLSTM 混合模型的社交媒体虚假信息识别研究[J/OL]．情报科学.
https://link.cnki.net/urlid/22.1264.G2.20240126.1809.008

网络首发时间：2024-01-29 16:26:28
网络首发地址：https://link.cnki.net/urlid/22.1264.G2.20240126.1809.008

基于 Bert-BiLSTM 混合模型的社交媒体虚假信息识别研究

冯由玲[1,2]，康鑫[1,2]，周金娉[1,2]，李军[1,2]

（1.吉林财经大学 管理科学与信息工程学院，长春市 130117；2.吉林省商务大数据研究中心，长春市 130117）

摘要：【目的/意义】探索信息疫情背景下社交媒体中真伪信息的主题特征，研究社交媒体平台评论信息特征及真伪识别问题，为用户和社交媒体平台信息识别提供参考依据。【方法/过程】针对社交媒体平台上疫情相关多主题数据，以 Twitter 平台推文为数据集。运用 LDA 模型，提取真实信息和虚假信息的主要表述和语义特征。引入 Bert 预处理方式，融合双向长短时记忆网络算法，构建 Bert-BiLSTM 混合模型，识别虚假疫情信息。【结果/结论】基于 LDA 主题模型的对比研究，发现真实和虚假信息在主题和表述特征上存在显著差异。通过与传统机器学习算法进行比较，Bert-BiLSTM 模型对虚假疫情信息识别具有显著优势，准确率达到 0.960，F1 值为 0.961。因此，本文构建的 Bert-BiLSTM 模型将为虚假信息识别提供更精准、高效的解决方案。【创新/局限】以社交媒体平台疫情信息为研究对象，综合运用 LDA 主题模型探究了疫情信息的特征，在小规模数据集上以较低成本实现了多主题数据的有效识别，为信息疫情治理提供了高效的解决方案。

关键词：社交媒体；多主题数据；LDA 模型；对比研究；虚假信息识别； Bert-BiLSTM；
分类号：G206

0 引言

随着移动互联网的深度普及和互联网技术的不断创新，社交媒体平台由于其便捷性和互动性等特点成为公众获取资讯的首要渠道，也正因其信息的易获取性，使社交媒体平台上虚假信息大量泛滥[1]。而虚假信息一旦被接受，便很难被更正，这将对公众认知产生严重负面影响。因此，虚假信息也被世界经济论坛列为对未来社会的主要威胁[2]。

中国公共卫生应急条例将突发公共卫生事件定义为“突然出现的无法解释的疾病，饮水，食物中毒等，并导致大规模群体感染的传染性疾病以及其他更严重的事件”[3]。而人们在突发公共卫生事件中面临的信息环境可能与以往有所不同，有学者指出，公共卫生事件同时也是下一场“信息危机”[4]。以新冠疫情为例，在疫情面前公众卫生安全面临巨大挑战，根据世卫组织 2022 年 11 月 7

日的数据，全球新冠肺炎共计 6.2 亿人确诊，657 万人因新冠疫情死亡[5]。然而随着疫情防控工作的开展，大量关于疫情和病毒本身的虚假信息也涌现在网络中。由于公众缺乏对新冠病毒的了解以及恐慌心理，关于新冠病毒的虚假信息传播广泛而迅速，用户暴露在大量不准确、充满误导性的信息中，使用户难以辨别疫情相关信息的真假[6]。2020 年 2 月世卫组织的一份报告中称这种现象为“信息疫情（Infodemic）”。“信息疫情”对公众的心理产生负面影响，降低了公众对公共卫生机构的信任，同时令信息接受者产生一些不合理的预防措施，间接影响公众卫生治理的效果，其破坏效果不亚于疫情本身[7]。

社交媒体平台在疫情期间扮演着信息实时分享和新闻传播的角色，由于隔离管控原因，更多用户选择通过社交媒体平台参与疫情相关话题讨论来获取相关信息[8-9]。本文选取 Twitter 平台疫情信息作为研究对象主要考虑一下原因。首先，Twitter 作为一种使用较为广泛的社交媒体平台针对虚假信息采取的是人工识别的方式[10]，识别效率不理想；其次，推文中除包含医疗信息外，还包含大量的民生相关信息[11]，这使得疫情信息成分复杂，增加了信息鉴别成本；另外，存在于推特上的虚假信息多来自含有“V”字标志的认证用户以及一些拥有大量关注者的意见领袖[12]，而虚假疫情信息的采纳者由于自身信息素养较低[13]，也会成为虚假信息的发布者[14]。最后，虚假信息往往涵盖多个主题，这使得“信息疫情”现象在 Twitter 平台上更为明显。而 Twitter 平台的疫情相关推文多为国内社交媒体平台用户关注度较高的国际问题，这使得针对 Twitter 平台的疫情信息研究对国内虚假疫情信息治理有一定借鉴意义。

综上，本文以 Twitter 平台中疫情主题相关推文为研究对象，构建基于 Bert，TF-IDF 和 Tokenizer 文本预处理方式和多种机器学习算法融合的虚假信息识别模型，通过比较研究得到最优模型，为虚假疫情信息识别提供指导。同时，利用 LDA 主题词模型提取真实疫情信息和虚假疫情信息的主要特征，为用户提供更为直观的信息描述。

1 相关研究

虚假信息识别是一种通过构建模型的文本分类任务，通过对不同标签的文

本内容进行特征提取，将文本分为虚假信息类和非虚假信息类。目前，学者们针对虚假信息识别的手段主要有三种：基于传统方法的文本特征识别、基于机器学习算法的识别以及基于深度学习算法的识别。

1.1 基于文本特征识别研究现状

学者们对于虚假信息的特征描述往往从表述，语义以及用户信息等方面展开。张帅针对微信辟谣助手上爬取的虚假信息研究中发现信息模糊、信息不完整、夸大事实以及假借权威等为虚假信息的主要特征【15】。学者 Syed 针对 Twitter 平台的虚假信息特征研究中，将信息特征分为三类，第一类为基于内容、第二类基于平台特征、第三类包括语句的心理特征。研究表明，虚假信息中存在大量自相冲突的内容以及最高级副词，部分存在人称代词和发誓性词语【16】。曾子明对微博平台虚假信息进行 LDA 主题识别，较好地提取出了虚假信息的主题特征【17】。

目前，社交媒体平台主要采取的是基于文本特征的人工识别，对识别人员的经验要求高，识别方法费时费力，消耗大量人工成本。同时过度依赖用户的举报，覆盖面窄【18】。

1.2 基于机器学习识别研究现状

最早针对疫情数据的虚假信息识别研究出现于 2020 年， Al-Rakhami 结合推文及用户特征进行对比试验后，发现支持向量机与随机森林的集成算法准确率最高，达到 0.978【19】。学者 Liu Y 利用 GBDT 算法对虚假信息进行预测，准确率高达 82%以上【20】。赵月华在针对虚假信息的预测中，将数据的特征分类为中心特征和外围特征，并利用多维数据特征的支持向量机取得了较高的准确率【21】。

以上研究皆基于传统机器学习算法，机器学习算法在处理由文本特征转化的低维数据上准确高效且训练成本低，但对数据特征要求高，人工处理费时费力，同时特征提取上未考虑文本本身，普适性较弱。因此有学者在机器学习算法的基础上，开发了基于深度学习的虚假信息识别模型。

1.3 基于深度学习识别研究现状

近几年，学者利用深度学习算法对虚假信息进行了研究，但针对虚假疫情信息的研究并不多。詹骞采取长短期记忆网络算法对在线医疗平台虚假信息进行识别[22]。赵月华对微博平台构建的医疗信息数据研究中发现经预训练的 Bert 模型识别效率最高[23]。Roy PK 通过长短时效记忆网络对在线新闻网站虚假信息进行了有效的识别[24]。

由相关研究可知，传统机器学习算法虽然在准确率上表现令人满意，但对于虚假信息数据的要求较高，需要收集用户相关信息。此外，学者针对疫情信息识别的研究主要为单一主题的信息，而多主题数据往往需要较大量的训练数据，识别成本高。鉴于此，本文以社交媒体平台多主题小规模疫情数据集为研究对象，通过 Bert 预训练语言模型对文本进行向量化表示的方式构建迁移模型，结合机器学习和深度学习算法进行对比研究，寻找虚假信息识别的低成本高效解决方案。

2 研究设计

2.1 研究框架

本文以社交媒体平台 Twitter 上新冠疫情信息数据集为研究对象，首先利用 LDA 对数据进行特征总结；然后采取不同预处理方式对文本数据进行向量化表示，将其引入到不同机器学习算法中进行对比研究，总结出虚假信息识别的高效模型；最后通过模型的构建增强用户和平台对虚假疫情信息的识别能力。研究框架如图 1。

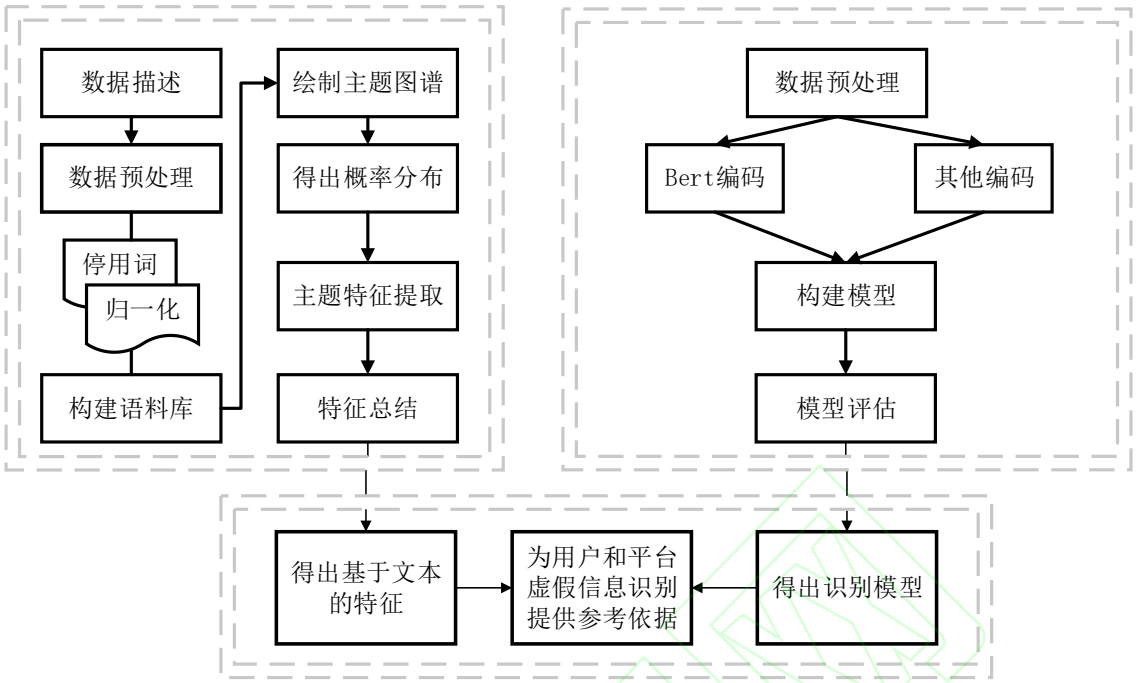

图 1　研究框架

Figure 1 The framework of research

2.2 研究方法

2.2.1 LDA 主题词模型

LDA（Latent Dirichlet Allocation）模型是由 Blei 提出的分析文本中隐含主题的模拟器[25]，其原理如下：

$$p\big((\theta,z,w|\alpha,\beta)\big)=\prod_{n=1}^{N}p(z_n|\theta)\,p(w_n|z_n,\beta) \quad (1)$$

其中，α，β为给定参数，α 为控制文档的主题分布，β 为每个主题的词分布，$p(z_n|\theta)$代表的是主题分布为θ的情况下，从某篇文档 w 中得到词 n 的主题 z_n的概率。

2.2.2 Bert 预处理方法

Bert 模型是谷歌 AI 于 2018 年 10 月提出的一种基于深度学习的语言表示模型，其全称是 Bidirectional Encoder Representation from Transformers(基于转换器的双向编码表征)，包括两个模块遮蔽式语言模块（Masked Language Modeling）和 NSP 模块（Next Sentence Prediction）[26]。区别于传统模型的

单向训练方式，Bert 采取双向训练使每个单词都能在训练中“看到自己的位置”。如图 2 所示， Trump contracts virus 这句话，经过分词和加入特定符号后，分别经过 Token Embedding（词嵌入），Segment Embedding（段落嵌入）以及 Position Embedding（位置嵌入）[27]转换为向量并逐层添加，得出训练完毕的 Bert 向量。本文所采取的预训练模型为 Bert-base-uncased。它的特点是模型的识别成本较低且识别效率高，同时涵盖文本特点比较广泛。

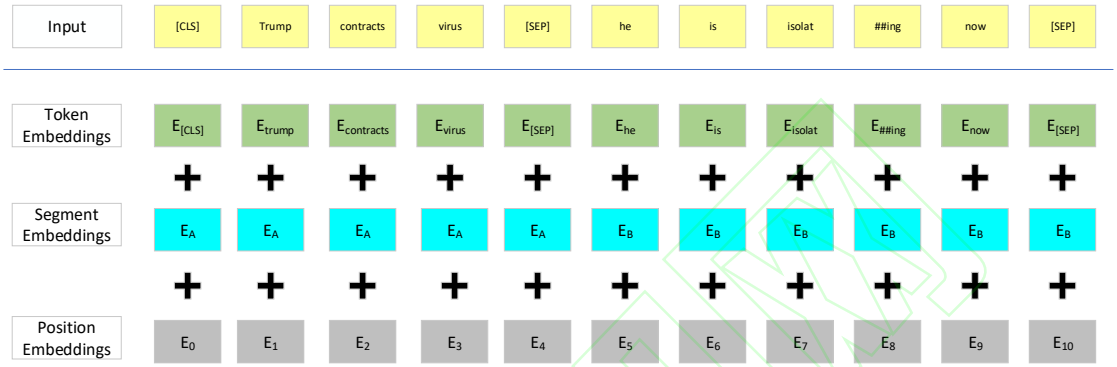

图 2 Bert 模型输入特征

Figure 2 The input feature of Bert Model

2.2.3 BiLSTM 算法

LSTM 是一种处理序列数据的神经网络结构，相比于传统的神经网络，更擅长处理长序列数据，同时也解决了 RNN 中容易出现的梯度爆炸问题。而 Bi-LSTM 由前向 LSTM 和后向 LSTM 构成，更适合需要整合前后信息的数据。BiLSTM 假设向量为X，则有$X\epsilon R^{B\times T\times D}$，其中 B 为单次传递给模型用以训练的样本个数，T 是序列长度，D 是输入数据的维度。

BiLSTM 可以用如下公式表示：

$$h_t^f = LSTM_f(x_t^{enc}, h_{t-1}^f) \tag{2}$$

$$y_t^f = h_t^f \tag{3}$$

$$h_t^b = LSTM_b(x_t^{enct}, h_{t+1}^b) \tag{4}$$

$$y_t^b = h_t^b \tag{5}$$

其中，x_t^{enc}是在时间 t 内编码后输入，，h_{t+1}^b和h_{t-1}^f分别是前一个时刻的隐状态和后一个时刻的隐状态，y_t^f和y_t^b是在时间 t 内的前向输出和后向输出。$LSTM_f$

和$LSTM_b$分别是 LSTM 的前向函数和后向函数。据此，我们可以计算双向 LSTM 的输出如下：

$$h^f = BILSTM(X, LSTM_f, LSTM_b) \quad (6)$$

$$H = [H^f, H^b] \quad (7)$$

其中，H^f和H^b是前向和后向 LSTM 的输出，而 H 是 BiLSTM 结合H^f和H^b的输出维度。模型的输出可以被表示为$y_T \epsilon R^{B \times H}$，其中 H 是 BiLSTM 的输出维度。

由于本文模型的目的是解决二分类问题，因此损失函数选取二元交叉熵函数，其定义如下：

$$L(y, \hat{y}) = -\frac{1}{n}\sum_{i=1}^{n}[y_i \, log \, \hat{y} + (1 - \hat{y}) \, log(1 - \hat{y})] \quad (8)$$

其中 y 是真实标签，$\hat{y}$是预测概率，n 是批次中的样本数。

2.3 Bert-BiLSTM 疫情信息混合识别模型构建

本文构建 Bert-BiLSTM 模型如图 3，经过 Bert 编码后，输入数据变成了 768 维的向量，并经过左向 LSTM 和右向 LSTM 处理后得到的隐状态进行拼接后，通过 Sigmod 函数转换为概率并进行二分类。

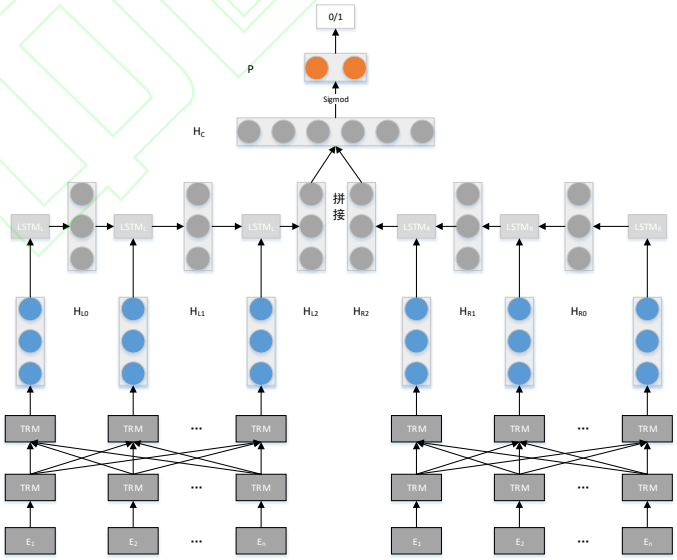

图 3 Bert-BiLSTM 混合识别模型结构

Figure 3 Bert-BiLSTM mixture model structure

3 基于 LDA 主题模型的疫情信息特征分析

本文数据来自 Kaggle 数据集网站上的 Twitter 疫情信息数据集[28]，数据涵盖 2019-2021 年 Twitter 上疫情主题英文推文 8560 条，其中真实信息 4480 条，虚假信息 4080 条。通过正则表达式清洗掉原始数据中的大量链接锚和表情符号，得到本文的初始数据。

3.1 疫情信息描述性分析

3.1.1 疫情数据文本长度分析

通过对数据中虚假信息和真实信息进行分类，并分别绘制数据相关的小提琴图，得出关于数据文本长度的情况如下图 4。

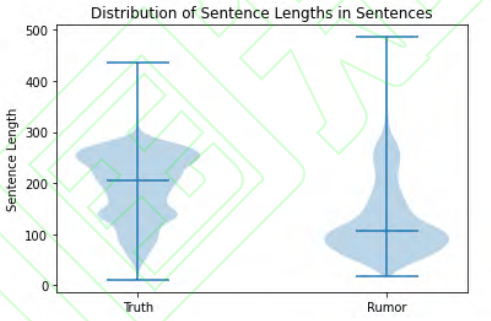

图 4 文本长度描述

Figure 4 The description of text length

杨丹等人的研究表明，虚假信息多为用户自创信息，文本长度较短[29]。图 5 中真实信息比虚假信息的文本长度更长，这表明了虚假疫情信息更多地是用户自创的信息。

3.1.2 疫情数据情感分析

本文通过 Snownlp 库对虚假疫情信息和真实信息进行情感打分，分别绘制小提琴图，得出疫情相关推文整体情感得分如图 5。

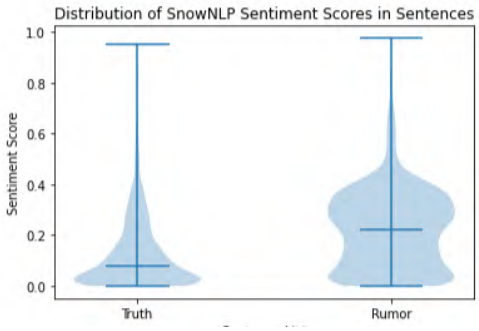

图 5 情感得分描述

Figure 5 The description of emotional score

由图 6 可知，真实信息情感得分较低反映了疫情的形势不容乐观，虚假信息表述方式相对积极，对于疫情信息的整体情感比真实信息更加乐观。Jemielniak 指出，Twitter 虚假疫情信息更多是对负面新闻的不当报道，而不是披露悲观事实[30]。同样，陈思静在他的研究中指出，人们更愿意接受相对乐观的信息，虚假信息往往通过共情和较为委婉的表述说服信息接收者[31]，因此虚假信息相对真实信息情感得分较高。

3.2 数据准备

为保留能体现表述特征的语气词，本文未采用 NLTK 英文停用词表，而是使用正则表达式过滤不能反映实际文本内容的介词、冠词、名词等。然后通过 NLTK 中的 Porter Stemmer 对数据进行词干提取，得到归一化的文本数据，分别提取其中真实数据和虚假数据，并利用 LDA 模型对两组数据进行分析。

3.3 基于真实信息的 LDA 模型构建

由图 6 可知，在选取主题数为 5 的时候连贯分数较高同时主题困惑度较低，因此本文选取 5 为主题数绘制 LDA 主题模型，利用 Pyldavis 库绘图后不同主题分布较均匀，表现良好，因此本文选取 5 为真实信息主题数。用 Pyldavis 库绘制真实信息主题图得出主题图中主题词如表 1 所示。

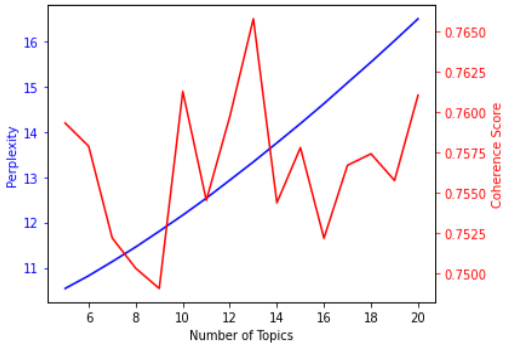

图 6 真实疫情信息困惑度和连贯系数

Figure 6 Real epidemic information perplexity and coherence score

表 1 真实疫情信息主题词

Table 1 Real epidemical information topic terms

	确诊情况	疫情传播	印度抗疫	死亡情况	疫情检测
主题词	Cases	Covid19	Tests	Covid19	Covid19
频率	0.032	0.037	0.043	0.016	0.014
主题词	New	More	Covid19	Cases	Tests
频率	0.021	0.013	0.032	0.015	0.011
主题词	Covid19	People	Cases	Confirmed	Number
频率	0.020	0.010	0.020	0.015	0.010
主题词	Number	Coronavirus	Indiafightscorona	Deaths	Completed
频率	0.018	0.009	0.017	0.014	0.009
主题词	Confirmed	Spread	Daily	Discharge	Laboratories
频率	0.017	0.009	0.017	0.013	0.008

根据各个主题所反映的内容如表 1，本文将主题 1-5 依次总结为确诊情况、疫情传播、印度抗疫、死亡情况和疫情检测。确诊情况主题主要反映新冠肺炎的确诊病例数等相关信息；疫情传播主题倾向于反映疫情的传播路径和预防措施；印度抗疫主题主要反映印度在抗击疫情方面进行了大量的检测；死亡情况主题主要反映了权威机构发布的疫情导致死亡的案例；疫情检测主题表达了实验室完成新冠肺炎检测的人数等信息。以上各主题中，主题词 covid19，cases 重复出现，这表明真实信息主要是对新冠疫情实际情况的客观陈述。

表 2 为真实疫情信息各主题代表推文部分截取，由表可知，表达日期和数量的词出现频率较高，说明真实信息在表意上比较倾向于精确表达。在人称代词上惯用第一人称复数，通常指代发布推文的主体，用以引出观点，表述更负责任。

表 2 真实疫情信息部分截取

Table 2 Real epidemic information segments(partial)

主题	推文
确诊情况	**COVID-19** Update There are 7 new **confirmed cases** of **COVID-19** to report in the community. In total **we** have 56 active **cases** …
疫情传播	To protect yourself from **COVID19** avoid being exposed to the virus and use everyday prevention action to prevent the **spread** of respiratory virus…
印度抗疫	**IndiaFightsCorona**: **We** have 1524 **COVID** testing **laboratories** in India and as on 25th August 2020 36827520 **tests** have been done
死亡情况	**We** have an update to today's chart of daily **deaths** with California's (late arriving) **number** included. The 7-day average was almost unchanged at 520
疫情检测	**Our** daily update is published. States reported 586k **tests** 28k **cases** and 224 **deaths**

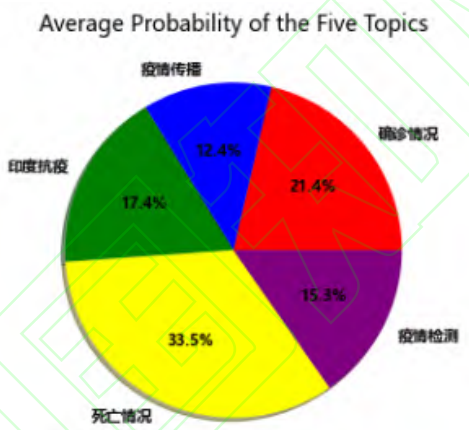

图 7 真实疫情信息主题概率平均值

Figure 7 Probability mean of real epidemic information topics

图 7 为真实信息主题概率的平均值，主要反映各个主题在真实信息中所占的比重。其中死亡情况主题和确诊情况主题平均值相对较高，死亡情况主题对应的是各种疫情导致死亡的相关信息，而确诊情况主要反映了各地确诊新冠肺炎的情况。由表 2 可知这两者的表述往往是连续出现的。

3.4 基于虚假信息的 LDA 模型构建

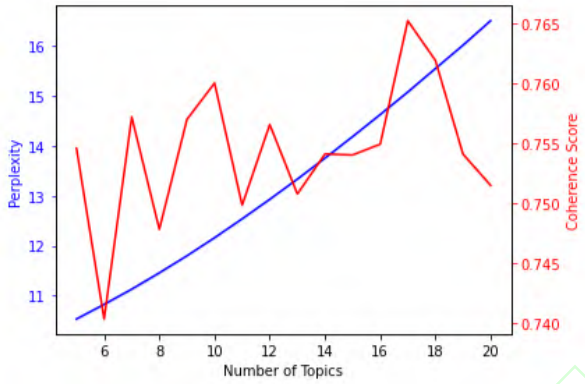

图 8 虚假信息困惑度及连贯系数

Figure 8 Epidemic misinformation perplexity and coherence score

虚假信息的主题模型构建方法与真实信息相同如图 8。虚假疫情信息的主题数确定在 5 时困惑度较高而连贯系数较低，实际绘图后各主题存在大量交叉，最终经实验后发现主题数为 3 时各主题在主题图上较为分散，因此将虚假信息主题数确定为 3。去掉各主题中重复出现的 covid19，coronavirus 以及大量无实意词汇后，将词频较高的单词依词频顺序展示如表 2 所示。

表 3 虚假疫情信息主题词

Table 3 Epidemical misinformation topic term

	引用佐证	政府政策	负面新闻
主题词	New	Trump	People
词频	0.020	0.012	0.020
主题词	Video	President	lockdown
词频	0.018	0.010	0.017
主题词	India	Vaccine	News
词频	0.017	0.009	0.016
主题词	Claim	Donald	Mask
词频	0.014	0.009	0.016
主题词	Facebook	Patients	Cure
词频	0.012	0.009	0.015

由表 3 可知虚假疫情信息与真实信息相比每个主题包含的内容较为集中且彼此独立，本文根据主题词特征将三类主题分别命名为引用佐证，政府政策以及负面新闻。

引用佐证主题中，高频主题词包含 video、India、claim、 Facebook 等，结合数据集实际内容，video 一词主要反映了虚假疫情信息声称存在视频佐证推文的内容，claim 反映了虚假疫情信息多假借权威组织发布推文，Facebook 反映了虚假疫情信息习惯性引用来自 Facebook 的内容。

政府政策主题高频主题词主要包括美国前任总统的姓名，vaccine 以及 patients。结合实际文本内容，针对美国前任总统的虚假信息主要集中于将要实行的政策，关于直接的政治诋毁和个人情况相对较少。vaccine 主题词主要包括疫苗相关负面信息。patients 主题词包含内容较为复杂，并未发现较为具体的规律。

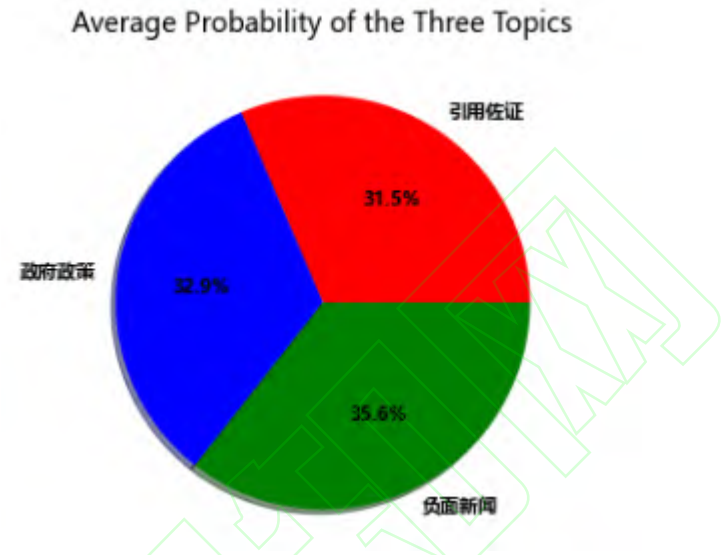

图 9 虚假疫情信息主题概率平均值

Figure 9 Probability mean of epidemic misinformation topics

负面新闻高频主题词包括 lockdown、mask 和 cure 等主题词。其中 lockdown 主题词主要包含的是涉及各国封闭管理的负面信息，目的在于反对封闭管理。mask 主题词主要反映的是关于口罩的负面信息，旨在反对强制佩戴口罩。Cure 主题词更多地反映了新冠肺炎患者被治愈的不实信息，以降低公众对于疫情风险的提防性。另外，主题词中包含大量否定词，转折词，语气较为绝对且具有诱骗性。

表 3 为各主题中代表性的推文，由内容可以发现，假借权威和引用其他信息源是虚假疫情信息在内容上的主要特征。另外，在内容上，虚假信息往往是对虚构事件的描述。而在表述上，虚假信息往往包含大量的从句引导词（as，that），人称上倾向于使用第三人称。

总结虚假信息的主要主题特征，可以得出虚假信息主要通过制造焦虑和恐慌达成某种政治目的，反映了 Twitter 上的虚假信息主要为不同个体用户抒发的政治主张。

表 4 虚假疫情信息部分截取

Table 4 Real epidemic information segments(partial)

主题	推文
引用佐证	**Video** shows doctors and nurses turning their back to the Health ministry of Spain **as** he arrives to a Hospital
政府政策	In Latest Move to Stop Coronavirus, **Trump** Urges Fed to Cut Interest Rates His Heavily Indebted Company Pays
负面新闻	A **video** of police officers immobilizing a screaming woman. The caption states **that** they were being beaten by the police because they refuse to go to their houses during **lockdown** in Minas Gerais Brazil

图 9 为虚假疫情信息主题概率平均值，反映了所有虚假信息在不同主题中分布的概率。由不同主题的概率分布可知虚假信息整体分布相对均匀。

3.5 真实和虚假信息主题特征对比

通过对真实信息和虚假信息的研究，我们找到了两者在内容特征、表述特征和主题特征三个维度的主要差别，见表 5。

另外，我们发现真实信息主题词中多包含防护，战“疫”等相对积极的词汇，同时也包括 symptom 等专业医学词汇，主要目的为陈述事实。而虚假信息主题词多为治愈等诱骗性词汇。同时，数据集大量涉及到国际问题，而虚假信息在针对国际问题上多为负面表述，带有一定政治目的。

以上分析为用户识别虚假疫情信息提供了较为直观的指导方案，同时，本文数据涉及多元主题，对其进行识别的模型需要具有较好的泛化能力。

表 5 真实和虚假信息特征对比

Table 5 Feature comparison

	内容特征	表述特征	主题特征
真实信息	多为对疫情状况的客观描述以及防护措施。一般不涉及具体的人物，推文长度相对较长，推文情感得分相对较	表达方式比较客观，程度副词少，往往引用数据佐证内容，倾向于使用第一人称复数	主题词分布均匀 主题词含义明确

	低		
虚假信息	多为对某种现象和事件的描述，描述对象往往是具体的人物，推文长度相对较短，多为用户自创信息，推文情感得分偏向中性	往往假借权威以及引用视频增强信息的迷惑度。多使用从句，程度副词以及多种人称代词	不同主题间主题词重复无实意主题词居多

4 Bert-BiLSTM 模型的实验设计与分析

4.1 实验准备

本文主要通过 Pytorch 和 Tensorflow 框架进行模型搭建，主要硬件环境如下：CPU：Intel Core I7-3970x 3950GHz；内存：16GB；GPU：GTX 1050 4GB。

本文通过 NLTK 对原始数据进行去停用词和分词，并利用 Porter Stemmer 对数据进行归一化，以便降低数据维度，减小噪声干扰。

实验中训练集 6420 条推文，测试集 2140 条推文，在机器学习算法构建的过程中，我们采取 10 折交叉验证划分训练集。

4.2 对比实验构建

本文通过构建对比实验，运用多种文本预处理方式和算法探索虚假疫情信息识别的高效模型。首先通过 TF-IDF 和 Tokenizer 传统文本预处理方式得到与预处理文本包含字段数相对应的词向量，并分别引入到支持向量机，随机森林以及 BiLSTM 机器学习算法中进行识别，以准确率、精确率、召回率、F1 值对模型效果进行评估。随后，采取 Bert 文本预处理方式，得到 6280（训练集文本数量），512（最大文本长度），768（Bert 模型隐藏单元数）的三维 Tensorflow 张量后重复输入到上述三种算法中，进行总体对比，得出 Bert-BiLSTM 混合模型有明显优势。

4.2.1 机器学习模型构建

本文运用 TF-IDF 和 Bert 将文本数据进行向量化处理后，分别引入到支持向量机和随机森林算法中，选取 10 折交叉验证后利用 GridSearchCV 对算法参数进行优化，训练模型后用测试集四项评估指标对算法进行评估，得出 TF-IDF 预处理后的随机森林模型有最佳表现。

4.2.2 深度学习模型构建

本文首先采取 Keras 中的 Tokenizer 对文本进行预处理，通过消融实验后确定输入最佳长度为 170，并添加一个包含 128 个单元，dropout 参数为 0.2 的 BiLSTM 层，之后添加全连接层将每层的数据进行连接，选取激活函数 Sigmoid 对模型进行二分类。在数据进行 10 次迭代后选取损失值最低的迭代次数对数据模型进行评估。

随后采取 Bert 预处理将文本数据向量化后，由于 Bert 的最佳数据长度为 512，在其他参数不变的情况下将 BiLSTM 的输入长度调整为 512，对模型进行评估。

4.2.3 Bert 模型参数优化

本文选取包含不同平台英文文本特征的 Bert-base-uncased 作为预训练模型。Bert-base-uncased 涵盖大量英文文本中常见的文字表述，具有较好的泛化能力。在 Bert 预训练中将 Batch size 设置为 128，最大长度设置为 512，并开启了最大填充以契合深度学习模型。为了能更好的识别填充数据，降低数据中的噪声，添加了 attention_mask 标记填充数据，并选择输出数据为 Tensorflow 张量。

4.3 实验结果对比讨论

4.3.1 基于传统编码算法的模型评估

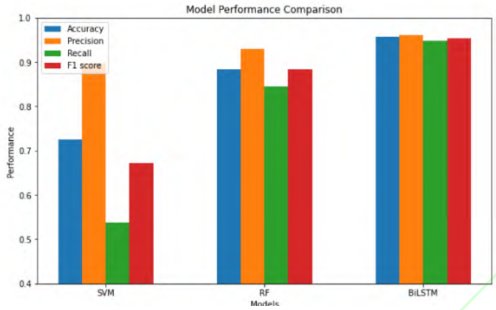

图 10 模型主要指标对比

Figure 10 Comparison of model major index

本文根据模型的各项指标构建了三种基于传统预处理方式的机器学习算法在数据集上的表现如图 10。其中，支持向量机作为传统机器学习算法表现较差，集成机器学习算法随机森林表现较好，深度学习模型 BiLSTM 表现最佳，准确率高达 95.7%，这反映了 BiLSTM 在处理较长文本数据时识别效果更好。另外，三种算法的精确率都比较理想，反映了部分真实疫情信息在特征上更为明显，更容易识别，这是因为真实疫情信息包含更多含有实际含义的词。而虚假疫情信息包含大量不含实意的助词、代词和介词等，它们在去停用词阶段被去掉了，这也增加了虚假信息识别的难度。

4.3.2 融合 Bert 预处理的模型评估

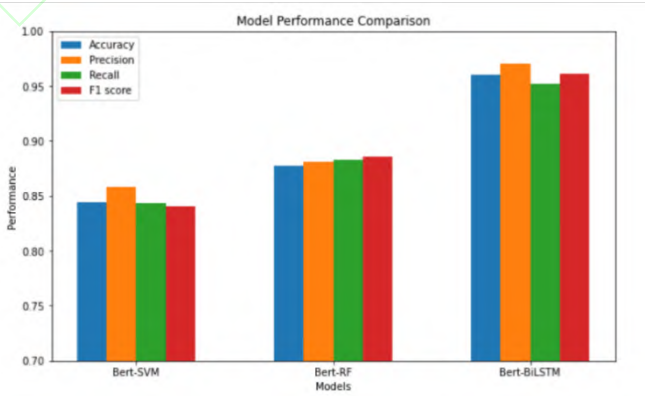

图 11 混合模型主要指标对比

Figure 11 Comparison of mixture model major index

图 11 为 Bert 和三种算法混合模型在测试集上的表现，对于 Bert 预处理后的数据， BiLSTM 有最好的表现并且优于传统预处理方式下的三种模型。

4.3.3 实验结果总结

由表 6 可知，经过 Bert 预处理后，支持向量机的表现有了显著的提升，这反映了 Bert 具有强大的语义表达能力。在随机森林模型中，由于 Bert 向量维度过大，导致随机森林中出现维度灾难[32]，从而使模型性能表现下降。与传统预处理方式相比，Bert-BiLSTM 模型在精确率上提高了 1%。

表 6 算法对比

Table 6 Algorithm comparison

	准确率	精确率	召回率	F1
SVM	0.726	0.897	0.538	0.671
RF	0.884,	0.929	0.844	0.884.
BiLSTM	0.957	0.960	0.947	0.954
Bert-SVM	0.844	0.858	0.843	0.841
Bert-RF	0.877	0.881	0.883	0.886
Bert-BiLSTM	0.960	0.970	0.952	0.961

5. 结论

针对社交媒体上虚假疫情信息泛滥问题，本文首先通过 LDA 模型结合主题词和文本实际内容对虚假疫情信息和真实疫情信息的特征进行了描述，总结了 Twitter 上虚假疫情信息的主要特点,为平台识别虚假疫情信息提供了理论依据，也增强了用户的虚假疫情信息识别能力，对国内虚假疫情信息研究也有一定启示。随后针对多主题数据，本文对 Bert 模型进行了迁移构建了混合模型，相对于现有虚假疫情信息识别方法，该模型能对多主题英文文本进行低成本有效识别。通过模型对比，Bert-BiLSTM 表现优秀，这表明融合 Bert 预处理的深度学习算法可以在较低的成本下对多主题数据进行精确识别，同时，模型具有较好的鲁棒性，可以扩展到其他领域复杂主题虚假信息的研究中。

研究中的不足之处是，由于数据集长度较短，Bert 模型在训练词向量过程中需要进行大量填充，增加了计算量并产生大量噪音，影响 Bert 模型语句推理功能的充分发挥[33]。未来的研究中应该运用词向量长度在 400-500 的文本数据，

以提高 Bert 的性能。另外，由于基于中文社交媒体平台的虚假疫情信息数据量过小，本文仅针对英文社交媒体平台数据进行了分析。由于实验条件和数据的限制，在对比实验中仅使用了支持向量机、随机森林以及 BiLSTM 几种机器学习，集成学习以及深度学习的代表性算法，可以探究其他更适合社交媒体虚假疫情信息的识别算法，减少虚假疫情信息传播，为信息疫情治理提供有效解决方案。

参考文献

[1] 王晰巍, 李文乔, 韦雅楠, 等. 社交媒体环境下网络虚假信息国内外研究动态及趋势[J]. 情报资料工作, 2020, 41(2):39-46.

[2] Del Vicario M , et al. The spreading of misinformation online[J]. Proceedings of the national academy of sciences of the united states of America，2016, 113(3):554-559.

[3] Hu JX. Improvement of Emergency Management Mechanism of Public Health Crisis in Rural China: A Review Article[J]. Iranian journal of public health, 2018，47 (2):156-165.

[4] 陈琼, 宋士杰, 赵宇翔. 突发公共卫生事件中信息过载对用户信息规避行为的影响：基于 COVID-19 信息疫情的实证研究[J]. 情报资料工作, 2020, 41(3):76-88.

[5]WHO 2022 Coronavirus (COVID-19) Dashboard[DB/OL]. https://covid19.who.int/2022- 12-25.

[6]Gruzd A. Studying the COVID-19 infodemic at scale[J]. Big data & society, 2021, 8(1) :1-6

[7] Li K. The Effects of Personality Traits on Online Rumor Sharing: The Mediating Role of Fear of COVID-19[J]. International journal of environmental research and public health，2022, 19(10) :6157

[8] Kudchadkar, SR. Using Social Media for Rapid Information Dissemination in a Pandemic: Peds ICU and Coronavirus Disease 2019[J] Journal of medical internet research，2020, 21 (8):538-546

[9] Bari, A. Exploring Coronavirus Disease 2019 Vaccine Hesitancy on Twitter Using Sentiment Analysis and Natural Language Processing Algorithms[J]. Clinical infectious diseases，2022, 74:4-9.

[10]吴非, 李旋. 后全球化时代虚假信息成为西方国家“认知战”的主要呈现手段[J]. 中国广播电视学刊, 2022, (12):35-39.

[11]Zhao YH, Zhu SC, Wan Q. Understanding How and by Whom COVID-19 Misinformation is Spread on Social Media: Coding and Network Analyses[J]. Journal of medical internet research，2023，224(66):e37623

[12] Pierri F, DeVerna MR, Yang KC. One Year of COVID-19 Vaccine Misinformation on Twitter: Longitudinal Study[J].Journal of medical Internet

research,2023,25
[13] 王世雄,朱明旻,骆彦余. 信息疫情中真假信息竞争性传播研究[J]. 现代情报,2023,43(9):124-136.
[14]Sallam M, Dababseh D. Conspiracy Beliefs Are Associated with Lower Knowledge and Higher Anxiety Levels Regarding COVID-19 among Students at the University of Jordan [J]. International journal of environment research and public health，2020，17(14)：4915
[15]张帅.社交媒体虚假健康信息特征识别[J].图书情报工作,2021,65(9):70-78.
[16]Mahbub Syed. COVID-19 Rumor Detection Using Psycho-Linguistic Features[J]. IEEE ACCESS, 10:117530-117543.
[17]曾子明,王婧. 基于 LDA 和随机森林的微博虚假信息识别研究——以 2016 年雾霾虚假信息为例[J]. 情报学报,2019,38(1):89-96.
[18]李妍蓉. 基于深度强化学习的网络虚假信息检测系统的研究与实现[D].北京：北京邮电大学,2021.
[19]Al-Rakhami MS. Lies Kill, Facts Save: Detecting COVID-19 Misinformation in Twitter[J]. IEEE ACCESS，2020，8,155961-155970.
[20]Liu, Y. Analysis and Detection of Health-Related Misinformation on Chinese Social Media[J]. IEEE ACCESS, 2019，7:154480-154489.
[21]Zhao, YH. Detecting health misinformation in online health communities: Incorporating behavioral features into machine learning based approaches[J]. Information processing & management，2021，58（1）.102390-102414
[22]詹骞,赵冰洁. 健康类虚假信息的人工神经网络识别与治理[J].现代传播(中国传媒大学学报),2022,44(8):155-161.
[23]赵月华,朱思成,苏新宁. 面向网络虚假医疗信息的识别模型构建研究——一种基于预训练的 BERT 模型[J].情报科学,2021,39(12):165-173.
[24]Roy K,Tripathy K,Weng T, et al. Securing social platform from misinformation using deep learning [J]. Computer Standards & Interfaces, 2022,84：103674
[25]David M Blei, Andrew Y Ng, Michael I Jordan. Latent Dirichlet Allocation[J]. Journal of Machine Learning Research, 2003, 3:993-1022.
[26]Devlin, J, Chang M W, Lee K, Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805,2018
[27]宋冠仪.基于 BERT 的多任务文本分析研究[D].济南：山东大学, 2021.
[28]杨丹. 网络虚假信息传播特点及治理——基于 2016 年上半年的大数据分析[J]. 新闻记者,2016,(8):38-43.
[29]COVID19 Fake News Dataset NLP[DB/OL]. https://www.kaggle.com/datasets/2df1d719501a13c90f94ad041022837e4bd39ad6a79464b51e6d6960e701fe5d
[30] Jemielniak D, Krempovych Y. An analysis of AstraZeneca COVID-19 vaccine misinformation and fear mongering on Twitter[J]. Public Health, 2021, 2021, 8(1):4-6.
[31]Chen, SJ. Persuasion strategies of misinformation-containing posts in the

social media[J]. Information processing & management, 2021, 58:5-28
[32]杨平安, 林亚平, 祝团飞. AdaBoostRS:高维不平衡数据学习的集成整合[J]. 计算机科学, 2019, 46(12):8-12.
[33] Acheampong FA，Nunoo-Mensah H，Chen, WY. Transformer models for text-based emotion detection: a review of BERT-based approaches [J].Artificial intelligence review，2021,54（8），5789-5829

基金项目：本文系吉林省科技厅 2024 年创新发展战略研究项目“基于深度学习的社交网络虚假健康信息识别及对策研究”、吉林省教育科学规划课题““互联网+”背景下高校毕业生求职平台服务质量提升策略研究”（项目编号：GH23166）、国家自然科学基金“社交媒体环境下学术成果的影响力研究-基于社会网络和传播视角”（项目编号：71971096）、吉林省教育厅社科基金“大数据背景下吉林省“互联网+冰雪旅游”游客感知分析及优化策略研究”（项目编号：JJKH20230196SK）、吉林省高等教育教学改革课题“产教融合、协同育人的新工科人才培养模式研究与实践”（项目编号：JLJY202319646619）

作者简介：冯由玲（1979- ），女，吉林省梅河口人，博士，教授，主要从事数据分析与数据挖掘研究；康鑫(1994-)，男，内蒙古通辽人，硕士研究生，主要从事自然语言处理，深度学习研究；周金娉（1985- ），女，吉林省吉林市人，博士，副教授，主要从事知识管理研究，通讯作者；李军（1974- ），内蒙古通辽人，男，博士，教授，主要从事人工智能及其应用及深度学习方向研究。

A Study on Social Network Misinformation Recognitional Mixture Model Based on Bert-BiLSTM

FENG Youling[1,2] KANG Xin[1,2] ZHOU Jinping[1,2] LI Jun[1,2]

(1.School of Management Science and Information Engineering, Jilin University of Finance and Economy, Changchun 130117, China; 2.Jilin Province Business Big Data Research Center.

Changchun 130117, China)

Abstract：【Purpose/significance】 This research aims to explore the thematic features of real and false information, study the problem of identifying the authenticity of comment information, and provide reference basis for information recognition on social media platform under the background of public health events.【Method/process】For epidemic related multi topic data on social media platforms, LDA models are used to extract thematic features of real and false information. By introducing a Bert preprocessing method, we construct a Bert-BiLSTM hybrid model to identify false epidemic information.【Result/conclusion】 We found that there are significant differences between real and false information in theme features and expression methods, providing opinions and references for identifying false information. In addition, compared with traditional machine learning algorithms, Bert-BiLSTM model has significant advantages in identifying epidemic misinformation, with an accuracy rate of 0.960 and an F1 value of 0.961. The Bert-BiLSTM model will provide a more efficient and accurate solution for misinformation recognition.【Innovation/limitation】Taking epidemic information on social media platforms as the research object, the LDA model was comprehensively used to explore the main characteristics of real and false epidemic information. Effective identification of multi topic data was achieved at a lower cost on small-scale datasets, providing an efficient solution for infodemic management.

Keywords：social media; multi topic; LDA model; comparison research; epidemic misinformation identification; Bert-BiLSTM