doi: 10. 3969 / j. issn. 1674 - 1242. 2023. 04. 010

基于国际新闻的疾病趋势预测算法

裴莹[1]，王鏖清[2]，韩霄松[2]
（1. 长春财经学院信息工程学院，吉林长春 130122;
2. 吉林大学计算机科学与技术学院，吉林长春 130012）

【摘要】文本分类作为自然语言处理领域的核心内容，已经成为文本处理的重点研究问题。该文主要针对社会上出现的大规模常见疾病进行预测。该文通过获取全球各大新闻媒体报道的新闻文本，分别统计新闻文本中出现次数排名前十的疾病，分析原始数据分布的特征。该文将基于 CNN 和 LSTM 网络的文本模型与基于 LSTM 网络的疾病趋势模型进行融合，综合分析文本中新闻文本的文本信息和疾病的时间序列，并使用了一种特殊的疾病选择策略。实验结果表明，该策略在 7 种不同的新闻数据集上获得了 70%以上的准确度。该文提出的融合策略和疾病选择策略对疾病的趋势预测具有一定的意义，有助于提高疾病趋势预测的准确性。

【关键词】人工智能；自然语言处理；文本分类；长短期记忆网络；卷积神经网络
【中图分类号】TP391.1 【文献标志码】A 文章编号：1674-1242（2023）04-0398-07

Disease Trend Prediction Algorithm Based on International News

PEI Ying[1], WANG Aoqing[2], HAN Xiaosong[2]
(1. College of Information Engineering, Changchun University of Finance and Economics, Changchun, Jilin 130122, China;
2. College of Computer Science and Technology, Jilin University, Changchun, Jilin 130012, China)

【**Abstract**】As the core of natural language processing, text classification has become a key research issue in text processing. This paper focuses on the prediction of large-scale and common diseases in society. Obtain news texts reported by major news media around the world, respectively count the top ten disease rankings in the news texts, and analyze the characteristics of the original data distribution. In this paper, the text model based on CNN and LSTM networks and the disease trend model based on LSTM networks are merged to comprehensively analyze the text information of news and the time series of diseases, and use a special strategy for selecting diseases. The experimental results show that the strategy achieves more than 70% accuracy on seven different news data sets. The fusion strategy and disease selection strategy proposed in this paper have certain significance for disease trend prediction and can help improve the accuracy of disease trend prediction.

【**Key words**】Artificial Intelligence; Natural Language Processing; Text Classification; Long Short-Term Memory Networks; Convolutional Neural Networks

收稿日期：2023-06-26。
基金项目：国家自然科学基金重点研发项目（No. 2021YFF1201200），国家自然科学基金（No. 62172187），吉林省科技发展计划（No. 20220201145GX、No. 20200708112YY 和 No. 20220601112FG），吉林省教育厅科学技术研究项目（No. JJKH20221262KJ），广东省高校创新团队项目（No. 2021KCXTD015），广东省重点学科项目（No. 2021xDJS138）。
作者简介：裴莹（1990—），女，吉林省长春市人，博士研究生，从事大数据分析研究。
通信作者：韩霄松，男，副教授，邮箱（E-mail）：hanxiaosong@jlu. edu. cn。

0　引言

Pew 研究中心最近的一项网络调查显示，超过72%的网民浏览过医学健康方面的网页[1]。从那些在医学网站上被普遍关注的问题到几年前医学实验室所做的课题研究，甚至到仅在局部环境下流通的文献都受到了人们的关注[2]。在网络新闻中，疾病出现的频率很好地说明了人们对疾病的关注程度，进而反映出疾病本身的流行和发展趋势。将网络新闻中疾病的出现频率作为对社会热点疾病的把控，对疾病趋势的预测有一定的研究意义。

同时，疾病的治疗费用和代价极其昂贵。例如，在 2013 年 5 岁前死亡的 6 300 万名儿童中，仅传染病一类疾病导致的死亡人数就占 51.8%（3 270 万人）[3]。即使在发达国家，疾病也会造成大量的人员伤亡和高昂的治疗费用，对经济的发展也有很大的影响。例如，美国每个流感季节都会死亡 3 000 ~ 49 000 人，平均经济产出减少 163 亿美元[4]。因此，有效和准确的疾病趋势预测有助于政府和社会更好地防控疾病，进一步减少生命财产损失。

数据分析领域的著名学者 Foege[5]曾在 1976 年发表的文章中论述了收集、分析疾病控制数据及据此做好疾病控制的重要性。该文章认为，监控并收集公众行为的有效数据可以为决策者的政策制定提供具体、有效的支持。为了对疾病进行监控、研究和预测，多个国家建立了电子监控系统。其中，美国引入国家疾病电子监控系统[6]，提高了公共卫生监测的有效性。该系统的运行依赖在职医生和护士撰写的报告[7]。

随着互联网的普及和大数据时代的来临，人工处理速度开始跟不上数据量的增加速度。越来越多的研究人员开始进行基于大数据的传染病监测研究，以补充现有的系统和设计缺陷。在这些研究中，目前国外正在使用大数据（如因特网）搜索、查询、监测传染病的发生，可以接近实时速度收集并处理因特网数据。根据 Towers 等的研究，利用因特网可以比传统的监控系统更快地获取疾病监控数据。除此之外，Huang 等[8]使用广义加性模型（Generalized Additive Models，GAM）获取疾病的发病情况，可以快速识别传染病发展趋势。李薇等[9]分析了天津市河北区 2013—2020 年疑似预防接种异常反应案例，可以有效地观测流行病发病趋势。

在 Tenkanen 等[10]的研究中，使用具有实时访问性的社交媒体大数据来预测疾病的趋势。Shin 等[11]的一项研究发现传染病和推特（Twitter）数据高度相关，未来有可能使用该数据监控系统来监测疾病。除了这些研究，还有人使用深度学习领域的技术来实现传染病的预测。在 Xu 等[12]的一项研究中，使用深度学习模型比广义线性模型、最小绝对收缩和选择算子模型及自回归移动平均模型具有更好的预测性能。白金川等[13]利用网络爬虫构建了异步式程序、暂态存储程序来处理医学影像，有效减轻了服务器的工作压力。这些研究表明，互联网大数据和深度学习相结合可以更有效地预测传染病。

本文主要针对社会上出现的大规模常见疾病趋势进行预测。通过获取全球各大新闻媒体报道中的新闻文本，分别统计新闻文本中出现次数排名前十的疾病，分析原始数据分布的特征，并对分析结果进行可视化展示。现阶段对于疾病的预测与防控，学术界的研究更关注疾病的传播模式、发病机理和治疗措施，使用新闻或其他舆论辅助疾病趋势预测的研究成果较少。

本文提出的疾病趋势预测方法总体流程如图 1 所示。该方法分为预处理、神经网络模型和评估 3 个阶段。其中，预处理阶段的输入为爬取的新闻信息和两个专业疾病分类编码系统中的名词库。经过预处理，将新闻中的疾病名词进行替换，获得替换后的新闻文本信息。得到新闻文本信息之后，从两个方面获得疾病特征信息：从文本信息中查找、提取疾病名词，获得疾病时间序列信息；使用 Doc2Vec 将新闻文本表征为 300 维的特征向量。本文分别使用 3 种神经网络模型对疾病趋势进行分析，并分别从训练集合、验证集合中对不同神经网络模型的效果进行评估。

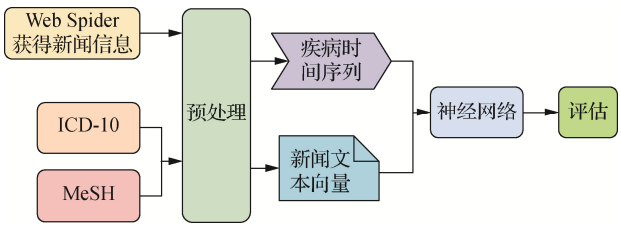

图 1　疾病趋势预测方法总体流程

Fig. 1　Flow chart of disease trend prediction

1　数据的获取和预处理

实验所需数据源来自 3 个部分，分别为新闻数据

源、ICD-10 和 MeSH 疾病数据源。其中，新闻数据构建爬虫从七大新闻媒体进行爬取，ICD-10 和 MeSH 疾病数据源则直接从官网获取。

1.1 新闻数据源

新闻语料库的新闻选自三大地区，即北美、欧洲和中东地区。本文选取了代表不同政治阶级、载体不同的多个专业新闻媒体进行新闻爬取，以保证新闻来源的广泛性和代表人群的普遍性。具体的新闻媒体统计信息如表 1 所示。

表 1 新闻媒体统计信息

Tab. 1 News media statistics

新闻媒体	起始时间	终止时间	新闻条数
CNN	2011/8/4	2020/3/20	124 394
FOX	2001/1/22	2020/3/2	524 695
Daily Beast	2008/10/5	2019/12/31	99 071
BBC	2011/4/11	2020/4/8	363 957
Reuters	2005/7/22	2017/2/10	2 599 837
Daily Telegraph	2005/1/1	2020/3/31	1 102 600
Al Jazeera	2003/4/19	2020/3/9	206 537

本文选取了美国的 3 个新闻媒体，分别为有线电视新闻网（CNN）、福克斯新闻网站（FOX）、每日野兽（Daily Beast）。CNN 注重即时、逼真的现场新闻播报，而忽视了现象与实际情况的联系。同时，CNN 与美国自由派关系密切，很多观点和数据更偏向左派政党。FOX 宣扬新闻的公平公正，事实上其在政治立场方面代表美国的右派势力和共和党的利益。Daily Beast 是一个更加自由的平台，注重评论的原创性和网络的多样性。英国广播公司（BBC）、路透社（Reuters）、每日电讯（Daily Telegraph）为本文选取的欧洲媒体，目标人群各不相同，BBC 作为英国最大的新闻广播机构之一，以对新闻的严格要求著称。Reuters 是英国最大的通信社之一，主要报道各类新闻和金融数据。而 Daily Telegraph 作为欧洲电信业的重要成员，主要报道形式是报纸，其网站内容大多是新闻报纸的图片，目标人群为中产阶级。半岛新闻台（Al Jazeera）立足于阿拉伯半岛，主要报道中东地区的新闻。中东地区作为世界力量的重要组成部分，由于其地理位置、石油资源、饮食习惯、宗教文化等多方面原因，其新闻背后的疾病信息也与美国、英国等西方国家大为不同。

1.2 ICD-10 疾病数据源

《国际疾病分类》（International Classification of Diseases，ICD）是为了分析世界各国人口健康情况和死因，面对各种疾病做出的国际通用的统一分类[14]。ICD 是由世界卫生组织提出的，是确定全球健康趋势、统计数据、疾病和健康状况国际标准的基础，是大部分临床和研究活动的诊断分类标准[15]。

ICD-10-CM 是目前世界上应用最广泛的 ICD 修订本[15]，其中 CM 表示使用的版本是美国当地的版本。ICD-10-CM 收录了 77 545 种疾病和疾病症状。ICD-10-CM 呈现为一种类似 XML 的形式。其文件中存在 29 133 种一级条目，21 850 种二级条目，11 257 种三级条目，4 995 种四级条目，1 853 种五级条目，483 种六级条目，78 种七级条目，4 种八级条目[15]。为了构建 ICD-10 的疾病集，将结构化分层的数据转换为疾病名称。

1.3 MeSH 疾病数据源

医学主题词表（Medical Subject Headings，MeSH）是一部权威、规范的主题词表，每个记录均由 3 个字段组成，分别为 CUI、语义标识符、概念的语义。在 MeSH 中，CUI 表示在一体化医学语言系统（Unified Medical Language System，UMLS）内的标号，由一个英文字母和数字组成，其中以字母 D 开头的 CUI 表示疾病信息。另外，CUI 可以作为疾病的唯一标识，标识不同名称的相同疾病。所以，本文选取 CUI 作为疾病的替换码，可以更好地统计疾病的名称[16,17]。使用正则表达式可以得到所有类别标签为疾病的 CUI 和对应的语义。最后，得到 27 885 种不同的疾病相关词汇类别和 228 519 种不同的疾病相关词汇、词组。

1.4 构建疾病词汇表

MeSH 中存放了 228 519 种疾病相关的词组。很多词组本身并不是疾病名称，却在疾病预防、发病、治疗中发挥了不可忽视的作用，这些词汇也会被收录其中。所以，通过 ICD-10 的数据集合与 UMLS 集合做交集运算，筛选出 MeSH 中真正表示疾病名称的词汇和词组，作为本文的疾病词汇表。

1.5 信息预处理过程

信息预处理过程如图 2 所示。预处理过程分为两部分：疾病信息的预处理过程和新闻文本的预处理过程。

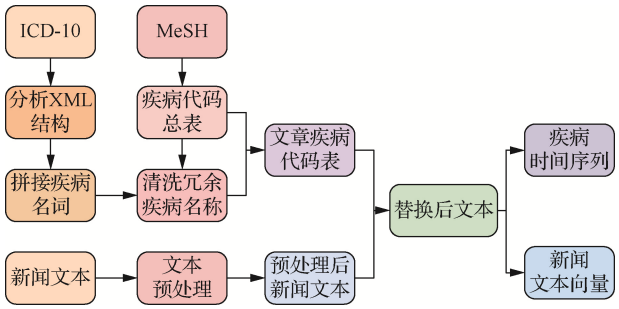

图 2　信息预处理过程

Fig. 2　Preprocessing flow chart

第一部分为疾病信息的预处理过程。在疾病信息的预处理过程中，综合了 ICD-10 和 MeSH 构建文章的疾病名词表。MeSH 中的大量词汇并不全是本文所需要的疾病名词，其还包括疾病相关症状等名词。从网站上下载 ICD-10 文件，使用正则表达式提取疾病的名称，将结构化分层的数据转换成疾病名称。拼接不同层次的词汇。利用 ICD-10 对 MeSH 中的疾病表进行过滤，获得 MeSH 的文章疾病代码表。

第二部分为新闻文本的预处理过程，在网页中爬取的新闻信息首先经过基本的文本预处理，再综合第一部分得到的文章疾病代码表将文本中的疾病名词替换为 UMLS 中的代码。替换的目的主要是避免疾病别名带来的统计差异。新闻中的疾病类别数如表 2 所示。

表 2　新闻中的疾病类别数

Tab. 2　Number of disease category in news

CNN	FOX	Daily Beast	BBC	Reuters	Daily Telegraph	Al Jazeera
105	144	82	96	136	111	57

2　模型

本文构建了深度学习融合模型，将新闻文本引入对疾病趋势的预测。模型上部分是针对新闻文本的处理过程，对新闻进行向量化表示后使用 CNN 的卷积池化，提取新闻之间的空间信息；模型下部分使用 LSTM 网络对疾病时间信息进行抽取。上下两部分经过融合层合并，经过 Softmax 层输出。融合模型如图 3 所示。

2.1　疾病时序模型

使用 LSTM 记忆元来处理新闻中排名前十的疾病的时间序列数据。在使用 LSTM 网络训练疾病的时序信息的过程中，数据从输入层流入模型，经过 LSTM 层综合时间信息，获取数据的特征向量。为了防止模型训练中的过拟合，在添加 LSTM 层时添加 Dropout 层，随机选取节点舍弃。

2.2　文本处理模型

对于新闻文本的处理，采用 CNN-LSTM 模型，CNN 用来提取文章中的特征值，LSTM 用来保留新闻文本的时序关系。具体模型如图 4 所示。

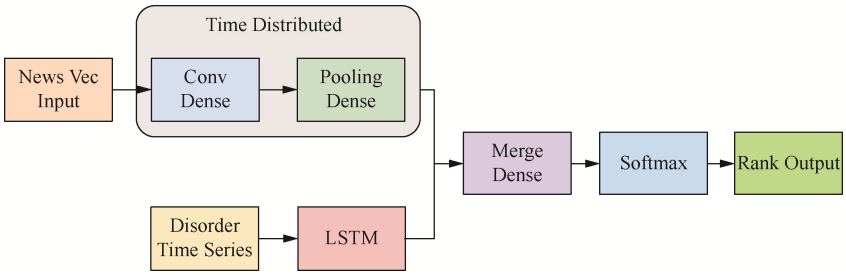

图 3　融合模型

Fig. 3　Merge model

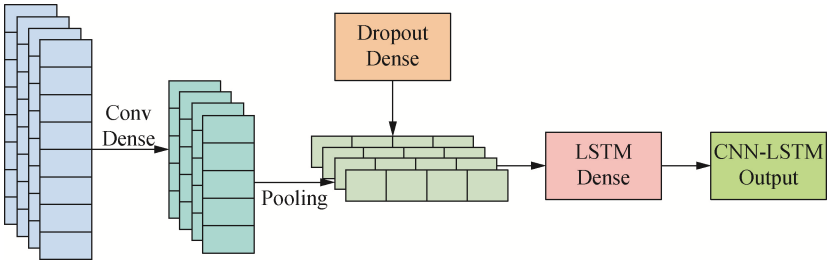

图 4　文本处理模型

Fig. 4　Text processing model

使用 Time Distributed 层将多天的新闻卷积池化，可以使模型具有多对一的能力。多对一表示有多个输入，但是只输出一个结果。在本模型中，输入层为 30 天的新闻向量，预测是下个月排名前十的疾病，将多层卷积池化后输出的特征矩阵作为 LSTM 层的输入，输出为 LSTM 网络对特征矩阵提取的信息。

由于卷积层和池化层的独特性，CNN 在自然语言处理方面具有巨大的潜力，可以捕捉句子的整体语义信息，并保留句子的位置信息[18]。不同于其他卷积处理，本模型使用按行卷积的方式。将新闻文本经过 Doc2Vec 处理获得的新闻向量输入 CNN-LSTM 模型中作为卷积层的输入，获取简单局部特征向量，通过池化层得到较深层次的特征向量。

3　实验和结果分析

3.1　实验数据

首先对预处理后的新闻文本中的疾病信息进行初步统计。以天为粒度统计新闻中的疾病数量，并统计总体新闻的疾病出现数据集合。其中，每行的疾病类别选自不同的新闻媒体。统计表 2 中的所有新闻文本中出现的全部疾病，同时对疾病出现的次数进行排序，选取在全部疾病中出现次数排名前五的疾病名称，其出现总次数如表 3 所示。

表 3　部分疾病出现总次数统计

Tab. 3　Statistics of total number of partial diseases

新闻媒体	HIV	AIDS	Depression	Polio	Gambling
CNN	5 207	3 445	5 238	1 555	1 116
FOX	11 936	11 086	17 797	1 798	5 586
Daily Beast	2 874	2 565	3 463	418	1 061
BBC	1 555	3 843	3 756	1 135	2 886
Reuters	20 303	16 982	23 112	2 973	16 139
Daily Telegraph	2 587	2 270	124	463	4 473
Al Jazeera	3 568	3 151	1 543	1 049	675

3.2　初步数据分析

以自然月作为粒度，计算每种疾病在所有新闻媒体中出现次数排名前十的疾病名词，构造自然月排名前十疾病列表的频度，根据列表中疾病出现的频度绘制疾病云图，如图 5 所示。图中的词语字体越大，说明其出现在新闻媒体中的频度越高。

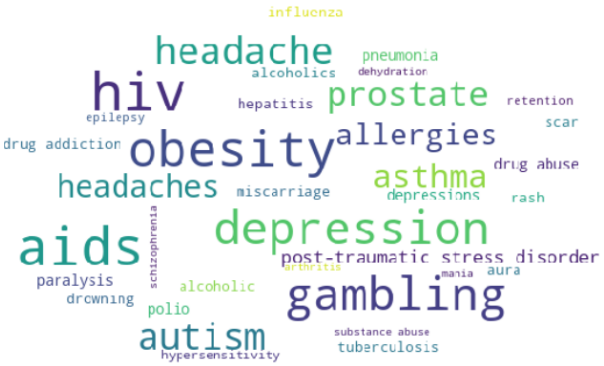

图 5　疾病云图

Fig. 5　Word cloud of diseases

在专业的医学表述中，HIV 和 AIDS 是两种不同的疾病，前者为艾滋病病毒，后者为获得性免疫缺陷综合征，是艾滋病病毒入侵人体很长时间之后，引起死亡免疫缺陷，进而引起其他病毒或细菌侵染的综合病症。但是，在面向大众的新闻媒体中，HIV 和 AIDS 通常都指艾滋病，所以在使用神经网络的趋势分析中，本文将 HIV 和 AIDS 视为一种疾病。

在以自然月对不同的新闻进行统计时，我们发现结核病在每年春季和秋季均存在高点，如图 6 所示。两个高点之间的距离较为固定（6 个月）。在大多数年份中，若春季结核病爆发早，则其在秋季的高点也会提前。这证明以结核病为代表的部分传染病存在时间上的规律性，也从侧面证明了 LSTM 方法的可行性。

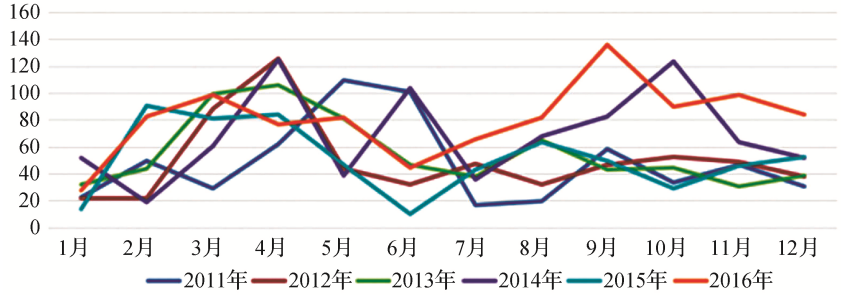

图 6　Al Jazeera 中每年结核病的折线图

Fig. 6　Trends of Tuberculosis From Al Jazeera

3.3　评价标准

规定在某媒体中出现的某疾病在未来 30 天排名前十的概率为该疾病的正样本概率。对 30 天的正样本概率进行排序，取正样本概率最高的 10 种疾病作为预测结果。命中率为预测正确的数量与预测目标的数量的比。

3.4　备选疾病策略

通常情况下，构建疾病集合包括多个年份期间新闻媒体中报道的所有疾病。但是，疾病是不断产生和发展的。例如，2020 年年初爆发的新型冠状病毒（COVID-19）并没有出现在之前的疾病列表中。因此，定义备选疾病和新发疾病这两个新的概念可以解决疾病发展和增加的问题。

备选疾病是指在每月排名前十的疾病中出现概率较高的疾病。新发疾病是指在每月排名前十的疾病中出现次数较少的疾病，可以被视为偶发疾病或突然爆发的疾病变种。模型的输入数据基于备选疾病和新发疾病的概念，改进 LSTM 模型的目标就变成了从新发疾病和多个备选疾病中预测排名前十的疾病。备选疾病集合选择所有月度排名前十疾病并集的前 85%所对应的疾病，选取的备选疾病数目如表 4 所示，其他疾病则统一作为偶发疾病。

表 4　选取的备选疾病数目

Tab. 4　Number of candidate diseases

新闻媒体	备选疾病数目	新闻媒体	备选疾病数目
CNN	23	Reuters	15
FOX	23	Daily Telegraph	23
Daily Beast	23	Al Jazeera	24
BBC	24		

3.5　实验细节及对比

首先，实验仅对每个新闻媒体中出现的所有疾病总数目 n，使用 n+1 个 LSTM 模型，仅利用疾病的趋势数据判断每个疾病是否出现在下个月排名前十的疾病中，可规约为二分类预测。然后，在改进的 LSTM 过程中采用备选疾病策略，结果表明，加入备选疾病策略后有效缩小了疾病预测空间，进一步提高了命中率。最后，结合新闻语义信息和疾病趋势数据，使用 Doc2Vec 对每日新闻进行编码，将每日的 50 篇新闻折叠为 CNN-LSTM 模型所需输入的张量。其中，单天新闻数量不足 50 篇的，使用零矩阵填补；单天新闻数量多于 50 篇的，删除多余的新闻文本向量。将 CNN-LSTM 新闻文本向量与带有备选疾病策略的 LSTM 的疾病趋势向量合并，综合判断疾病的趋势。

在实验过程中，总共训练了 1 280 254 个网络隐含层参数，其中单天新闻最多 356 篇，使用前 70%的数据作为训练集合，后 30%的数据作为测试集合。CNN-LSTM 模型的输入为步长为 1 天的 30 天的新闻向量（362，30，356，300），300 为新闻所对应的 Doc2Vec 编码。单天新闻少于 356 篇的，使用（1，300）的零矩阵将其填补为 356。LSTM 模型的输入作为新闻对应天的所有疾病数目，将 30 个单天的排名前十疾病出现次数矩阵折叠，其输入矩阵规模为（362，30，95）。本实验使用 24 种备选疾病和一种新发疾病，共计 25 个分类模型，即模型运行 25 次，分别判断某个疾病是否为排名前十的疾病。不同模型的对比结果如表 5 所示。结果表明，融合模型在具有备选疾病策略的模型的基础上进一步提升了命中率，有效地利用了新闻中的隐含语义。

表 5　不同模型的对比结果

Tab. 5　Target score of different models

新闻媒体	LSTM	LSTM 备选疾病	融合模型
CNN	0.741	0.898	0.91
FOX	0.745	0.76	0.77
Daily Beast	0.698	0.713	0.754
BBC	0.713	0.717	0.743
Reuters	0.773	0.834	0.865
Daily Telegraph	0.624	0.65	0.71
Al Jazeera	0.758	0.84	0.86

4　结语

CNN 模型和 LSTM 网络模型在处理自然语言问题方面有天然的优势，降低了对人工的依赖。本文主要研究了利用 CNN 和 LSTM 预测新闻文本疾病的方法，对社会上出现的大规模常见疾病进行预测。通过获取全球各大新闻媒体报道的新闻文本，分别统计新闻文本中出现次数排名前十的疾病，分析原始数据分布的特征。本文搭建了一个新闻数据集来存放所爬取的新闻信息，新闻数据集中的数据来自 CNN、FOX、Daily Beast、BBC、Reuters、Daily Telegraph、Al Jazeera。这些代表不同国家和政党利益的新闻媒体几乎囊括了全球的新闻。为了确保疾病种类的正确性和广泛性，本文综合 UMLS 和 ICD-10 的数据形成了一个疾病数

据集，得到疾病名称 29 643 种。将所获得的新闻数据中的疾病名称替换为特殊的码值，再将所获得的新闻文本作为基本新闻数据集合。基本新闻数据集合构建之后，统计基本数据集中的疾病关键词并分析数据背后的隐含意义。对新闻中排名前十的疾病建立 LSTM 模型。新闻文本通过 Doc2Vec 转化为新闻向量，对新闻向量和疾病新闻个数构建合并的 CNN-LSTM 模型。实验结果表明，候选疾病和新闻向量的引入有效地提升了模型的命中率，可以更有效地观察并预测疾病的发展趋势，为疾病防控部门制定相关政策提供有效的依据。本文针对不同的新闻媒体构建了不同的模型，后续将整合更多的新闻数据，探索一种基于复杂网络的新闻和疾病趋势协同作用的模型，进一步提升模型的预测效果。

参考文献

[1] PEW Internet. Health fact sheet[EB/OL]. http://www.pewinternet.org/fact/sheets/health-fact-sheet/.

[2] LAVRENKO V, ALLAN J, DEGUZMAN E, *et al*. Relevance models for topic detection and tracking[C]. Proceedings of the Human Language Technology Conference (HLT). San Francisco: Morgan Kaufmann Publishers Inc., 2002:104-110.

[3] LIU L, OZA S, HOGAN D R, *et al*. Global, regional, and national causes of child mortality in 2000-13, with projections to inform post-2015 priorities: an updated systematic analysis[J]. **The Lancet**, 2015, 385(9966): 430-440.

[4] TOWERS S, AFZAL S, BERNAL G, *et al*. Mass media and the contagion of fear: The case of Ebola in America[J]. **PLOS ONE**, 2015, 10(6): e0129179.

[5] FOEGE W H, HOGAN R C, NEWTON L H. Surveillance projects for selected diseases[J]. **International Journal of Epidemiology**, 1976, 5(1): 29-37.

[6] Centers for Disease Control and Prevention. National notifiable diseases surveillance system (NNDSS)[EB/OL]. https://www.cdc.gov/nndss/index.html, 2022.

[7] WARD J, HILDEBRANDT C, PATEL A. NEDSS base system (NBS): electronic data exchange and workflow decision support[J]. **Online Journal of Public Health Informatics**, 2017, 9(1): 47.

[8] HUANG D C, WANG J F. Monitoring hand, foot and mouth disease by combining search engine query data and meteorological factors[J]. **The Science of the Total Environment**, 2018, 612: 1293-1299.

[9] 李薇，杨晴晴，刘红英. 天津市河北区 2013—2020 年疑似预防接种异常反应监测分析[J]. **生物医学工程学进展**，2023，44（1）：73-81.
LI Wei, YANG Qingqing, LIU Hongying. Surveillance and Analysis of Adverse Events Following Immunization in Hebei District, Tianjin City from 2013 to 2020[J]. **Progress in Biomedical Engineering**, 2023, 44(1):73-81.

[10] TENKANEN H, MININ E D, HEIKINHEIMO V, *et al*. Instagram, Flickr, or Twitter: assessing the usability of social media data for visitor monitoring in protected areas[J]. **Scientific Reports**, 2017, 7(1): 17615.

[11] SHIN S, SEO D W, AN J, *et al*. High correlation of Middle East respiratory syndrome spread with Google search and Twitter trends in Korea[J]. **Scientific Reports**, 2016, 6(1): 32920.

[12] XU Q, GEL Y R, RAMIREZ L L R, *et al*. Forecasting influenza in HongKong with Google search queries and statistical model fusion[J]. **PLOS ONE**, 2017, 12(5).

[13] 白金川，王豪，焦宝园，等. Python 网络爬虫在医学影像领域的发展现状与趋势研究[J]. **生物医学工程学进展**，2023，44(3): 260-266.
BAI Jinchuan, WANG Hao, JIAO Baoyuan, *et al*. Research on the Development Status and Trends of Python Web Crawlers in Medical Imaging[J]. **Progress in Biomedical Engineering**, 2023, 44(3): 260-266.

[14] 李亚鹏，刘启伟. 关于提高病案管理的方法及效率的研究[J]. **齐齐哈尔医学院学报**，2013，34（11）：1666-1668.
LI Yapeng, LIU Qiwei. Research on improving the methods and efficiency of the medical record management[J]. **Journal of Qiqihar Medical College**, 2013,34 (11): 1666-1668.

[15] WHO International Classification of Diseases,11th Revision (ICD-11) [EB/OL]. (2022-05-22) [2023-06-26]. https://www.who.int/classifications/icd/en/.

[16] Unified Medical Language System (UMLS)[EB/OL]. (2022-05-22) [2023-06-26]. https://www.nlm.nih.gov/research/umls/index.html.

[17] 董小芸. 基于一体化医学语言系统（UMLS）的语义检索实验研究[D]. 上海：上海大学，2004.
DONG Xiaoyun. Experimental study on semantic retrieval based on the Integrated Medical Language System (UMLS)[D]. Shanghai: Shanghai University, 2004.

[18] 孔晓风，李莹，李昊旻，等. 基于自然语言处理技术的消化科内窥镜检查报告的结构化[J]. **中国医疗器械杂志**，2008（5）：44-47.
KONG Xiaofeng, LI Ying, LI Haomin, *et al*. Structured reports of gastroenterology endoscopy based on natural language processing technology[J]. **Chinese Journal of Medical Instrumentation**, 2008 (5): 44-47.