基于 SVM 模型实现 GHSI 科技论文的情感识别

王雅娇　曾骏程　李俊翰　崔基哲*

（延边大学）

摘要：全球卫生安全指数（Global Health Security Index，GHSI）是 2019 年美国约翰·霍普金斯大学发布的用于国家级别预防、检测和应对传染病威胁能力的综合性评价工具，在指导如何防范和应对突发公共卫生事件的长期机制方面存在重要的指导意义。但对照当前卫生安全传播的实际背景下，该指数的预测能力与实际情况有较大偏差。故本文以 Web of Science 为文献检索数据库，对站内关于 GHSI 的文献进行筛选整理，并基于 SVM 模型，识别不同文献中对于 GHSI 的情感态度，并与人工分类结果进行比对改正，进一步分析得出文献作者对于 GHSI 的情感态度及其深层原因。以此为 GHSI 的改进方向提供理论支持，分析该指数对公共卫生领域的可借鉴程度，对后疫情时代世界范围内推广 GHSI，并统一公共卫生领域指标有较强的积极意义。

关键词：全球卫生安全指数；GHSI；SVM 模型；情感识别

Emotion recognition of GHSI scientific papers based on SVM model

Wang Ya-jiao　Zeng Juncheng　Li Junhan　Cui jizhe*

(Yanbian University)

Abstract: The Global Health Security Index (GHSI) is a comprehensive assessment tool for national capacity to prevent, detect and respond to infectious disease threats, which was released by Johns Hopkins University in 2019. It has important guiding significance in guiding long-term mechanisms on how to prepare for and respond to public health emergencies. However, compared with the actual background of the current health security communication, the prediction ability of the index is significantly different from the actual situation. Therefore, in this paper, Web of Science is used as the literature retrieval database, the literature on GHSI is screened and sorted, and based on the SVM model, the emotional attitude towards GHSI in different literatures is identified, and compared and corrected with the manual classification results, and the author's emotional attitude towards GHSI and its deep causes are further analyzed. This provides theoretical support for the improvement direction of GHSI, analyzes the referable degree of this index to the field of public health, and has a strong positive significance for promoting GHSI worldwide in the post-epidemic era and unifying the indicators in the field of public health.

Key words: Global Health Security Index; GHSI; SVM model; Emotion recognition

1 引言

2019 年 10 月美国约翰·霍普金斯大学发布《2019 全球卫生安全指数》报告（下文简称为 Global Health Security Index，GHSI）[1]。该报告是首个结合多组织资料报告创立的用以评估各国对大流行病的防范能力的指数。世界各国家、各地域组织建立的公共卫生机制在应对突发公共卫生事件中发挥着至关重要的作用。而全球卫生安全指数正是用于国家级别预防、检测和应对传染病威胁能力的综合性评价工具。但在 GHSI 被颁布后，各国学者对全球卫生安全指数的评价褒贬不一。一方面，有学者认为 GHSI 不仅作为一篇启发性的卫生安全评估报告，而且提供了相关数据库，里面详细记载了 GHSI 的评估细节，列举了各国评分的依据和参考文献：等等，其详尽的数据以及启发性的指标对于各国的医疗改革有着重要的指导意义；另一方面，也有学者认为 GHSI 聚焦模糊，在公共卫生领域的形式大于实质，并带有一定“以西方为尺”的主观性，缺乏对于公共卫生事件的整体把控。因此本文在现有的研究基础上，基于 SVM 模型对于各文献进行情感识别并进行分类，实现对 GHSI 的推广以及改进的数据支持。

2.研究思路

本文首先将从 Web of Science 站内收集各学者针对 GHSI 的相关文献进行整理，对得到的文献数据进行断句处理，对其中的 100 条来源于不同文献的语句进行人工判断。在此基础上，基于 SVM 模型对各文献的情感识别，再与人工判断结果进行比对分析得出深层原因，研究过程包括文献数据构建、识别模型生成、建立 SVM 模型、对比人工结果、GHSI 情感分析等。具体研究过程及研究意义如下图所示：

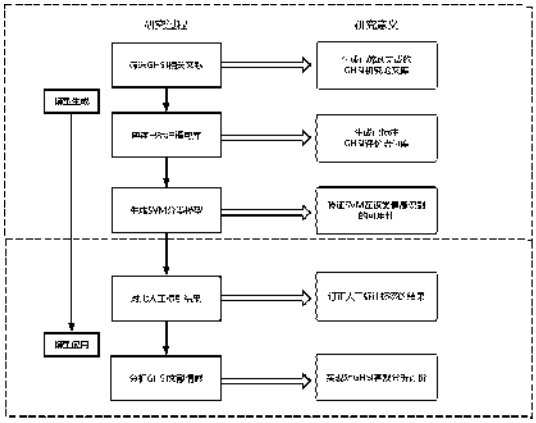

图 1　研究过程及研究意义

3.文献数据构建

3.1 数据集构建

本文构建了文献摘要集和摘要句子集两个层次的文献数据库，该数据来源于 Web of Science 文献数据库，以 GHSI 为主题词检索得到。考虑到文献的可获得性和情感分析的价值性，本文选择该批文献的摘要进行分析。通过条件筛选，共获得相关文献 55 篇。筛选剔除部分无摘要的文献，和有相关主题词但与该评价指标并无直接关系的相关文献，共获得文献 51 篇，由此构建文献的摘要数据库。在摘要数据集的基础上，本文对摘要中的句子进行了分句处理，进而形成句子集的文献数据库，共分句得到 482 句。

3.2 情感标注

本文对摘要句子集数据库的每条数据，逐句进行了情感标注。本文所指的情感标注是指该句话对 GHSI 的态度是积极的，中性的还是消极的。本文标注语句的具体依据如下：积极语句（pos），指该条语句是对 GHSI 指数进行肯定或利用该指数进一步预测研究；消极语句（neg），指该条语句是对 GHSI 指数进行否定或说明其不符合实际情况；中性语句（neu），指该条语句是对论文背景的描述或该句不包含对 GHSI 的具体描述等。基于上述情感标注的原则，本文对于每条数据进行了一轮态度标注。

3.3 文本预处理

对 GHSI 的文献信息文本进行预处理的目的是将无用的或可能干扰分类效果的因素从数据集中剔除出去。文献的预处理中最主要的环节是分词，针对英文语句的复杂性比较适用的分类模型主要包括以下几种：①大小写处理；②词干提取和词型还原：这是英文文本预处理的特色。两者其实有共同点，即找到词的原始形式，使得文本能还原成一个正确的词再进行处理；③设立停用词表，并剔除停用词，进而完成文本预处理过程。

4.情感识别模型

4.1 模型概述

支持向量机（Support Vector Machine，SVM）本质是一类按监督学习方式对数据进行二元分类的广义线性分类器[2]。给定一组训练实例，每个训练实例被标记为属于两个类别中的一个或另一个，SVM 训练算法创建一个将新的实例分配给两个类别之一的模型，使其成为非概率二元线性分类器。模型是将实例表示为空间中的点，这样映射就使得单独类别的实例被尽可能宽的明显的间隔分开，将新的实例映射到同一空间，并基于它们落在间隔的哪一侧来预测所属类别。

4.2 模型构建

该模型的输入为摘要句子集数据库的每条数据，及该条分句的标注情感标签，该模型的输出为对每个分句的标签预测值。在模型的调试及参数设定中，采用 70%的数据训练模型的样本集合训练模型参数，和 30%测试集用来对模型效果进行评估。

4.3 数据输入

对文本分类来说，特征就是文本中表达了文本类别属性的词语，因此特征的选择较大程度上决定了文本分类效果的好坏。经过对预处理完成之后的文本进行人工的判断的得到的训练数据集表示{T_1，T_2，T_3}分别代表积极的，中性的和消极的三个类别的数据集，第 i 个数据集：T_i={（a_{i1}，b_{i1}），（a_{i2}，b_{i2}），...，（aim，bim）}，其中 a_{ij}表示第 i 个数据集第 j 个文本的词向量，b_{ij}表示第 i 个数据集第 j 个文本的是否为积极的，是则取值为 1，不是则取值为 0。分别用 SVM 算法对数据{T_1，T_2，T_3}进行训练得到对应的分类器。SVM 训练成功后。利用训练好的 SVM 分类器对预处理之后的测试数据集进行关于 GHSI 文献的情感识别，以确定参试文献的情感的积极与否。从而生成已标注 GHSI 的评价语句库。

4.4 模型结果

通过上述情感识别模型得到如下预测结果，该结果图显示 neg、neu、pos 三类标签的真实值和预测结果的对比，通过该图可以认为该模型具有一定的预测能力，在预测中性语句时具有较强的准确性。与此同时，该模型的测试准确率的平均值达到 79%，进一步验证该类型的问题可以通过 SVM 模型进行合理的预测分类，并可以得到较为理想的效果。

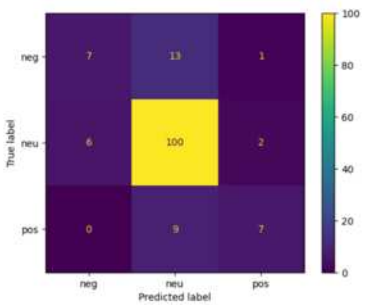

图 2　模型结果对比图

4.5 模型应用

本文基于上述情感识别模型的搭建，进一步采用该模型的输出结果，即每个分句的 SVM 标签预测值与初始人工标注的情感标签进行比对。以期望通过该模型校正人工标注的错误。并对初始值与预测值不同的分值进行人工筛选，进一步核查。发现有极少句子是由于人工标注疏忽所导致的标注错误由 SVM 模型可以进行订正，有部分错误则是由于模型的学习能力不强所导致。进一步提高了文献摘要集和摘要句子集情感识别的准确性。

5. GHSI 情感分析

根据文献摘要集数据库和摘要句子集数据库之间的交互比对，可以得出在 Web of Science 网站中所获取的对于 GHSI 的文献研究中，有 78.43%的文献对于 GHSI 持有中立态度，此类文献中对于 GHS 指标大多数是予以学习性的描述，或者是分析该指数在新冠疫情中所展现出来的辅助作用，此类中性立场文献大多数更关注的是 GHSI 所提出的标准是否科学，但对标准于实际情况中实施是否顺利，未收到外界影响并没有做出太多分析。

有 13.73%的文献对于 GHSI 进行了批判与主观层面的修正，持有消极意见的学者指出在最近的新冠肺炎疫情中 GHS 指标评级情况与实际情况截然不同，代表性的就是 GHS 报告中将美国和英国评为应对灾难性大流行的最佳国家，但在实际的应对公共卫生突发事件的表现并不突出。这些主观性评价指标狭隘地将全球卫生安全概念化为检测新出现的传染病并防止其传染的技术基础设施的可用性，但深刻地低估了公共卫生的更广泛的社会和政治决定因素，使得 GHSI 有“以西方为尺”，评判世界的嫌疑，因此研究认为 GHSI 的指数标准分类较少，其中应包括其他社会人口、政治和治理变量，以提高其表述备灾情况的能力。

最后 7.84%的全球卫生安全指数研究文献对于 GHSI 在疫情期间的贡献做出了肯定，认为 GHSI 作为一个首创的指数为世界强调了公共卫生方面的重要性，并在引导构建完善具有强大的预防、检测和应对疫情能力的国家安全卫生体系方面有效降低了突发危机所带来的许多社会、政治、经济和卫生系统成本，并且有研究表示官方尝试结合 GHS 报告与全民健康覆盖指数（UHCI）对于 GHSI 所创立之后所出现的问题给出协同解决方案，结合初级卫生保健原则，致力于加强公共卫生系统，实施“一个健康”方针，将逐步推进各国能够实现全民健康和全球统一卫生制度。

6.结论

本文通过构建文献摘要集和摘要句子集两个层次的文献数据库，筛选 Web of Science 关于 GHSI 相关文献，对 51 篇文献中的 482 条语句进行情感分析，建立了正确的分析评价体系，验证了 SVM 在该类情感识别的可用性，反映出 GHSI 评价指标在一些方面的不足和缺点，指出当今的评价指标应更加全面和系统，对 GHSI 的提供一定的改进思路，从而推动全球卫生安全高效进行，加快全球卫生安全建设的进程，使 GHSI 在公共卫生领域发挥更大价值。

参考文献：

[1]NTI，Johns Hopkins，Economist Impact. 2019 Global Health Security Index [EB/OL]. https：//www.ghsindex.org/.

[2]Sholkopf B，Sung K，Burges C J C，et al. Comparing support vector machine with Gaussian Kernels to radial basis function classifiers [J].IEEE Trans，Signal Processing，1997，45：2758-2765.

基金项目：2021 年度吉林省大学生创新创业计划训练项目延边大学大学生创新创业训练资助项目“国际评价指标体系与制度性话语权关系的探究——以疫情期间 GHS 指数为例”（项目编号：202110184108）