

1. 3315 软件架构方向的补充与更新

Ovidiu Serban 等学者提出了一个结合了深度学习的基于 Twitter 的健康分类综合监控系统, 该系统可以从 Twitter 数据中检测出疾病爆发, 并建立以及显示有关这些爆发的信息 [1]。3315 项目所提出的“基于大数据和人工智能技术面向新型传染病的重大场景融合预测预警系统”参考了其提出的系统架构, 并对其进行了完善与改进。例如, Ovidiu Serban 等提出的框架在数据来源上仅仅使用 Twitter 数据和英文新闻数据, 而 3315 项目所提出的项目在数据来源上更加多源, 包括: 中英文的新闻数据, 中英文的医疗期刊数据以及中英文的社交媒体数据 (使用百度舆情服务) 等。

3315 提出的预测预警系统是一个 web 应用, 系统部分包括服务器以及数据库 (MongoDB) 的搭建, 并拟采取前端后端分离的方式进行开发。在前端方面上, 通过多种可视化技术, 将传染病预测预警信息可视化, 拟采用百度可视化开源库 Echarts 以及 Twitter 公司的前端开源框架 Bootstrap4 进行开发; 在后端上, 主要分为多个模块, 具体包括: 多源大数据信息的收集模块 (爬虫, 百度舆情服务, CDC 数据), 大数据处理模块 (数据清洗, 数据预处理), **自然语言处理分类模型模块 (机器学习模型, 深度学习模型等)**, 传染病预警分析处理模块。此外, 本项目拟在后端编写 RESTFUL 接口对接前端, 前后端间的数据以 JSON 格式进行传输。

3315 预测预警系统会定期从社交媒体上 (国内的微博, 国外的 Twitter), 中国疾控中心网站, 世界卫生组织网站等, 以及专业的医疗期刊杂志获取多源数据, 然后使用大数据处理模块进行数据处理, 并将处理后的数据输入到自然语言处理模型中。自然语言处理分类模型, 使用监督学习的方法将文章进行分类, 模型会将文本分类为“与传染病有关”和“与传染病无关”两类, 并从“与传染病有关”的文章中进行信息提取 (information retrieval), 如提取时间信息、地理信息、传染病名称信息、传染病症状等信息。下一步, 系统会将在“与传染病有关”等文本中提取到的关键信息发送到传染病预警分析处理模块。在此模块中, 我们将录用各类传染病的特征信息并配合专业的医生以及医疗专家等设定预警规则 (如使用阈值法等) 来分析统计各类信息, 从而完成对预警信息的处理。例如, 如果 3315 软件在 2019 年建成, 那么本软件传染病预警分析处理模块会在 2019 年 12 月份多次监测到如“咳嗽”, “肺炎”等信息, 并且地理位置为“武汉”, 那么该预警分析模块就会将“咳嗽”与“肺炎”等信息与“非典”联系起来, 从而将会预警武汉在 2019 年 12 月可能有“非典”传染病爆发的情况。

2. 大数据 NLP 科研方向的贡献与创新点

在大数据 NLP 科研方向上，创新点主要为数据源的创新，自然语言处理文本分类模型的创新以及传染病预警分析方法的创新，具体描述如下。

2.1 传染病相关的多源数据集(语料库): 目前在 text classification 这个方向上比较流行的数据集为: Reuters 数据集 [2] 和 arXiv Academic Paper 数据集 [3] 等。就我们所知，在医疗健康方向，特别是传染病相关的文本分类方向，并没有公开的数据集。因此本项目通过多源数据收集模块，使用网络爬虫，Twitter 数据，CDC 数据等，将构建一个中文英文两种语言的传染病相关的多源数据集。如果项目将来开源一部分数据集，**即可为后续相关研究人员研究医疗健康类的文本分类 (text classification) 提供来源可靠的数据集**，在领域内作出贡献。

2.2 自然语言处理文本分类模型:自然语言处理文本分类 (text classification)是指为句子 或文档分配适当类别的任务。模型的输入为一个文件 d (通常表示为特征向量形式)，以及固定的输出类别 $C = \{c_1, c_2, \dots, c_k\}$ (在本项目为“与传染病有关”和“与传染病 无关”两类)，模型的输出为预测结果。如下图 1 所示，一个文本分类模型通常包括特征提取 (Feature Extraction)，降维 (Dimensional Reduction)，分类 (Classification) 以及模型评估 (Evaluation)。本项目在自然语言处理文本分类模型上的创新点为在分类器上：**创建新的基于 CNN+MLP 的深度学习网络模型**并用来进行传染病的文本分类任务。

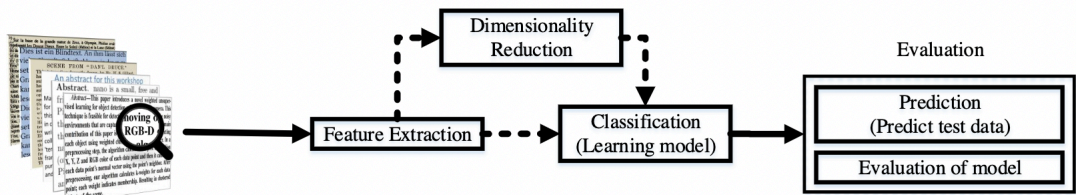


图 1: 文本分类模型架构图 [4]

比较常见的分类模型为机器学习模型如:朴素贝叶斯分类器 (Naïve Bayes Classifier), K 近邻(K-Nearest Neighbour) 以及支持向量机 (Support Vector Machine) 等。在本项目中，我们将以这些机器学习的方法作为 baseline，并在我们组建的数据集上进行实验。深度学习中的卷积神经网络 (CNN) 原本是用来应用在机器视觉中的图像处理任务，但是近年来，CNN 在自然语言处理中的文本分类任务上，也表现出了出色的成绩，比如，Alexis Conneau 等提出了一个在词层面上的 CNN 模型 [5] 在文本分类上取得了不错的成绩; Rie Johnson 等则提出了一个使用 LSTM 作为局部词嵌入的 CNN 网络用来进行文本分类 [6]。Alexis Conneau 的模型目前在文本分类任务是现有技术的最高水平之一 (SOTA)，因此我们拟基于 Alexis Conneau 模型上提出四点创新：

2.2.1 Alexis Conneau 使用的 Encoded Characters 的特征，而我们将在特征提取上使用 TF-IDF 与 Encoded Characters 相结合的方式。特征提取的基本思路是根据某个评价指标独立的对原始特征项（词项）进行评分排序，从中选择得分最高的一些特征项，过滤掉其余的特征项。Encoded characters 使用的是字符嵌入与词嵌入相结合的方法，目的是将文本数据转化为数值数据，获取一个单词的数值表示，并使用这些数值向量得到句子/段落/文本等数值表示。然而训练字符嵌入在计算上十分昂贵，因此我们考虑结合 TF-IDF 方法。Kowsari 等人在他们提出的文本分类模型中，就用到了 TF-IDF 作为特征提取方法 [7]。TF-IDF 方法中 TF 为 Term Frequency 用来表征一个词的重要度与在类别内的词频成正比，而 IDF 为 Inverse Document Frequency 用来表征一个词的重要度与所有类别出现的次数成反比。TF-IDF 可以很好的表达特征权重，因此我们认为 TF-IDF 与 Encoded Characters 相结合的特征提取方式会取得更好的效果。

2.2.2 Alexis Conneau 在池化方法中使用 K-Max pooling，我们将使用 Chunk-Max pooling 的方法进行改进实验，看能否取得更好的效果。池化（Pooling）是卷积神经网络中的一个重要的概念，它实际上是一种形式的降采样。有多种不同形式的非线性池化函数，池化层会不断地减小数据的空间大小，因此参数的数量和计算量也会下降，这在一定程度上也控制了过拟合。Max-pooling 即最大池化是对邻域内特征只求最大，而 K-max pooling 可以取每一个 filter 抽取的一些列特征值中得分在前 K 大的值，并保留他们的相对的先后顺序。Chunk-Max pooling 是把某个 filter 抽取到的特征向量进行分段，切割成若干段后，在每个分段里面各自取得一个最大特征值，比如将某个 filter 的特征向量切成 3 个 chunk，那么就在每个 chunk 里面取一个最大值，于是获得 3 个特征值。相比于 K-Max pooling，Chunk-Max pooling 的方法能够保留了多个局部最大特征值的相对顺序信息。因此我们期望使用 Chunk-Max pooling 的方法能够取得比 K-Max pooling 更好的表现。

2.2.3 Alexis Conneau 的输入是新闻类的长文本，而我们则将使用“基于规则”的方法，缩短输入的文本长度。如果一篇文章的内容过长，很可能一些不必要的文章内容会影响文章的主体思想判断。因此结合人们阅读文章的经验，我们认为一篇文章的第一句话，或者是一篇论文的 abstract 通常会涵盖整篇文章的内容。在模型实验中，我们将使用“基于规则”的方法，自定义文本处理流程，使用文本第一句话或者 Abstract 代替文章整体，使得输入的文本既保留原意，又缩短了文本长度，我们希望通过这种方法能够提升模型的准确率。

2.2.4. 多层感知机 (MLP) 是一个多层神经网络，包括输入层，隐藏层和输出层。我们的实验将另外训练一个多层感知机模型，并将全新的 CNN 模型（如图 2 所示）与多层感知机模型 (MLP) 相融合，然后将它们的预测结果取均值，提出一个 CNN+MLP 的融合后的新模型（如图 3 所示）。

综上，我们将基于以上四点创新提出一个全新的 CNN + MLP 的深度学习网络模型。我们将使用准确率（accuracy）作为模型评估的方法，预计该模型在我们的数据集上将取得 90 以上的准确率。

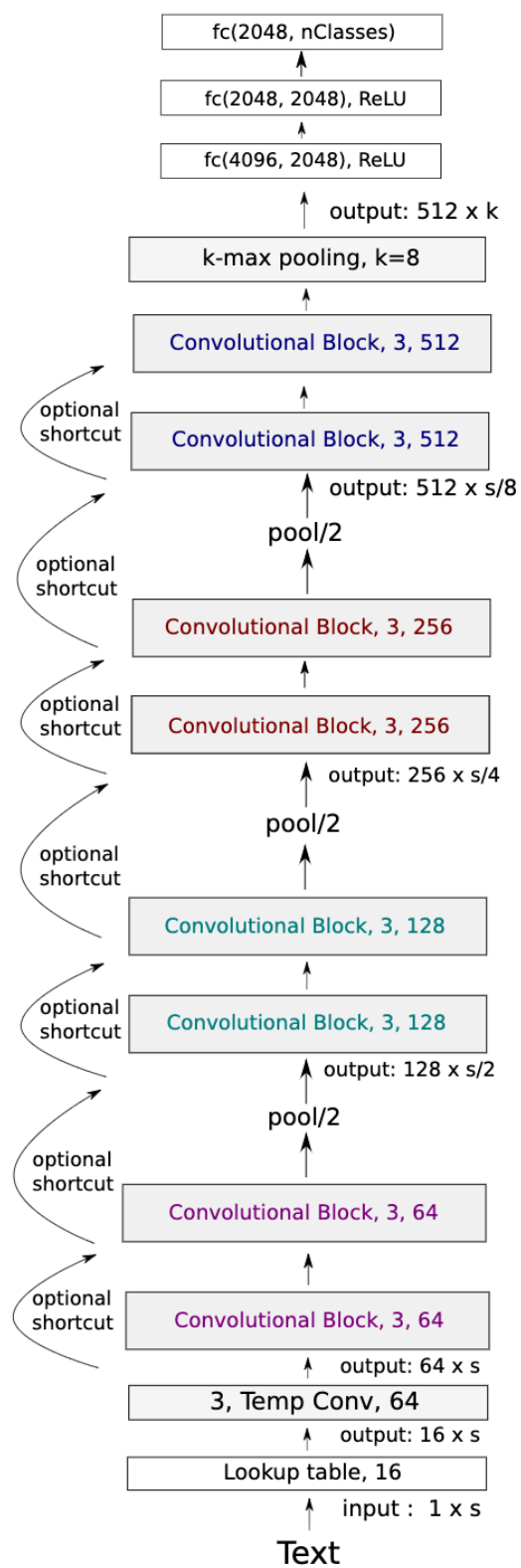


图 2: 文本分类模型使用卷积神经网络架构示意图 [5]

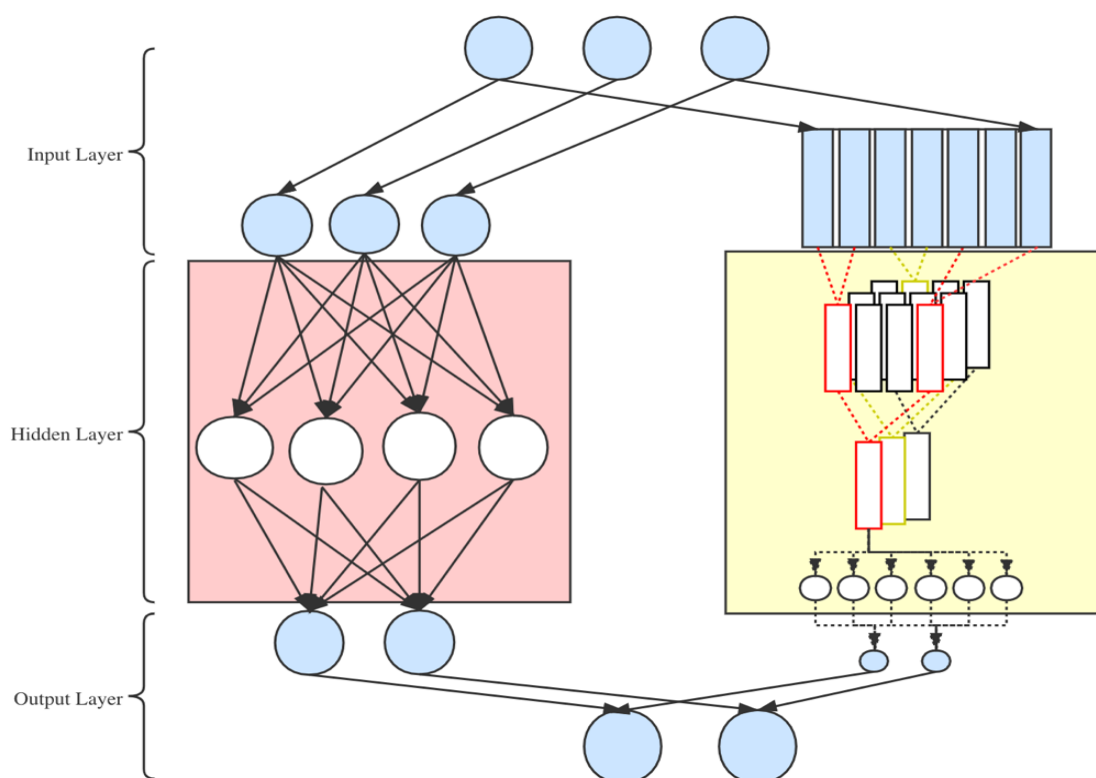


图 3:CNN+MLP 模型融合架构示意图

2.3 传染病预警分析方法: 3315 团队中有英国的传染病学教授，因此，我们团队可以针对利用自然语言处理所提取的关键信息进行专业的分析处理，**并提出新颖有效的预警分析方法**。例如，我们可以通过专业分析，提出在新闻报道以及社交媒体中的哪些关键词与“新型冠状病毒”(COVID-19)最相关，如果自然语言处理的模型提取到了很多同样的关键词或者与其近似的词，那么我们的预警分析模块会进行相应的预警处理。此外，我们也可以通过实验，探究如何设定阈值，例如是否当某一地区针对某个传染病有关的 Twitter 或微博数量达到 20 条以上才进行预警等。

参考文献

- [1] O. Şerban, N. Thapen, B. Maginnis, C. Hankin, and V. Foot, “Real-time processing of social media with sentinel: A syndromic surveillance system incorporating deep learning for health classification,” *Information Processing & Management*, vol. 56, no. 3, pp. 1166– 1184, 2019.
- [2] C. Apté, F. Damerau, and S. M. Weiss, “Automated learning of decision rules for text categorization,” *ACM Transactions on Information Systems (TOIS)*, vol. 12, no. 3, pp. 233– 251, 1994.

- [3] P. Yang, X. Sun, W. Li, S. Ma, W. Wu, and H. Wang, "Sgm: sequence generation model for multi-label classification," *arXiv preprint arXiv:1806.04822*, 2018.
- [4] K. Kowsari, K. Jafari Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, "Text classification algorithms: A survey," *Information*, vol. 10, no. 4, p. 150, 2019.
- [5] A. Conneau, H. Schwenk, L. Barrault, and Y. Lecun, "Very deep convolutional networks for text classification," *arXiv preprint arXiv:1606.01781*, 2016.
- [6] R. Johnson and T. Zhang, "Supervised and semi-supervised text categorization using lstm for region embeddings," *arXiv preprint arXiv:1602.02373*, 2016.
- [7] Kowsari, Kamran, et al. "Hdltex: Hierarchical deep learning for text classification." 2017 16th IEEE international conference on machine learning and applications (ICMLA). IEEE, 2017.