Digital Health and Informatics Innovations for Sustainable Health Care Systems
J. Mantas et al. (Eds.)

doi:10.3233/SHTI240425

A Multi-Label Text Classifier at Publication Level Based on "PubMedBERT + TextRNN" for Cancer Literature

Zhang YING[a], Xia GUANGHUI[a,1], Li XIAOYING[a] and Tang SHISHI[a]
[a] *Institute of Medical Information, Chinese Academy of Medical Sciences, China*
ORCiD ID: Xia Guanghui https://orcid.org/0000-0003-4587-0344

Abstract. There is a rapid growth in the volume of data in the cancer field and fine-grained classification is in high demand especially for interdisciplinary and collaborative research. There is thus a need to establish a multi-label classifier with higher resolution to reduce the burden of screening articles for clinical relevance. This research trains a multi-label classifier with scalability for classifying literature on cancer research directly at the publication level. Firstly, a corpus was divided into a training set and a testing set at a ratio of 7:3. Secondly, we compared the performance of classifiers developed by "PubMedBERT + TextRNN" and "BioBERT + TextRNN" with ICRP CT. Finally, the classifier was obtained based on the optimal combination "PubMedBERT + TextRNN", with P= 0.952014, R=0.936696, F1=0.931664. The quantitative comparisons demonstrate that the resulting classifier is fit for high-resolution classification of cancer literature at the publication level to support accurate retrieving and academic statistics.

Keywords. Text classification; Publication level classifier; Cancer literature; PubMedBERT; TextRNN

1. Introduction

There is a rapid growth in the volume of literature published in the cancer field [1]. However, most of the existing literature classification (WOS, Scopus) were usually carried out at the journal level [2]. There is a need for a more precise classifier as the articles from one journal always present a diverse range of topics [3]. In this research, we introduce a method of multi-label classification at the publication level for cancer research, based on the model of "PubMedBERT + TextRNN".

2. Methods

The overall framework of the study is illustrated in Figure 1. Firstly, a set of corpuses consisting of the publications titles and abstracts captured from PubMed database, will be divided into a training and a testing subset at a ratio of 7:3 after pre-processing.

[1] Corresponding Author: Xia Guanghui; E-mail: xiagh@imicams.ac.cn.

Secondly, in order to capture sufficient text features for multi-label classification, the titles and the abstracts of cancer publications will be taken separately as two independent layers, which would be called the tuple in this study. Finally, a “PubMedBERT + TextRNN” classifier will be trained with the comparison of “BioBERT + TextRNN”. PubMedBERT and BioBERT are effective pre-trained models in the field of biomedical natural language processing, and "BERT+TextRNN" is proved to be an effective classification model [4].

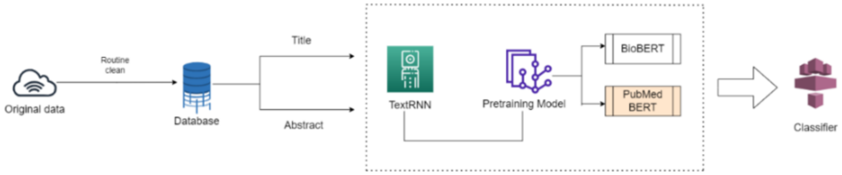

Figure 1. The study framework

3. Results

The testing results of the combined classification downstream model demonstrate that all indexes of “PubMedBERT + TextRNN” are consistently at a high level with P= 0.952014, R=0.936696, F1=0.931664.“BioBERT + TextRNN” are consistently at a high level with P= 0.854138, R=0.878695, F1=0.871694.

4. Discussion

There are several reasons for “PubMedBERT + TextRNN” obtaining optimal performance in cancer text classification: the cancer publications are quite suitable for TextRNN. PubMedBERT has already been successful trained on the PubMed. The titles and the abstracts would be taken as two independent layers to capture text features.

5. Conclusions

This research presents a scalable and extensible model that is suitable for high-resolution subject classification of the cancer literature at the publication level. Verification of the multi-label classifier for literature at the publication level indicates that it could provide effective support for academic statistics and clinical research.

References

[1] Daniels H. Exploring the Use of Genomic and Routinely Collected Data: Narrative Literature Review and Interview Study. J Med Internet Res .2021;23(9): e15739. DOI: 10.2196/15739.

[2] Jie Kong. Research on Automatic Literature Classification System Based on Deep Learning and Chinese Library Classification. New Century Library.2021;(05):51-56.

[3] Luo W. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view.2016;18(12): e323.

[4] Zhang Y, Li X, Liu Y, Li A, Yang X, Tang X.A Multilabel Text Classifier of Cancer Literature at the Publication Level: Methods Study of Medical Text Classification. JMIR Med Inform 2023;11:e44892.