Chapter 18
Creating a Sustainable Large-Scale Content-Based Biomedical Article Classifier Using BERT

Aakash Jayakumar
SRM Institute of Science and Technology, India

Kavya Saketharaman
SRM Institute of Science and Technology, India

J. Arthy
SRM Institute of Science and Technology, India

S. Jayabharathi
SRM Institute of Science and Technology, India

ABSTRACT

Given the scarcity of labeled corpora and the high costs of human annotation by qualified experts, clinical decision-making algorithms in biomedical text classification require a significant number of costly training texts. To reduce labeling expenses, it is common practice to use the active learning (AL) approach to reduce the volume of labeled documents required to produce the required performance. There are two methods for categorizing articles: article-level classification and journal-level classification. In this chapter, the authors present a hybrid strategy for training classifiers with article metadata such as title, abstract, and keywords annotated with the journal-level classification FoR (fields of research) using natural language processing (NLP) embedding techniques. These classifiers are then applied at the article level to analyze biomedical publications using PubMed metadata. The authors trained BERT classifiers with FoR codes and applied them to classify publications based on their available metadata.

INTRODUCTION

The vast corpus of biomedical literature accessible on PubMed poses a formidable challenge for researchers, healthcare providers, clinicians, and the general public when it comes to locating relevant informa-

DOI: 10.4018/979-8-3693-5951-8.ch018

tion (Ghozali et al., 2022a). A standard search on PubMed yields hundreds to thousands of documents, impeding physicians from promptly accessing pertinent data during patient care. Hence, the need arises for a literature repository that is not only intuitive but also well-organized, ensuring ease of comprehension to aid clinical decision-making (Awais et al., 2023). Research has underscored the importance of presenting large document collections in an easily digestible manner, underscoring the necessity for human-friendly access (Bhuva & Kumar, 2023).

Machine learning stands as the predominant methodology for predictive analysis and data classification, assisting individuals in making critical decisions. Machine learning algorithms undergo training through instances wherein they assess historical data and derive insights from past experiences (Boopathy, 2023). With repeated training on instances, these algorithms become proficient in recognizing patterns that enable future predictions. At the core of machine learning algorithms lies data, and further data generation is achievable through training on historical datasets (Elaiyaraja et al., 2023). Generative adversarial networks, an advanced concept in machine learning, have been utilized to create additional visual content by learning from previously generated images, extending their utility to text and speech synthesis (Ghozali, 2022). Consequently, machine learning has significantly broadened the scope of data science applications, incorporating computer science, mathematics, and statistics for data-driven inferences (Ghozali et al., 2022b).

The scientific literature landscape is rapidly expanding, with over a million paper citations added to PubMed in the past year alone (Tak et al., 2023). Effective techniques are imperative to automatically identify entities, link them to standardized concepts within knowledge bases, and index key subjects, simplifying information retrieval for readers (Ravi et al., 2023). The task of named-entity recognition (NER), entity linking, and topic indexing, focusing on chemical names and themes within full-text PubMed publications, has been incorporated into BioCreative VII (Kothuru, 2023). Named entity recognition (NER) is a pivotal phase in the information extraction process from text. The latest NER methodologies employ BERT-based models, with demonstrated performance enhancements achieved through pretraining BERT on domain-specific texts and employing domain-specific lexicons (Krishna Vaddy, 2023). In the realm of biomedical NLP, larger models tend to perform better as NER is particularly sensitive to alterations in the model vocabulary. Following NER, entity linking assumes critical importance, involving the mapping of natural language concepts to their unique identifiers and canonical forms preserved in knowledge bases (Kumar et al., 2023). Older entity linking methods rely on heuristics such as string matching and edit distance calculations (Senbagavalli & Arasu, 2016). At the same time, modern deep learning techniques encompass a multi-step pipeline integrating an NER model, candidate generation, candidate selection, and entity ranking (Kumar Nomula, 2023).

Text classification is a common machine learning approach employed to structure the vast expanse of unstructured digital data (Veronin, et al., 2020a). Algorithms like Support Vector Machine and Naïve Bayes are frequently used due to their simplicity and high accuracy. The advent of pre-trained language models, such as BERT, founded on deep neural networks, has brought about a revolution in natural language processing (Vashist et al., 2023). However, irrespective of the method employed, the demand for appropriately labelled training data remains constant. Manual annotation of training examples can be prohibitively resource-intensive, particularly in domains like biomedicine, necessitating the exploration of alternative strategies, including active learning, to reduce annotation efforts (Thallaj & Vashishtha, 2023).

Natural language processing (NLP) techniques empower computers to undertake an array of language and speech-related tasks, with biomedicine reaping substantial benefits from NLP applications (Veronin, et al., 2020b). NLP has been instrumental in diverse applications, encompassing the analysis