Multilabel-classification task for Medline abstracts

Nelson Quiñones [1,2,3], Cesar Canales [1], Javier Torres [1], Dietrich Rebholz-Schuhmann [3,4], Leyla Jael Castro [3], Andrés Aristizábal [1]

[1] *Universidad ICESI, CL 18 122-135, Cali, 760031, Colombia*
[2] *Leibniz University of Hannover, Welfengarten 1, Hannover, 30167, Germany*
[3] *ZB MED Information Centre for Life Sciences, Gleueler Str 60, Cologne, 50931, Germany*
[4] *University of Cologne, Albertus-Magnus-Platz, Cologne, 50923, Germany*

Abstract

Assigning categories to scholarly articles is a common approach to help researchers navigate the continuously growing scientific literature. This is a task well-covered in biomedical literature thanks to efforts assigning Medical Subject Headings to biomedical abstracts, particularly from Medline. Here we propose a multilabel-classification approach to assign major topics to biomedical literature with the purpose of later applying transfer learning to cover conference papers and preprints, as well as the agricultural domain. In this short paper, we present some preliminary results.

Keywords

Multilabel-classification, literature categorization, MeSH topics

1. Introduction

ZB MED Information Centre for Life Science hosts literature for the biomedical and agricultural domains. They are currently implementing a new topic-based recommender system. To this end, we are first exploring the literature in the biomedical domain by taking advantage of the Medical Subject Headings (MeSH) [1] descriptors assigned to Medline abstracts and available in the PubMed repository. MeSH is a comprehensive controlled vocabulary for the purpose of indexing literature in the biomedical domain. Here we present preliminary results from our initial experiments on assigning Unified Medical Language System (UMLS) Semantic Network Types (STY) [2] to Medline publications annotated with MeSH terms. With the lessons learned from this approach, we will move to transfer learning approaches to cover biomedical publications outside the Medline scope.

2. Materials and Methods

We worked with title, abstract and MeSH descriptors corresponding to a subset of the PubMed Central Open Access [3] articles retrieved with the Biopython library [4]. From the initial set of 7.4 million abstracts from 2015 to 2022, we retained only 2.8 million corresponding to those with all the elements, i.e., abstract, title, and MeSH descriptors, available in machine processable form. Data was further cleaned and transformed to create word embeddings. We then translated the MeSH terms to UMLS STYs to (i) reduce the number of prediction classes, from 348,860 in MeSH to 127 in UMLS STY, and to (ii) prioritize those types that could be more meaningful to biomedical researchers. The dataset creation process took 2 days in an AMD Ryzen 5 3400G. Our method corresponds to a fine-tuning of transformer models.

First, we initialized the models with pre-trained parameters, and then we fine-tuned such parameters by using labeled data from the downstream tasks. A new layer on top of the based model abstracts the knowledge enclosed by our dataset. Preliminary exploration was done on the HugginFace

Proceedings Semantic Web Applications and Tools for Healthcare and Life Sciences, February 13–16, 2023, Basel, Switzerland
EMAIL: ljgarcia@zbmed.de (A. 5)
ORCID: 0000-0002-1018-0370 (A.4); 0000-0003-3986-0510 (A.8)

platform. We then perform hyperparameter optimization with an algorithm called Hyperband [5] to find the best configuration to train the final model. We kept track of metrics including Hamming Score, Accuracy Score, macro F1, micro F1, and Hamming Loss. Still, our main metric to guide the hyperparameter optimization process was the F1 micro as our corpus exhibits a high imbalance in the classes. The Mobster algorithm (available in the Syne Tune Library) was used in over 10% of the articles in the corpus to find the values corresponding to the best hyperparameter configuration. We allowed the algorithm to run for four days in a virtual machine part of the deNBI Cloud platform, equipped with an RTX6000 GPU and 128GB of ram. The optimal configuration was used to train our model on our dataset. In addition, we created a proof-of-concept web application[1] to use the model and display predicted STYs for a given PubMed identifier. We used vanilla JS for the web application and uploaded the model to HuggingFace[2].

3. Results and Discussion

The hyperparameters used in the optimization process were the following: Learning rate (LR) between [5e-6 ~ 1e-4], dropout rate (DR) between [0 ~ 1], model selection [biobert-v1.1, distilbert-base-uncased, scibert_scivocab_uncased, Bio_ClinicalBERT, bert-base-uncased], the maximum length of input tokens (L) between [100 ~ 512], batch size between (BS) [4 ~ 64], and the number of threads (NT) used for processing between 1 and 8. The best-performing model had an LR of 2.0-05, a DR of 0.0, used the scibert_scivocab_uncased model, an L of 403, an BS of 23, and an NT of 5. After training the previously mentioned model with the training dataset, we obtained the following results with the validations dataset: an F1 micro of 0.489, an accuracy score of 0.196, an F1 macro of 0.416, Hamming score of 0.389, and Hamming Loss of 0.016. Although the metric scores do not show high values, our approach can still be further developed and improved. Multi-classification and multi-labeling with the number of classes, i.e., STY labels, that we are dealing with do not commonly show high scores as happens with binary classification. Still, pre-trained data opens new possibilities for this sort of task.

5. Acknowledgements

This work was partially supported by the BMBF-funded de.NBI Cloud within the German Network for Bioinformatics Infrastructure (de.NBI) (031A532B, 031A533A, 031A533B, 031A534A, 031A535A, 031A537A, 031A537B, 031A537C, 031A537D, 031A538A)

4. References

[1] Dhammi IK, Kumar S. Medical subject headings (MeSH) terms. Indian J Orthop. 2014 Sep;48(5):443-4. doi: 10.4103/0019-5413.139827. PMID: 25298548; PMCID: PMC4175855.

[2] National Library of Medicine (US); 2009 Sep-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK9676/

[3] PMC Open Access Subset [Internet]. Bethesda (MD): National Library of Medicine. 2003 - [cited 2022 11 20]. Available from https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/

[4] Cock, P.J.A. et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 2009 Jun 1; 25(11) 1422-3 https://doi.org/10.1093/bioinformatics/btp163 pmid:19304878

[5] Li, Lisha, Kevin Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, and Ameet Talwalkar. “Hyperband: A Novel Bandit-Based Approach to Hyperparameter Optimization.” arXiv, June 18, 2018. https://doi.org/10.48550/arXiv.1603.06560.

[1] https://github.com/zbmed-semtec/topic-categorization-system

[2] https://wandb.ai/javtor/huggingface and https://huggingface.co/datasets/Javtor/biomedical-topic-categorization