OJS Field of Study Classification using Transformers

Stanford CS224N Custom Project

Jon Ball
Graduate School of Education
Stanford University
jonball@stanford.edu

Abstract

Although Open Journal Systems (OJS) is the world's largest open source academic publishing software by user volume, little is currently known about the different fields of study supported by OJS. Researchers working with the Public Knowledge Project have discovered that a majority of the more than 30,000 journals actively publishing with OJS are based in middle income countries such as Indonesia, India, and Brazil. A majority of OJS journals are also published in English. And yet, these journals' *fields of study* are only just being revealed. Weber et al.'s (2020) field of study (fosc) classifier[1] is a large multilayer perceptron that has been successfully applied to a sample of English-language article abstracts representing more than 20,000 active OJS journals. With the hope of improving upon fosc's baseline classification the first truly global view of the different fields of study favored by OJS users this paper introduces a new disciplinary classifier, built using AllenAI's pre-trained SciBERT model[2] and Hugging Face's popular transformers library. Whereas Weber et al. implemented their neural net using tensorflow, a pytorch implementation using transformers shows significant promise when trained and evaluated on Weber et al.'s (2020) labeled data.[3] The transformers-based model discussed in this paper was trained on significantly fewer examples (n = 256,000 article abstracts) than Weber et al.'s for significantly less time, yet it achieves better precision[4] on a subset of Weber et al.'s evaluation set. However, it achieves worse recall scores. The classifier takes as its input short samples of scholarly text such as article abstracts, and it outputs a multi-label classification for each sample, represented as a one-hot vector of length 20 corresponding to the Australian and New Zealand Standard Research Classification system's 20 fields of study.[5] In the relatively niche research area of disciplinary tagging, this paper offers an easily scalable model for further training, as well as circumstantial evidence that transformers may be the optimal architecture for field of study classification tasks.

1 Key Information to include

- Mentor: Ben Newman
- External Collaborators (if you have any): Public Knowledge Project (PKP)
- Sharing project: N/A

[1] https://github.com/tgweber/fosc/
[2] https://github.com/allenai/scibert
[3] https://zenodo.org/record/3490460.YhlU6hPMI0Q
[4] Both micro- and macro-averaged precision were measured with scikit-learn's precision$_{s}coremethod.$
[5] https://www.abs.gov.au/ausstats/abs@.nsf/0/6BB427AB9696C225CA2574180004463E

2 Introduction

Because Open Journal Systems adheres closely to open source principles, user data have historically been inaccessible to researchers. However, a new ‘beacon’ feature introduced in the latest version of OJS has made it possible to collect OAI-PMH metadata[6] for the millions of articles published in OJS journals. Labeling a sample of these data would be costly, and so Weber et al.’s (2020) labeled data are invaluable for the downstream task of classifying OJS journals by their field of study. OJS presents perhaps the world’s largest open access use case for field of study classification models, meaning that models successfully applied to OJS beacon data can reveal previously undiscovered trends in academic publishing at an unprecedented scale. Weber et al.’s large multilayer perceptron is a first-of-its-kind neural model for disciplinary tagging, and the obvious choice of classifier for application to OJS beacon data. However, once adequately trained and scaled, the pre-trained models provided in Hugging Face’s transformers library can match Weber et al.’s model’s performance, with clear potential to set a new baseline in field of study classification.

3 Related Work

Weber et al.’s (2020) large multi-layer perceptron sets the current baseline for performance on field of study classification tasks.[7] Golub et al. (2018) previously used Support Vector Machine and Multinomial Naive Bayes algorithms to classify a smaller sample of research article abstracts according to the Dewey Decimal Classification system. However, as Weber et al. note, the Dewey Decimal Classification system specifies just 10 primary field of study labels, which is an inadequate level of specificity ("Science" is a single label in DDC, for example). This paper agrees with Weber et al. in its use of the Australian and New Zealand Standard Research Classification (ANZSRC) system,[8] which specifies the following 20 disciplinary classes.

4 ANZSRC Labels

- 1 Mathematical Sciences
- 2 Physical Sciences
- 3 Chemical Sciences
- 4 Earth and Environmental Sciences
- 5 Biological Sciences
- 6 Agricultural and Veterinary Sciences
- 7 Information and Computing Sciences
- 8 Engineering and Technology
- 9 Medical and Health Sciences
- 10 Built Environment and Design
- 11 Education
- 12 Economics
- 13 Commerce, Management, Tourism and Services
- 14 Studies in Human Society
- 15 Psychology and Cognitive Sciences
- 16 Law and Legal Studies
- 17 Studies in Creative Arts and Writing
- 18 Language, Communication and Culture
- 19 History and Archaeology
- 20 Philosophy and Religious Studies

[6] https://www.openarchives.org/pmh/

[7] https://direct.mit.edu/qss/article/1/2/525/96148/Using-supervised-learning-to-classify-metadata-of

[8] https://www.abs.gov.au/ausstats/abs@.nsf/0/6BB427AB9696C225CA2574180004463E

5 Approach

The model proposed in this paper makes simple use of Hugging Face's Trainer, AutoTokenizer, and AutoModelForSequenceClassification classes (which helpfully have a new 'problem type' flag for 'multil label classification'). The 'multi label classification' models pass Hugging Face's standard BERT outputs through an additional layer for learning the 20 dimensions specified by 'num labels'.

6 Experiments

This section describes the experiment conducted.

6.1 Data

The multi-label field of study classification data used for this paper are a subset of the data[9] described in Weber et al.'s (2020) paper "Using supervised learning to classify metadata of research data by field of study." Weber et al.'s dataset consists of train (n = 497,003) and test (n = 55,223) data. Each row in the data contains both a sample of text and a multi-label vector with 1s and 0s corresponding to booleans that indicate each of the 20 different fields of study. The task is correct prediction of the 1s and 0s in each label vector, corresponding to the correct field of study labels for each example of scientific text.

6.2 Evaluation method

In line with Weber et al. (2020), the transformers-based model was evaluated using macro- and micro-averaged precision and recall.

6.3 Experimental details

The transformers-based model was trained for 1 epoch on 256,000 examples at learning rate $2e^{-5}$ using a free Google Colab Nvidia Tesla GPU. The model was evaluated on an eval set of 32,000 examples. Training time was approximately 5 hours. Examples were truncated at 128 tokens.

6.4 Results

Results for the transformers-based model are reported in the following table:

Labels	Recall	Precision
All (Macro)	0.45	0.85
All (Micro)	0.64	0.87

As compared to Weber et al.'s large multilayer perceptron results:

Labels	Recall	Precision
All (Macro)	0.65	0.83
All (Micro)	0.81	0.85

7 Analysis

The transformers-based model's micro- and macro-averaged precision scores are comparable to Weber et al.'s. The relatively lower recall scores can be partially explained by the nature of the multi-label classification task. The model is reasonably precise in assigning field of study labels, but because there can be multiple labels assigned to each example, the model frequently 'misses' labels. This error can be potentially be addressed by training on Weber et al.'s full train dataset, or by adjusting the threshold for label assignment (conventionally set to 0.5). However, on a *single*-label, multi-class field of study classification task, the transformers model may already perform on par with Weber et al.'s large multilayer perceptron.

[9] https://zenodo.org/record/3490460.YhlU6hPMI0Q

8 Conclusion

Transformers-based sequence classification models, such as those provided pre-trained by Hugging Face, show significant potential to aid in the classification of unlabeled, open access research data. Open Journal Systems (OJS) beacon data present an excellent use case for field of study classifiers, notably Weber et al.'s (2020) fosc neural net, applied at the level of articles, journals, or both. But OJS data also present a challenge for NLP researchers seeking to improve on the baseline set by Weber et al.'s simple feedforward neural net in classifying research data according to traditional disciplinary schemas. Trained on *all* of Weber et al.'s field of study classification task data for a sufficient amount of time, the transformers-based model described in this paper will surely establish a new baseline. It is the intention of the author to continue training a SciBERT/BertForSequenceClassification model until it achieves better performance on Weber et al.'s test set, and then, to upload the model to Github and Hugging Face Hub for others to use freely.