OJS Field of Study Classification using Transformers

Jon Ball
Stanford Graduate School of Education
Research Mentors: Ben Newman

Stanford | GRADUATE SCHOOL OF EDUCATION

Introduction

Open Journal Systems (OJS): open-source software with more than 30,000 active open access journals

How can these journals' fields of study be automatically classified?

Key Findings

Transformers-based model matches precision benchmarks on Weber et al.'s (2020) field of study classification task with ***less*** data and ***less*** compute

Methods

PRETRAINED MODELS:
AllenAI's SciBERT
Hugging Face's BertModel
- ForSequenceClassification
- 'multi_label_classification'

Analyses

- With HF transformers, ***less is more***
- HF model's lower recall scores explained by less training data for multi-label classification task

TASK: given article abstracts and titles, correct prediction of ANZSRC's 20 field of study class labels → [0,0,1,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0]

Results

Transformers-based SciBERT model:

Average	Precision	Recall
Micro	0.87	0.45
Macro	0.85	0.64

Weber et al.'s vanilla neural model:

Average	Precision	Recall
Micro	0.85	0.65
Macro	0.83	0.81

Conclusions

1. Transformers-based models show promise for classifying unlabeled, open access research data.
2. OJS data present an excellent use case.
3. OJS data also present a challenge for researchers seeking to improve on the baseline set by Weber et al.
4. Trained on *all* of Weber et al.'s field of study classification task data for sufficient time, the transformers-based model will surely establish a new baseline.
5. The author intends to continue training until the SciBERT model achieves benchmark performance, and then, to upload the model to Github and Hugging Face Hub for others to use freely.

References

1. Tobias Weber, Dieter Kranzlmüller, Michael Fromm, Nelson Tavares de Souza; "Using supervised learning to classify metadata of research data by field of study." *Quantitative Science Studies* 2020; 1 (2): 525-550.

Acknowledgments

Thanks are due to John Willinsky and the Public Knowledge Project (PKP) // Open Journal Systems (OJS) team, Ben Newman, and the entire CS224N teaching team!