Contributions to a system for Open Reproducible Publication Research

Topic Classification of Publications

Dayoung Lim <D.Lim-2@student.tudelft.nl>

1. Introduction

Background

- Volume of published journals and difficulty in finding journals are increasing [1] → correctly classify publications

Research Gap

- Existing classification works with short text are mainly sentence based [2]
- Abstract based classifications are mainly domain specific [3]

Research Question: *How can publication topics be identified and matched based on existing journal topic values?*

2. Methodology

Data

- April 2022 Public Data File from Crossref: includes data of multilingual journals, including abstracts, titles, and topic values

↓

Data Selection

- Extract works that have an abstract, titles, and topic values
- Use of Alexandria3k: provides efficient querying on large publication open data sets [4]

↓

Data Cleaning

- After examining the data format, unnecessary aspects* for the model are removed (e.g., HTML tags, trailing whitespace)

* Considered the nature of the data set

↓

Data Embedding

- OpenAI Embeddings (text-embedding-ada-002)

↓

Model Training

- XGBoost
- GridSearchCV: hyper-parameter tuning

↓

Evaluation

- Weighted F1: Harmonic mean of precision and recall
- Micro-average precision: $\frac{Total\ TP}{Total\ TP + Total\ FP}$
- Micro-average recall: $\frac{Total\ TP}{Total\ TP + Total\ FN}$

3. Results

Performance - Initial Run

- 10,000 data and 50 topic values
- Grid search → max_depth=6, eta=0.5, n_estimators=500
    - max_depth: depth of the tree
    - eta: learning rate
    - n_estimators: number of boosting rounds or trees to build
- Baseline model (BM25 + XGBoost) comparison
    - Difference is in data cleaning stage
    - XGBoost parameters for both experiment are the same
- Different Features: Abstract, Abstract + Title, Abstract + Title + Author

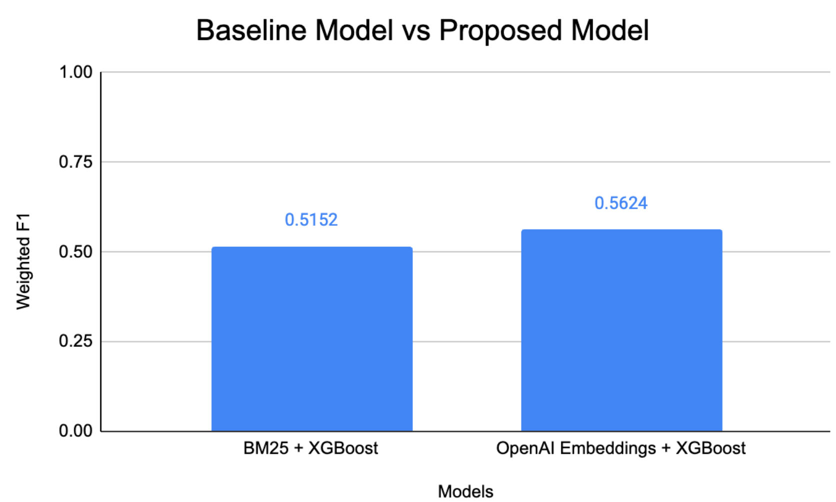

Figure 1. Comparison of weighted f1 for baseline and proposed model

Performance – Final Run

- 50,000 stratified data
- Abstract + Title
- Grid search → max_depth=20, eta=0.8, n_estimators=1000

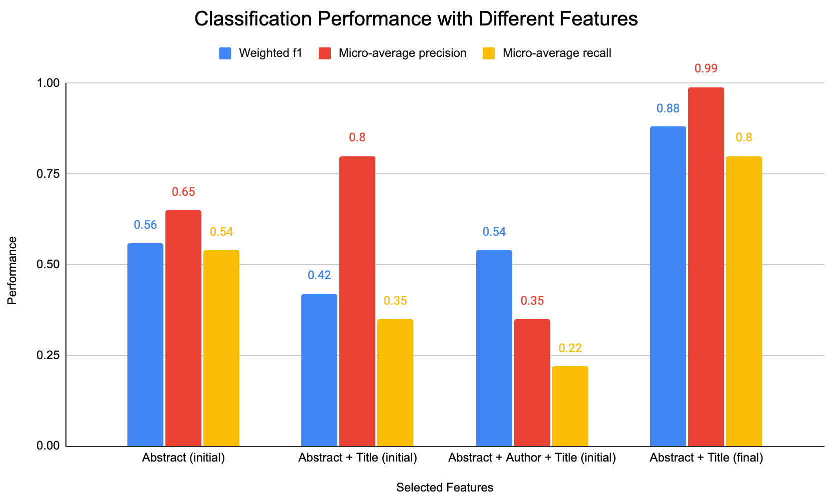

Figure 2. Performance comparison of initial runs with different features and the final run

Cost

- $0.0001/1k token
    - For 10,000 data (50 topic values): ~$0.2
    - For 50,000 stratified data~$2.8

4. Conclusion

Research Question Answer

- OpenAI Embeddings + XGBoost combination can be used for publication topic classification when the right features are chosen

Limitations

- High computational cost → only on a sample of Crossref
    - Classification verified on data with work names
- Correctness of original data has **not** been checked

Future works

- Test on works without work names
- Verify its performance on other publication data set
- Usage of newer model for embeddings: text-embedding-3-small/large [5]

References

[1] Ahmed Sajid, N., Ahmad, M., Rahman, A., et al. 2023. A novel metadata based multi-label document classification technique. *Computer Systems Science and Engineering 46*, 2, 2195–2214.
[2] Lee, K., Palsetia, D., Narayanan, R., Patwary, Md.M., Agrawal, A., and Choudhary, A. 2011. Twitter trending topic classification. *2011 IEEE 11th International Conference on Data Mining Workshops*.
[3] Deepika, A. and Radha, N. 2021. Performance analysis of abstract-based classification of medical journals using Machine Learning Techniques. *Computer Networks and Inventive Communication Technologies*, 613–626.
[4] Spinellis, D. alexandria3k: Local relational access to openly-available publication data sets. *GitHub*. https://github.com/dspinellis/alexandria3k.
[5] Zhuang, J., Braunstein, A., Neelakantan, A., et al. 2024. *New embedding models and API updates*. https://openai.com/blog/new-embedding-models-and-api-updates.

Responsible Professor: Diomidis Spinellis <D.Spinellis@tudelft.nl>
Supervisor: Georgios Gousios <G.Gousios@tudelft.nl>