Fail and try again: Return on topic modelling apply to archaeological scientific literature

Mathias Bellat, Ruhollah Tagizadeh-Mehrjardi, Thomas Scholten

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

FACULTY OF SCIENCES

CRC 1070 RESOURCECULTURES

Fail and try again: Return on topic modelling apply to archaeological scientific literature. CAA Conference, Auckland, 08-12 April 2024.

MATHIAS BELLAT[1,2,*], RUHOLLAH TAGHIZADEH-MEHRJARDI[1,2], THOMAS SCHOLTEN[1,2,3]

[1] CRC 1070 ResourceCultures, Eberhard Karls University of Tübingen, Germany
[2] Chair of Soil Science and Geomorphology, Department of Geosciences, Eberhard Karls University of Tübingen, Germany
[3] Cluster of Excellence Machine Learning „New Perpectives for Science“, Eberhard Karls Universität Tübingen, Germany

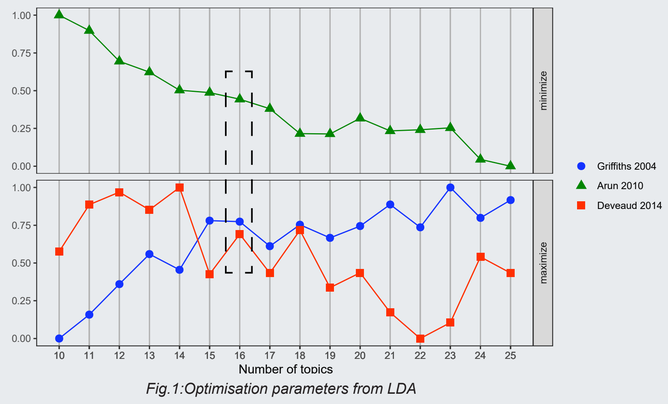

Fig.1:Optimisation parameters from LDA

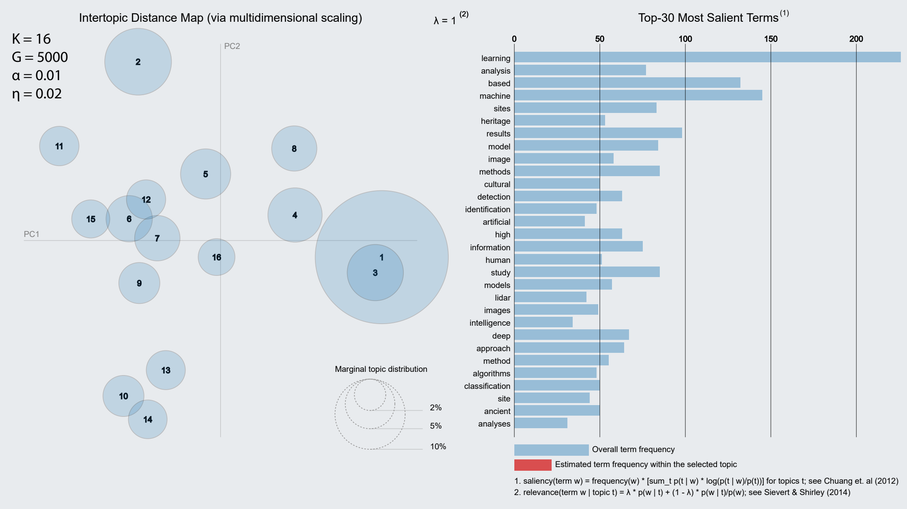

Fig.2: Inter-topic distance map wit LDAvis and Top 30 most relevant words.

*Fig.3: Clusters realised after the BERT model (**All-MiniLM-L6-v2**).*

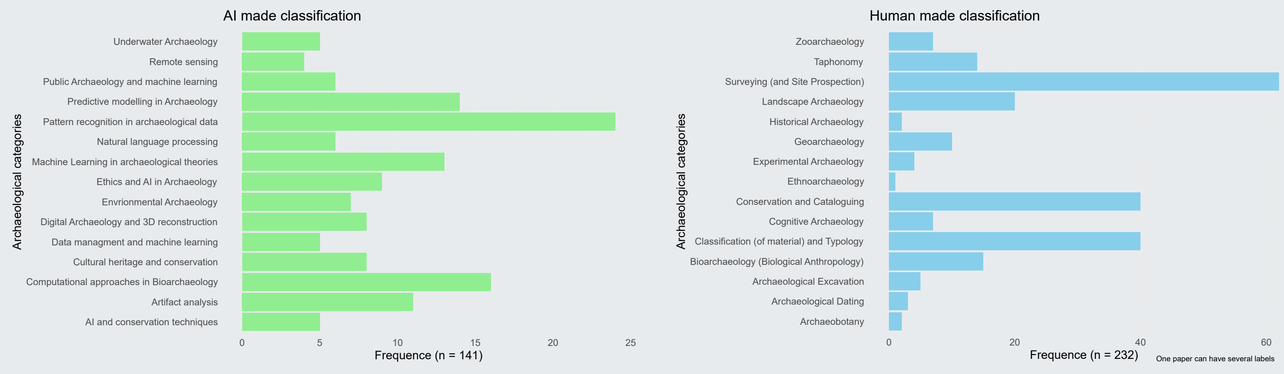

Fig.4: Comparaison between GPT 4.0. classification and human made classifcation.

1. INTRODUCTION

- For the review process, distinguishing different types of fields/subfiles inside the discipline helps understand the hidden patterns (Padarian et al., 2020).
- If expert-based classification is efficient and conducted according to naturalistic tradition, artificial intelligence can help in classifying.
- **Topic model** is part of **Natural language proccessing** which allows to uncovers the different topics of a *corpus* of documents thanks to statistical and machine learning approach (Brandsen, 2023).
- Two different models were tested for topic model and compared to **Large language model (LLM)** solution from GPT-4.0.

COULD TOPIC MODEL MODEL
HELP TO CLASSIFY ARCHAEOLOGICAL SUBDISCIPLINES**?**

2. MATERIEL AND METHODS

DATA

Scrapped documents	Filtered documents	Analysed documents
1460	731	141

METHODS

- **Latent Dirichlet allocation (LDA)**, an unsupervised model, was trained for several parameters **(Fig.1)**, and an optimal K = 16 was selected.
- With its triple-level architecture, **Bidirectional Encoder Representations from Transformers (BERT)** was adopted to cluster the different articles**.**
- **GPT-4.0** model from *OpenAI* was used to classify with supervised and unsupervised methods the documents.

3. RESULTS

- Running the topic model on full texts gave more sporadic topics than with the abstracts.
- LDA presented heterogeneous classes (**Fig.2**), but only 2 - 3 can be clearly identified.
- BERT model presented 3 clusters (**Fig.3**) with: bone surface marks; remote sensing, LiDAR, and automatic detection features; recognition and classification of different artefacts.
- The GPT-4.0 model presented an "ideal" topic selection of 12 labels and was able to create 15 classes to compare to the human made classification (**Fig.4**)

4. DISCUSSION AND CONCLUSION

- Topic modelling with "traditional" models presents the advance of being transparent and fully parametrisable, but their results in terms of topic classification are not always satisfactory.
- LLMs, such as GPT 4.0, are more powerful and flexible regarding human requests.
- However, LLMs can be seen as a "black box" where the process and optimisation can not be analysed in detail.
- The multi-labelling task is very demanding in terms of resources, even for LLMs, and can not always be performed (**Fig.5**).

Fig.5: Error message of GPT 4.0. while trying multi-labeling.

ACKNOWLEDGMENTS

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), Collaborative Research Centre *Resource-Cultures* - SFB 1070/3 - Project number 215859406.

CONTACT

*Mathias Bellat
mathias.bellat@uni-tuebingen.de

FURTHER READINGS

J. Padarian, B. Minasny, A.B. McBratney, (2020), "Machine learning and soil sciences: a review aided by machine learning tools", *SOIL* , 6(1). **A. Brandsen, (2023),** "Information Extraction and Machine Learning for Archaeological Texts. In C. Gonzalez-Perez, P. Martin-Rodilla, & M. Pereira-Fariña (Eds.), *Discourse and Argumentation in Archaeology: Conceptual and Computational Approaches*, Springer International Publishing.

CODE

Deutsche Forschungsgemeinschaft

Soil Science and Geomorphology
University ofTübingen

ACROSS THE HORIZON CAA Auckland 8 - 12 April 2024