Opportunities and Challenges in Classifying Hungarian Scientific Texts by Field of Science

Réka Dodé
ELTE Research Centre For Linguistics
Budapest, Hungary
dode.reka@nytud.elte.hu

Kristóf Varga
ELTE Research Centre For Linguistics
Budapest, Hungary
varga.kristof@nytud.elte.hu

Győző Zijian Yang
ELTE Research Centre For Linguistics
Budapest, Hungary
yang.zijian.gyozo@nytud.elte.hu

Gábor Madarász
Telekom System Integration Ltd.
Budapest, Hungary

Mátyás Osváth
ELTE Research Centre For Linguistics
Budapest, Hungary
osvath.matyas@nytud.elte.hu

Enikő Héja
ELTE Research Centre For Linguistics
Budapest, Hungary
heja.eniko@nytud.elte.hu

1 Introduction

The organization of knowledge is a fundamental necessity in the scientific world. Researchers and their work are classified into one or more scientific classification systems to help structure and organize research. These systems facilitate the identification, tracking, and comparison of scientific contributions, providing a framework for understanding the vast landscape of academic knowledge.

Traditionally, classification has been performed manually by humans. However, in order to support librarians' work, the need for automatic classification also emerged in this case some time ago.

2 Motivation and Challenges

Science is constantly evolving. New fields emerge, interdisciplinary research is expanding, and multidisciplinarity presents challenges to classification. Researchers often work at the intersection of multiple disciplines, making it difficult to categorize their work within a single predefined category.

The variety of classification systems highlights both the importance of this task and the challenges involved, as the categorization of scientific disciplines is often complex and ambiguous.

This challenge is further intensified by the increasing prevalence of multidisciplinary approaches in contemporary science.

2.1 Classification systems

The various repositories are curated by librarians. They classify incoming scientific publications using a scientific classification system they have selected in advance.

The REAL Repository of the Library and Information Centre of the Hungarian Academy of Sciences[1], for example, uses the Library of Congress Classification system (Library of Congress, n.d.).

The MTMT (Hungarian Scientific Works Database) and the Contenta repository in Szeged[2] use the Frascati classification system (OECD, 2007). In Hungary, the MTA's discipline nomenclature is also widely applied (MTA Magyar Tudományos Akadémia, 2017).

In addition to these, the Hungarian Academy of Sciences organizes the sciences into 11 sections. These sections are intended to cover the various scientific fields (MTA Magyar Tudományos Akadémia, n.d.).

[1] https://real.mtak.hu/

[2] https://www.ek.szte.hu/kezdooldal/mit-keres/contenta-repozitoriumok/

2.1.1 Frascati classification system

One well-known example is the Frascati classification system (OECD, 2007), which organizes scientific research into multiple hierarchical levels: from the 1st level with 6 main categories until the 5th level with 1,626 subcategories. The 2nd level contains 45 subcategories. The 6 main categories on the first level are: Natural Sciences, Engineering and Technological Sciences, Medical and Health Sciences, Agricultural Sciences, Social Sciences and Humanities.

2.1.2 The Scientific Sections of the (Hungarian) Academy

In practice, the role of the academic sections is to assign elected academicians to specific sections. The descriptions of these sections also reveal overlaps and ambiguities, as well as unclassified fields of science—such as Media and Communication.

The 11 sections are as follows: I. Section of Linguistics and Literary Scholarship, II. Section of Philosophy and Historical Sciences, III. Section of Mathematics, IV. Section of Agricultural Sciences, V. Section of Medical Sciences, VI. Section of Engineering Sciences, VII. Section of Chemical Sciences, VIII. Section of Biological Sciences, IX. Section of Economics and Law including sociology, demography, and political sciences, X. Section of Earth Sciences, XI. Section of Physical Sciences.

2.2 Challenges of the research

We are faced with several challenges. One of them is the already mentioned issue of categorization.

1. The complexity and multidisciplinarity of scientific fields justify the existence of multiple classification systems.

Although researchers are generally aware of which fields their work pertains to, they do not always find the appropriate label—especially in situations such as applying for grants, where they are required to indicate one (or preferably more) relevant scientific fields.

Naturally, librarians are also not expected to categorize with complete certainty. As a consequence, inconsistencies will inevitably appear in the training data as well.

2. All scientific fields are important and each must have its place within the scientific classification system. However, some fields include significantly more subfields, while others are narrower in scope and have fewer representatives. Despite this, in the classification system, they are treated on the same hierarchical level as broader fields with more researchers. We encountered this issue while compiling the training dataset, particularly in terms of lack of representativeness and imbalanced categories.

3. Although the existence of multiple scientific classification systems is justified, there are situations where a need arises to align or map these systems to one another. For instance, the Frascati classification is science-domain oriented, whereas the Library of Congress classification follows a more library-centric approach, where similarity may be based on temporal or geographical proximity. Regardless of which classification system is used, it is important to map categories across systems whenever possible to ensure compatibility and interoperability.

3 Methods and Data

For this study, we used an open-access dataset from the Szeged Contenta repository, specifically from the Doktori, Acta, and Publicatio collections.

These files, originally in PDF format, included Frascati classification labels in their metadata—assigned by librarians—which make the texts suitable for analyzing scientific classification.

Since the documents were in PDF format, we needed to extract the text for further processing. For this, we used Tesseract OCR engine 5.5 to perform optical character recognition (OCR)[3].

Since the data set was labeled using Frascati categories, this served as the starting point for the preparation of the training data. The labels were normalized and converted to the second level of the Frascati classification; in cases where a lower-level label was provided, the data was aggregated accordingly. We then examined the distribution of the 45 second-level Frascati categories within the dataset—that is, we analyzed how many texts were associated with each label.

[3] https://github.com/tesseract-ocr

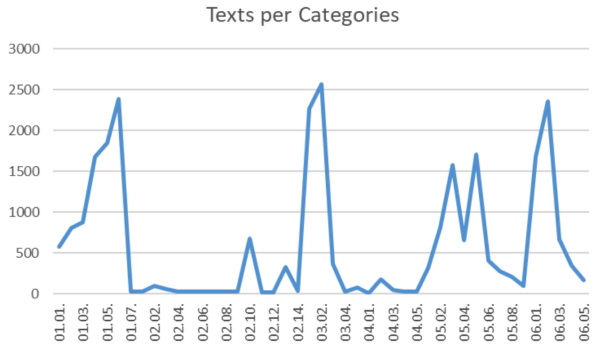

Figure 1: Texts per categories

During the processing, we encountered the issue that nearly half of the 45 Frascati categories were underrepresented (less than 20 texts) (see Figure 1). To address the imbalance, we implemented several steps:

1. The first and most important step was to reduce the 45 categories to the 11 sections of the Hungarian Academy of Sciences. The goal was to match as many second-level Frascati categories as possible with the appropriate section. We relabeled the publications with the new categories.

2. For sections with a very high number of publications, we removed the largest ones (i.e., those with the most texts).

3. We excluded non-Hungarian texts. We used Compact Language Detector v3 (CLD3)[4] for language detection.

4. In some of the experiments, a maximum of 200 texts per category was sufficient; however, in the other part of the research, a larger amount of data was important, so the underrepresented categories were oversampled.

3.1 Mapping between the Frascati scientific fields and the scientific sections of the Hungarian Academy of Sciences

The mapping was an important step to ensure that each scientific field would be represented by a sufficient quantity and quality of text, thereby generating high-quality training data for training and fine-tuning. Although the 11 sections do not represent a large number, they are twice as many as the first-level Frascati categories, thus providing a significantly more detailed and effective classification scheme.

The starting point was the 45 second-level Frascati categories. From these 45, we ultimately excluded the categories of Space, Other Humanities, Other Natural Sciences, and Other Social Sciences from being assigned to any subclass, as we judged these to be such mixed and ambiguous categories that they would degrade the quality of the training data.

Additionally, in one case, we had to go down to the third level: in the case of 06.04. Arts (arts, art history, performing arts, music), arts were assigned to the I. Section of Linguistics and Literary Scholarship, while art history was assigned to the II. Section of Philosophy and Historical Sciences.

For technical categories, theoretical affiliation was the main principle. For example, 02.14. Environmental Engineering Sciences was assigned to the X. Section of Earth Sciences rather than the VI. Section of Engineering Sciences.

4 Experiments

During the classification task, we had to make some decisions and experiment with multiple approaches. The first factor concerned the complexity of scientific domains—specifically, whether the classification

[4]https://github.com/google/cld3

should be single-label or multi-label. We decided to make single-label classification.

The next question was what kind of model to use for training: a large language model (LLM) or a BERT-based model, which tends to perform more efficiently on tasks of this nature.

We decided to experiment with fine-tuning a generative language model for classification. The result of this serves as our baseline. Additionally, we are also experimenting with ModernBERT (**aarsen2024modernbert**), an encoder-only architecture that performs efficiently on classification tasks.

4.1 Classification with LLMs

Llama 3.3 is a series of multilingual large language models (LLMs) developed by Meta AI, pushing the boundaries of language understanding and generation (Meta AI, 2024). It is a text only instruct-tuned model in 70B size (text in/text out). The Llama 3.3 model can be fine-tuned for various natural language processing tasks and datasets and it is open-source. Large language models (LLMs) have billions of parameters, making full fine-tuning computationally expensive and data-heavy. LoRA (Low-Rank Adaptation) (Hu et al., 2021) offers a more efficient alternative by freezing the original weights and adding a small set of trainable adapters that adjust the model for specific tasks. This reduces computation and data requirements while maintaining strong performance—like customizing a powerful tool instead of building one from scratch.

As a baseline, we are training the Llama 3.3 model using Low-Rank Adaptation.

4.2 Classification with ModernBERT

Various BERT models have been developed from the original encoder-only architecture, each introducing improvements in efficiency, scalability, or domain adaptation. For our research, we chose the ModernBERT model due to its proven downstream performance. Unlike older BERT variants, ModernBERT incorporates recent advancements in transformer optimisation and pre-training strategies, which enable it to capture deeper semantic relationships without significantly increasing computational overhead (Aarsen et al., 2024).

The ModernBERT model will be trained specifically on Hungarian texts to ensure optimal performance in the target linguistic context. Since the intended use case involves classifying longer scientific documents, the model's ability to handle extended input sequences and preserve contextual meaning across multiple sentences is a key advantage. This adaptation allows the model to generalize better across diverse domains within the corpus and maintain high classification accuracy even in cases where relevant information is spread across the document. To continue training it, the largest possible amount of high-quality training data is needed.

Our hypothesis is that ModernBERT will perform better in the task of scientific field classification than the generative language model. Our results will be useful for automatic scientific field classification in library work.

References

Aarsen, T., Cooper, N., Adams, G., Howard, J., Poli, I., et al. (2024, December). Finally, a replacement for bert: Introducing modernbert [Accessed: 2025-07-23].

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., & Li, H. (2021). Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*. https://arxiv.org/abs/2106.09685.

Library of Congress. (n.d.). Library of congress classification outline [Accessed: 2025-07-22].

Meta AI. (2024, December). Llama 3.3 70b instruct [Instruction-tuned multilingual LLM; release date: 2024-12-06].

MTA Magyar Tudományos Akadémia. (n.d.). Tudományos osztályok [Accessed: 2025-07-22].

MTA Magyar Tudományos Akadémia. (2017, October). Tudományági nómenklatúra [Accessed: 2025-07-22].

OECD. (2007, February). Revised field of science and technology (fos) classification in the frascati manual [Published: 2007-02-26].