# LLMs4Subjects: Automated Subject Tagging Using Large Language Models

Afaq Alam
21i-1700
i211700@nu.edu.pk

Shizra Burney
21i-2660
i212660@nu.edu.pk

Hammad Sikandar
21i-1684
i211684@nu.edu.pk

## Abstract

The exponential growth of digital scientific literature necessitates advanced automated subject classification methods. This research introduces an innovative approach to subject tagging using large language models (LLMs) for technical records from the Leibniz University's Technical Library (TIBKAT). By leveraging the T5 transformer model and the Gemeinsame Normdatei (GND) taxonomy, we develop a bilingual semantic processing system capable of accurately recommending subject tags for technical documents in both English and German.

***Categories and Subject Descriptors*** I.2.7 [*Artificial Intelligence* ]: Natural Language Processing - Text Analysis

***General Terms*** Machine Learning, Information Retrieval, Computational Semantics

***Keywords*** Large Language Models, Subject Tagging, , Digital Libraries

## 1. Problem statement

Digital libraries face significant challenges in efficiently organizing and classifying technical documents. One major issue is that manual subject tagging is both labor-intensive and time-consuming. Moreover, existing classification systems often struggle to handle the complexities of interdisciplinary research effectively. The primary research problem, therefore, revolves around developing an automated and semantically aware subject tagging system. Such a system should not only be capable of accurately classifying documents using a comprehensive taxonomy but also significantly reduce the need for human intervention in the document classification process.

## 2. Introduction

**Problem Details** Modern scientific literature encompasses diverse and complex research topics that defy traditional classification methods. The Leibniz Information Centre for Science and Technology (TIB) maintains over 100,000 records spanning multiple disciplines, highlighting the urgent need for advanced classification techniques.

**Motivation**

The proposed solution aims to address critical challenges in digital library management by improving document discoverability, enhancing the efficiency of information retrieval, and supporting multilingual semantic processing.

**Background**

- **TIBKAT**: Open-access bibliographic database for science and technology
- **GND Taxonomy**: Comprehensive authority file for information categorization
- **Large Language Models**: Advanced AI systems capable of nuanced semantic understanding

## 3. Related work

Subject classification techniques encompass a range of methods, including traditional rule-based classification systems, machine learning approaches to document categorization, and neural network-based semantic classification methods. Recent developments in the field highlight initiatives such as the Annif project in computational subject tagging, advances in transfer learning for multilingual processing, and emerging applications of large language models in information organization.

## 4. Your approach

Methodology

1. **Data Preprocessing**

The process involves extracting abstracts from the TIBKAT collection, normalizing and cleaning the technical documents, and preparing input-output pairs for model training.

2. **Model Architecture**

The base model for the system is the T5-small transformer, which is fine-tuned using a transfer learning strategy and incorporates bilingual processing capabilities.

3. **Training Process**

The input format for the model is structured as "tagging: abstract," while the output consists of comma-separated GND subject tags. This approach leverages pre-trained language understanding to enhance tagging accuracy.

**Key Innovations**

The system incorporates bilingual semantic processing to provide automated subject tag recommendations, requiring minimal manual intervention.



*CONF 'yy*, Month d–d, 20yy, City, ST, Country.

http://dx.doi.org/10.1145/nnnnnnn.nnnnnnn

## 5. Evaluation and Experiments

Experimental Setup

- **Dataset**: TIBKAT technical records
- **Languages**: English and German
- **Taxonomy**: GND subject headings

The benchmarks for the system include evaluating the precision of subject tag predictions and assessing the semantic relevance of the assigned tags. Performance metrics focus on semantic matching accuracy and computational efficiency.

## 6. Sidenotes

### 6.1 Transformer Model Working:

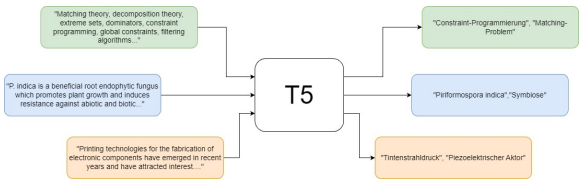


**Figure 1.** T-5 Transformer Model

### 6.2 Training Loss:

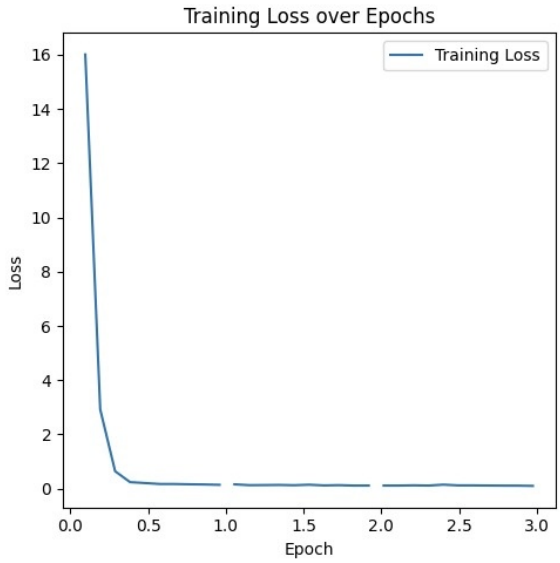


**Figure 2.** Training Loss over Epochs

| Epochs | Batch Size | Input Token Size | Label Token Size | Training Loss | Validation Loss | Time Taken (minutes) |
|---|---|---|---|---|---|---|
| 20 | 8 | 512 | 64 | 0.31 | 0.26 | 120 |
| 25 | 8 | 1028 | 256 | 0.30 | 0.26 | 48 |
| 3 | 4 | 2056 | 512 | 0.10 | 0.09 | 31 |

**Table 1.** Hyperparameter tuning results

## 7. Hyperparameter Tuning

The following table summarizes the hyperparameters used during the training process:

## 8. Testing:

**Abstract:** With the growing reliance on renewable energy sources such as wind and solar, power grid operators face unprecedented challenges in ensuring stability and reliability. This paper explores advanced machine learning techniques, including reinforcement learning and neural networks, to optimize energy distribution in smart grids. The study integrates historical weather data, energy consumption patterns, and grid topology to predict demand surges and adjust supply in real-time. Additionally, it examines the use of blockchain for decentralized energy trading, ensuring transparency and security. The findings highlight significant improvements in grid efficiency and reduction in carbon emissions, paving the way for sustainable energy management systems. **Predicted Subjects:** Grids, Stromverbrauch, Stromversorgung

## References

[1] Raffel, C., et al. "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer." Journal of Machine Learning Research, 2020.

[2] TIB Open Data Services Documentation

[3] GND Taxonomy Official Documentation

[4] Hugging Face Transformers Library

[5] https://github.com/jd-coderepos/llms4subjects/tree/main/shared-task-datasets

[6] https://drive.google.com/file/d/1wu6XXl604nFErOxp5zoy-tVi2tHtA$_u m/view?usp = sharing$