Jim at SemEval-2025 Task 5: Multilingual BERT Ensemble

Jim Hahn [1,2]
[1]University of Illinois at Urbana-Champaign, School of Information Sciences, USA
[2]University of Pennsylvania, University Libraries, USA
jimhahn@illinois.edu

Abstract

The SemEval-2025 Task 5 calls for the utilization of LLM capabilities to apply controlled subject labels to record descriptions in the multilingual library collection of the German National Library of Science and Technology. The multilingual BERT ensemble system described herein produces subject labels for various record types, including articles, books, conference papers, reports, and theses. For English language article records, bidirectional encoder-only LLMs demonstrate high recall in automated subject assignment.

1 Introduction

SemEval-2025 Task 5 utilizes Large Language Model (LLM) capabilities to assign controlled subject labels to multilingual German and English record descriptions (D'Souza et al., 2025). BERT models, such as ModernBERT, are a natural fit for subject tagging (Warner et al., 2024). When trained as classifiers, such as in this submission, they are less prone to hallucination—a common challenge in generative AI models. The ModernBERT model provides improvements of increased context sequences of 8192 tokens, over prior limits of 512 tokens in the original BERT (Warner et al., 2024). Another advance in ModernBERT is the use of flash attention (Dao et al., 2022).

The emergent capabilities of LLMs are not fully explained by existing theories (Li et al., 2022). BERT utilizes a transformer architecture, but does not stack together transformers as in GPT models (Devlin et al., 2019). The research community has been able to empirically inspect why BERT works so effectively (Tenney et al., 2019; Rogers et al., 2020). Similar "mechanistic interpretability" is underway for LLMs, but is not nearly as mature as the understanding of BERT models (Sharkey et al., 2025).

The BERT ensemble developed for this task consists of four models: two multilingual BERT models, one German-only BERT model, and one English-only BERT model. All models were fine-tuned with data from the TIB Technical Library's Open-Access Catalog. See Table 1 for the model card links.

For the average recall measures in the quantitative leaderboard, the BERT ensemble ranked 7th out of 11 teams in the "All Subjects" task group. In the qualitative results, this system's highest ranking was 5th out of 13 teams. The BERT models do not mimic reasoning and cannot correct labels in the way current state-of-the-art reasoning models can, which puts purely BERT ensembles at a disadvantage. Future work will investigate combining BERT outputs with reasoning over the labels using advances in chain of thought (CoT). The code for training, testing, and inference is available on GitHub (`https://github.com/jimfhahn/SemEval-2025-Task5`).

Returning to the call to research and develop a system that could be used in practice, the system is fully reusable from the Hugging Face platform (`https://huggingface.co/`). Noteworthy for its open source hosting, the Hugging Face platform enables hosting of models and datasets and has useful inference capabilities for machine learning projects.

2 Background

The task to assign a subject to a work requires a target vocabulary. In this case, the GND vocabulary is paired with the title and abstract data from the TIBKAT collection (D'Souza et al., 2024). The language of the title and abstract was both English and German. To train the BERT models that encompassed the ensemble powering the core of the inference stack all provided data from TIBKAT are processed into JSONL format and were modeled using title and abstract text along with corresponding

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 2407–2412

Model Name	URL
German BERT	https://huggingface.co/jimfhahn/bert-german-cased
Multilingual cased	https://huggingface.co/jimfhahn/bert-multilingual-cased
Multilingual uncased	https://huggingface.co/jimfhahn/bert-multilingual-uncased
ModernBERT base	https://huggingface.co/jimfhahn/ModernBERT-base-gnd

Table 1: The BERT ensemble is comprised of four GND-trained models developed for this task.

DNB labels. The curated dataset is available on Hugging Face with an open source license (https://huggingface.co/datasets/jimfhahn/SemEval2025-Task5-Curated-Data).

While LLMs excel at a wide range of generative AI tasks, the specific task at hand is generating subjects, which falls under classification. Therefore, BERT models are well-suited to act as the core inference engine powering an LLM-based subject indexing system.

3 System overview

The Hugging Face software package "AutoTrain Advanced" was configured for training the component BERT models (Thakur, 2024). The input training data, sourced from the "All Subjects" folder provided by the competition organizers, was incorporated into the dataset. Additionally, the supplementary dataset, "DNB SKOS Exports of the GND," was subsequently incorporated to enrich the input data. A roughly 25/75 split was applied, allocating 78,800 rows to testing and 245,000 rows to training. This decision reflects the GND technical staff's acknowledged expertise in curating high-quality resources. A departure point for prior work in semi-automated subject indexing is to reference existing professional skills while extending professional expertise (Hahn, 2021, 2024).

The combined dataset is multilingual mixing in both German language training with English language text. For the training, the software was installed in a compute environment at the University of Illinois campus compute cluster where GPU hardware, NVIDIA A100 Tensor Core GPUs are available to be scheduled. Training time for the largest BERT models, including ModernBERT, was completed within ten hours.

The system employs an ensemble of BERT models to generate classification results. Refer to the Inference folder of the GitHub repository for the methods described herein (https://github.com/jimfhahn/SemEval-2025-Task5/blob/main/Inference/inference.py).

During inference, the `classify_text_batch` function processes input texts in batches. Each input is tokenized with truncation applied, followed by generating probability scores for all possible labels using the `torch.softmax` function.

The system then identifies the top n labels (default is 50) and their associated confidence scores using `torch.topk`. While the models are trained for single-label classification, this approach enables the generation of multiple subject labels for each input. To aggregate classification results from individual models in the ensemble, the `filter_and_aggregate` function combines confidence scores for each label across models, summing them to produce a single combined score. The system then retains the top 50 labels based on the highest accumulated scores. The `get_top_50_subjects` function finalizes the process by extracting and validating these top 50 labels for each input. By leveraging the probabilistic confidence scores, this pipeline adapts single-label models to a multi-label context, effectively simulating a multi-label classification system through confidence score aggregation.

4 Experimental Setup

The training of BERT models began with loading and processing the dataset from Hugging Face. Refer to the Train folder of the GitHub repository for the methods described herein (https://github.com/jimfhahn/SemEval-2025-Task5/blob/main/Train/train.py). The processing code filtered out underrepresented labels, ensuring that each label had at least two examples. Subsequently, the code split the dataset into training and validation sets for BERT, ensuring that all classes were represented in both sets. The AutoTrain Advanced software package included a default configuration to stop training if there was no improvement after 5 epochs, to prevent overfitting. The threshold for measuring the new optimum to continue training was set to 0.01 by default. It ensured that the training

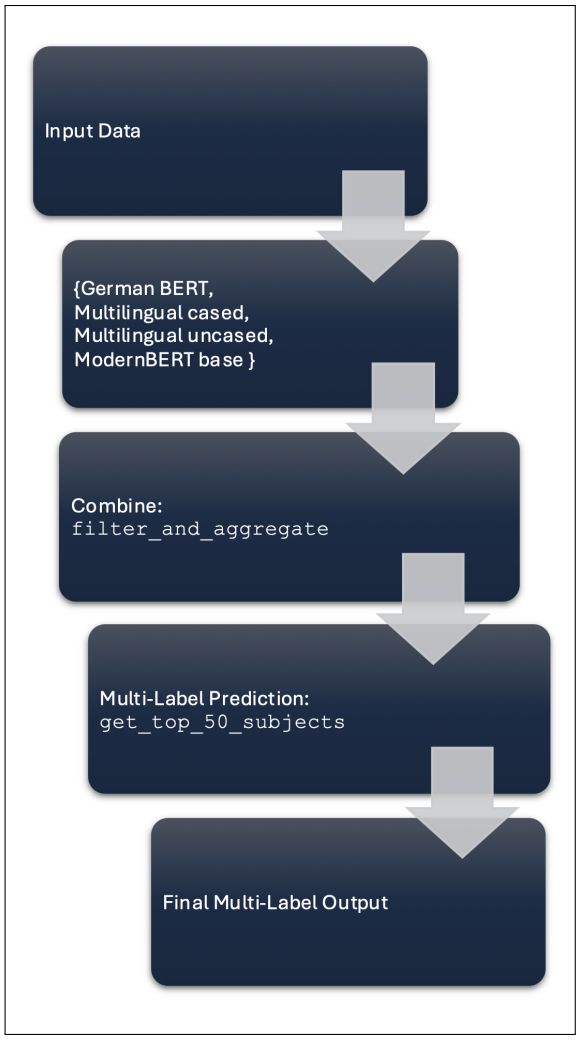

Figure 1: The Inference Pipeline.

```python
from autotrain.params import TextClassificationParams
from autotrain.project import AutoTrainProject

# Define the parameters for the AutoTrain project
params = TextClassificationParams(
    model="google-bert/bert-base-multilingual-cased",
    data_path="./jsonl",
    text_column="text",
    train_split="train",
    valid_split="validation",
    target_column="label",
    epochs=20, # drops out between epochs 10 through 20
    batch_size=16,
    max_seq_length=512,
    lr=3e-5,
    optimizer="adamw_torch",
    scheduler="linear",
    gradient_accumulation=2,
    mixed_precision="fp16",
    project_name="base-multilingual-gnd-bert",
    log="tensorboard",
    push_to_hub=True,
    )
```

Figure 2: The Settings for AutoTrain Advanced.

continued as long as the model's performance improved by at least 1% in each iteration. Figure 2 shows the settings of AutoTrain Advanced, as `TextClassificationParams` that were utilized to train each of the models on the curated dataset.

The AutoTrain Advanced software used Optuna for automated hyperparameter optimization (Akiba et al., 2019). In practice, two iterations of fixed learning rate settings (`lr=1e-5` and later `lr=3e-5`) were tested, with the latter yielding superior F1 scores during inference. Similarly, the trial and error included two batch size iterations in training. The initial tests used a batch size of 8, while the final, better-scoring model training parameters were trained using a batch size of 16. The BERT models were all trained as single-label classifiers where each input was assigned exactly one label.

5 Results

Recall@K was used as the central measure. Specifically, the average of Recall@K scores was used for the final leaderboard ranking of "Average Recall." According to the quantitative leaderboard, the Multilingual BERT ensemble ranked 7th out of 11 systems in the "All Subjects" category. See Table 2 for a selected set of metrics (K@50).

The system's performance on English language articles is noteworthy, as it was a standout in subject recall; the details are considered in section 5.1. Regarding the qualitative results, the system ranked 5th out of 13 systems in both Case 1 in Case 2. Detailed qualitative results are discussed in more detail in section 5.3.

5.1 Quantitative analysis

The "All Subjects" leaderboard was analyzed by record and by language. This analysis helps to identify where the BERT ensemble inference was most successful and where it was failing.

Record Type	Language	Recall
Article	de	0.2000
Article	en	**0.8329**
Book	de	0.5440
Book	en	0.5419
Conference	de	0.5165
Conference	en	0.5829
Report	de	0.5625
Report	en	0.4719
Thesis	de	0.4082
Thesis	en	0.3830

Table 2: K@50 by Record Type, Language, and Recall.

The system's standout performance was with English language articles. In the K@50 round, the system's recall for English language articles was 0.8329 of relevant subjects. Several teams in the competition had strong recall for this record type.

The BERT ensemble score of 0.8329 ranked sixth out of eleven scores for English language articles.

5.2 Why do English language article records have high recall?

An analysis was conducted on the readability of titles and abstracts in the English article record type, compared to other English language title and abstract records (Chall and Dale, 1995). Aggregated results, shown in Table 3, indicate that the title and abstract metadata from article records in the English training data is the least complex and most readable text among the record types, as evidenced by the lower Dale-Chall readability scores which indicate easier understanding. German language data was not evaluated for readability because the metric uses English language words.

Record Type	Average Readability	Record Count
Article	10.7815	1042
Book	11.5618	26966
Conference	12.5864	3619
Report	13.9757	1275
Thesis	12.6136	3452

Table 3: Average Dale-Chall Readability Scores and Record Counts by Record Type in the English Training Data.

Record Type	Average Readability	Record Count
Article	10.8531	423
Book	11.6002	7598
Conference	12.3372	808
Report	13.5789	334
Thesis	12.7133	833

Table 4: Average Dale-Chall Readability Scores and Record Counts by Record Type in the English Test Data.

In both English and German, the BERT ensemble struggled the most with thesis record types. However, the readability of the training data does not seem to fully explain the difficulties with thesis records. An analysis of subject groupings per record type in the training data was instructive. Figure 3 shows the distribution of subject counts by record type in the English training data. Notably, there are outlier points beyond the outer limits of the plot, suggesting greater variability or the presence of extreme values in that record type.

Two indicators for why English language articles scored among the highest recall in the task are considered here. First, the training examples for English language articles had a more consistent number of subjects per record. In contrast, there was greater variability for Thesis and Book record types in English, which had higher subject counts. Two additional box plot figures highlight the nuanced scores of reading complexity in the training data (Figure 4) and the reading complexity of the title and abstract test data (Figure 5).

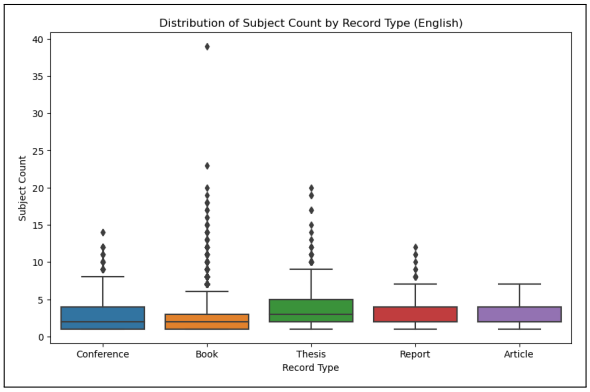

Figure 3: Subject Distribution in Training Data.

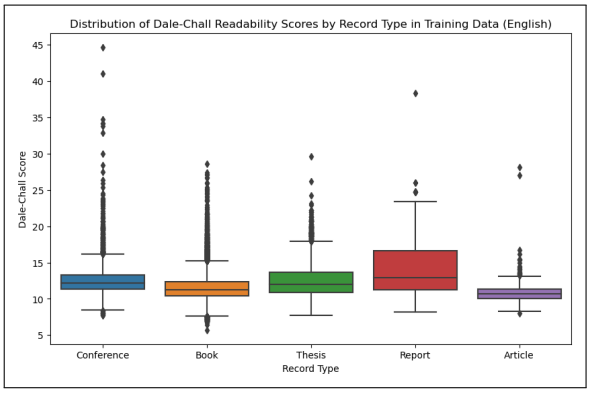

Figure 4: Readability Scores by Record Type in Training Data.

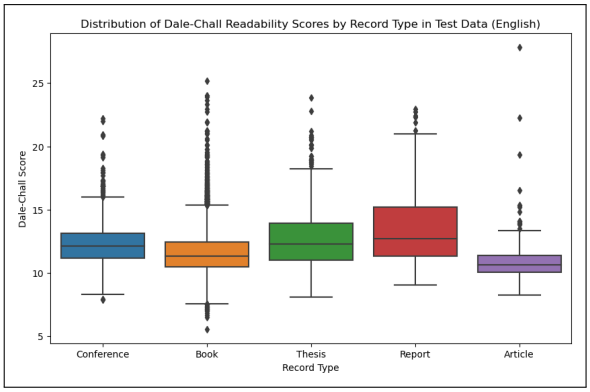

Figure 5: Readability Scores by Record Type in Test Data.

The thesis and report record types have both the most challenging readability and the lowest recall

scores by this system. When examining the scores for book and record types, which were the next two top-scoring record types for English language records, they have lower reading complexity. However, their subject distributions include a higher number of outlying values compared to article subject distributions.

This suggests that augmenting training data by rewriting abstracts for easier reading comprehension could result in performance gains at inference time. This represents a possible future use of generative AI in improving training. This needs more study particularly within those record types with a wide distribution of subjects. This analysis indicates that a lower incidence of wide subject distribution and lower complexity in abstract readability may improve recall scores.

5.3 Qualitative analysis

The highest qualitative performance, which was ranked 5th out of 13 teams, was scored by subject librarians at the TIB Technical Library. Scores are divided into two cases. In Case 1, a more expansive scoring criterion is used where both the correct keyword and irrelevant, but technically correct, subjects are considered correct. In Case 2, only correct subjects with no irrelevant subjects are scored as correct. The average of the qualitative recall scores in Case 1 (0.5263) was higher than the average score on the quantitative "All Subjects" leaderboard, which was 0.4686. However, in Case 2 (0.4258), the system did not surpass the average recall score of the quantitative leaderboard.

By design and by name, BERT is bidirectional, meaning that words are learned in context by looking both left and right. The recall performance in the results might be attributed to this bidirectional view of the training data, where the contextual notions of words are learned as part of the classification task. This idea also has theoretical grounding in the work of philosophers of language, such as Wittgenstein. In *Philosophical Investigations*, Wittgenstein theorized that context holds special importance to the meaning of words, specifically that the meanings of words are derived by their context (Wittgenstein and Anscombe, 2000). The notion of contextual relevance is particularly appropriate to consider in light of librarian scoring. Librarian expertise provides a valuable and necessary validation of the quantitative results of recall measures.

6 Conclusion

The system showed good performance in recall for English language articles. Evidence as to why these records are amenable to subject classification were considered. Specifically, an analysis of the subject distributions by record type and the readability or reading complexity of the record metadata was conducted. The system is completely portable from the Hugging Face platform. The ensemble models are all easily extensible into library systems, allowing experimentation to be taken into production without requiring extensive coding for adapting to local environments.

Acknowledgments

This work made use of the Illinois Campus Cluster, a computing resource that is operated by the Illinois Campus Cluster Program (ICCP) in conjunction with the National Center for Supercomputing Applications (NCSA) and which is supported by funds from the University of Illinois at Urbana-Champaign.

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Jeanne S. Chall and Edgar Dale. 1995. *Readability revisited : the new Dale-Chall readability formula*. Brookline Books, Cambridge, Mass. Section: 159 pages ; 26 cm.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. *Preprint*, arXiv:2205.14135.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Jennifer D'Souza, Sameer Sadruddin, Holger Israel, and Mathias Begoin. 2024. The SemEval 2025 LLMs4Subjects Shared Task Dataset.

Jennifer D'Souza, Sameer Sadruddin, Holger Israel, Mathias Begoin, and Diana Slawig. 2025. Semeval-2025 task 5: Llms4subjects - llm-based automated subject tagging for a national technical library's open-access catalog. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 1082–1095, Vienna, Austria. Association for Computational Linguistics.

Jim Hahn. 2021. Semi-Automated Methods for BIBFRAME Work Entity Description. *Cataloging & Classification Quarterly*, 59(8):853–867. Publisher: Routledge.

Jim Hahn. 2024. Bifurcation of Semi-Automated Subject Indexing Services. *Library Resources & Technical Services*, 68(3).

Kenneth Li, Aspen K Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2022. Emergent world representations: Exploring a sequence model trained on a synthetic task. *arXiv preprint arXiv:2210.13382*.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeff Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Bloom, Stella Biderman, Adria Garriga-Alonso, Arthur Conmy, Neel Nanda, Jessica Rumbelow, Martin Wattenberg, Nandi Schoots, Joseph Miller, Eric J. Michaud, Stephen Casper, Max Tegmark, William Saunders, David Bau, Eric Todd, Atticus Geiger, Mor Geva, Jesse Hoogland, Daniel Murfet, and Tom McGrath. 2025. Open problems in mechanistic interpretability. *Preprint*, arXiv:2501.16496.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Abhishek Thakur. 2024. AutoTrain: No-code training for state-of-the-art models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 419–423, Miami, Florida, USA. Association for Computational Linguistics.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *arXiv*.

Ludwig Wittgenstein and G. E. M. Anscombe. 2000. *Philosophical investigations: the English text of the third edition*, 3. ed edition. Prentice Hall, Englewood Cliffs, N.J.