YNU-HPCC at SemEval-2025 Task 5: Contrastive Learning for GND Subject Tagging with Multilingual Sentence-BERT

Hong Jiang, Jin Wang and Xuejie Zhang
School of Information Science and Engineering
Yunnan University
Kunming China
jianghong@stu.ynu.edu.cn, {wangjin,xjzhang}@ynu.edu.cn

Abstract

This paper describes YNU-HPCC (Alias JH) team's participation in the sub-task 2 of the SemEval-2025 Task 5, which requires fine-tuning language models to align subject tags with the TIBKAT collection. The task presents three key challenges: cross-disciplinary document coverage, bilingual (English-German) processing requirements, and extreme classification over 200,000 GND Subjects. To address these challenges, we apply a contrastive learning framework using multilingual Sentence-BERT models, implementing two training strategies: mixed-negative multi-label sampling, and single-label sampling with random negative selection. Our best-performing model achieves significant improvements of 28.6% in average recall, reaching 0.2252 on the core-test set and 0.1677 on the all-test set. Notably, we reveal model architecture-dependent response patterns: MiniLM-series models benefit from multi-label training (+33.5% zero-shot recall), while mpnet variants excel with single-label approaches (+230.3% zero-shot recall). The study further demonstrates the effectiveness of contrastive learning for multilingual semantic alignment in low-resource scenarios, providing insights for extreme classification tasks. Our implementation is publicly available at https://github.com/Jiangnaio/SemEval2025Task5.

1 Introduction

This task emerges in response to the challenges associated with manually tagging library bibliographic records (D'Souza et al., 2025). Our team is primarily involved in Task 2, which is dedicated to aligning GND Subjects with TIBKAT records. This task involves two datasets: the core dataset, which contains 79,427 subject labels, and the all dataset, comprising 204,739 subject labels. The data for the entire task is presented in two languages, German and English, necessitating the consideration of multilingual characteristics during the modeling process.

Data analysis highlights two critical characteristics that have significant implications for the task. Firstly, there is severe label underutilization, with only 31.9% (25,371 out of 79,427) of the core dataset labels and 16.5% (33,898 out of 204,739) of the all dataset labels appearing in the training and development subsets. Secondly, there is extreme subject sparsity, as over 50% of the topics contain two or fewer documents. This situation creates high-dimensional semantic space challenges.

Given the large number of subjects to classify in this task, traditional topic-classification approaches like Latent Dirichlet Allocation (LDA, (Blei et al., 2003; Jelodar et al., 2019)) struggle due to complex, non-parallelizable probability calculations that result in low computational efficiency.

Although BERT (Devlin et al., 2019; Koroteev, 2021; Zhou et al., 2024) can capture sentence semantics effectively, its generated sentence vectors suffer from anisotropy, being unevenly distributed in the vector space and concentrated in a narrow cone, leading to generally high calculated vector similarities. Most Transformer-based pre-trained models face this issue in the learned sentence-vector space (Gao et al., 2019; Ethayarajh, 2019).

Researchers (Cer et al., 2018; Conneau et al., 2018) have developed sentence-vector encoders using a "dual-tower" structure with sentence-task training datasets. With the advent of Sentence-BERT (Reimers and Gurevych, 2019), sentence vector representation has advanced significantly (Chi et al., 2023; Wang et al., 2025; Tavares and Ayres, 2025). It modifies BERT's structure and fine-tunes it via supervised tasks (Luo et al., 2022) , overcoming BERT's limitations in sentence-vector representation. Subsequently, (Gao et al., 2021) proposed contrastive learning for sentence-related tasks, often used to fine-tune Sentence-

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 2443–2448
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

BERT. (Huang et al., 2024) found that the loss function during Sentence-BERT training differs from that during prediction, causing poor performance in similarity determination. (Nielsen and Hansen, 2023) identified a hubness problem in Sentence-BERT's semantic space, which may also exist in this task. Future research could improve the loss function and optimize the training set to address these issues.

The pre-trained models used here are based on Sentence-BERT. The following sections detail our work and results.

2 Related Work

To address this task, we carried out a series of investigations. First, training datasets were created using three methods: the multi-label sample scheme, the single-label sample scheme with mixed negative samples, and the multi-label sample scheme with mixed negative samples. The prepared datasets were then utilized. Simultaneously, we explored and experimented with several pre-trained models from sentence-transformers (`https://www.sbert.net/`) and, prior to fine-tuning, tested them according to the official metrics (4.3). The test results are presented in Table 1.

Subsequently, we trained several sentence-transformers models using contrastive learning and reported the training results along with their analysis. Eventually, the average recall scores on the all-test and core-test datasets were 16.8% and 22.5% respectively (see Table 3). Moreover, we provided detailed experimental methodologies, procedures, some experimental results, and their analysis.

Furthermore, we proposed leveraging a translation model (Tiedemann and Thottingal, 2020) to conduct bidirectional translation between German and English, thus achieving data augmentation, and performing fine-tuning verification on the augmented data.

Lastly, we concluded the experiment, analyzed its limitations, and proposed subsequent improvement plans.

3 Methodology

Our architecture combines Sentence-BERT with contrastive learning to address the extreme classification challenge. As shown in Figure 1, the system processes TIBKAT-GND pairs through dual encoders with shared parameters.

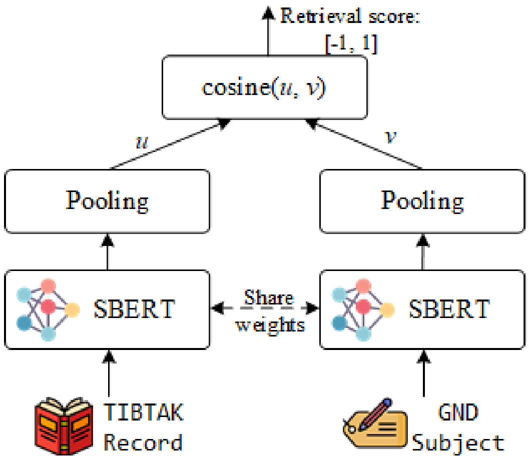

Figure 1: Flowchart illustrating the use of the Sentence-BERT model and cosine similarity to measure the similarity between two text segments: TIBTAK Record and GND Subject.

Semantic Encoding For a TIBKAT record d and GND subject s, we compute their embeddings:

$$\mathbf{h}_d = \text{SBERT}([d_{\text{title}}; d_{\text{abstract}}]) \tag{1}$$

$$\mathbf{h}_s = \text{SBERT}([s_{\text{Name}}; s_{\text{Alternate Name}}]) \tag{2}$$

where $[;]$ denotes concatenation and SBERT represents our fine-tuned model.

Balanced Sample Construction We create training pairs maintaining 1:1 positive-negative ratio:

- Positive: (d, s^+) where $s^+ \in \mathcal{S}_d^{\text{true}}$
- Negative: (d, s^-) where $s^- \in \mathcal{S}_d^{\text{false}}$

where $\mathcal{S}_d^{\text{true}}$ represents the set of all s that are related to d. $\mathcal{S}$ represents the set of all s or the GND Subjects used in the train and dev datasets. $\mathcal{S}_d^{\text{false}}$ represents $\mathcal{U}(\mathcal{S} \setminus \mathcal{S}_d^{\text{true}})$. (Li et al., 2024, 2025) We design two sampling paradigms for extreme classification scenarios:

Aggregate Multi-label Sampling: Aggregate all the true subjects of $\mathcal{S}_d^{\text{true}}$ into a single sample $(d, \mathit{Aggregate}(\mathcal{S}_d^{\text{true}}), 1)$, and randomly sample an equal number of $\mathcal{S}_d^{\text{false}}$ into a negative sample $(d, \mathit{Aggregate}(\mathcal{S}_d^{\text{false}}), 0)$ to construct 1:1 positive-negative pairs.

Instance-wise Disaggregated Sampling : Individually construct a positive sample $(d, s_i, 1)$ for each true subject s_i, and randomly select a negative subject s_i' from the candidate set $\mathcal{S}_d^{\text{false}}$ to create a negative simple $(d, s_i, 0)$.

Training Objective We optimize a temperature-scaled contrastive loss:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{\text{sim}(\mathbf{h}_d^{(i)},\mathbf{h}_s^{(i)+})/\tau}}{\sum_{j=1}^{K} e^{\text{sim}(\mathbf{h}_d^{(i)},\mathbf{h}_s^{(i)j})/\tau}} \qquad (3)$$

where $\tau = 0.05$ is learned during training, and $K = 2$ for our 1:1 sampling.

4 Experiments

4.1 Experimental Setup

We conducted systematic evaluations across three dimensions: (1) baseline performance of pre-trained models, (2) effectiveness of multi-label simple, and (3) effectiveness of single-label simple. Our implementation leveraged four multilingual Sentence-BERT variants from the sentence-transformers library, selected based on their zero-shot performance (Table 1).

Training Configuration

- **Batch size:** 32 (multi-label) / 16 (single-label)
- **Learning rate:** 2×10^{-5} with AdamW optimizer
- **Temperature parameter** τ**:** 0.05 (learned)
- **Training epochs:** 3 (core dataset) / 20 (all dataset)
- **Hardware:** 1×NVIDIA RTX 3060 GPU

4.2 Data Strategies

Upon analyzing the datasets, we discovered that even the GND subject covered by the datasets in the tib-core-subjects directory might not be found in the GND-Subjects-tib-core.json file. Therefore, we utilized the GND-Subjects-all.json file to create the datasets for training and evaluation.

4.3 Evaluation Metric

We employ three standard metrics for system evaluation:

$$Precision = \frac{Count(true\_set \cap pred\_set)}{k} \qquad (4)$$

$$Recall = \frac{Count(true\_set \cap pred\_set)}{Count(true\_set)} \qquad (5)$$

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall}, \qquad (6)$$

Model	P	R	F1
all-distilroberta-v1	0.0086	0.0820	0.0146
all-MiniLM-L6-v2	0.0112	0.1031	0.0190
all-mpnet-base-v2	0.0110	0.1043	0.0187
P-MiniLM-L6	0.0049	0.0439	0.0083
PM-MiniLM-L12	**0.0185**	**0.1751**	**0.0317**
P-mpnet-base-v2	0.0081	0.0773	0.0138
PM-mpnet-v2	0.0113	0.1069	0.0192

Table 1: Zero-shot performance comparison (Precision/Recall/F1 averages for k=5-50), the P of models represents ”paraphrase” and M represents ”multilingual”.

Among them, the set of true GND Subjects is represented as $true\_set$, and the set of the top k most relevant GND Subjects predicted by our model is represented as $pred\_set$. The number of elements in a set is counted by $Count(*)$, and the intersection of sets is denoted by $\cap$. $F1 - score$ will be 0 when $Precision + Recall = 0$.

For comprehensive analysis, we compute metric averages across k-values from 5 to 50 (incrementing by 5), reporting: Avg. Precision@k, Avg. Recall@k and Avg. F1-score@k

4.4 Baseline

We select the metrics of the PM-MiniLM-L12 (paraphrase-multilingual-MiniLM-L12-v2) model in Table 1 as the baseline.

5 Results and Analysis

5.1 Zero-shot Performance Baseline

Table 1 presents the zero-shot performance of various pre-trained models, establishing baseline metrics for subsequent comparisons. The paraphrase-multilingual-MiniLM-L12-v2 (PM-MiniLM-L12-v2) model demonstrates superior zero-shot recall (17.51%), outperforming other candidates by 63.8% relative to the second-best PM-mpnet-v2 model. This baseline analysis reveals significant performance variance across architectural variants, with MiniLM-series models showing particular promise for the target task.

5.2 Training Strategy Effectiveness

Table 2 compares the impact of different training approaches. The multi-label sampling strategy improves performance for all-distilroberta-v1 and P-MiniLM-L6, while both the all-MiniLM-L6-v2 and

Table 2: Training strategy comparison (Average Recall@k)

Model	Strategy	P	R	F1
all-distilroberta-v1	Multi-label	0.0135↑	0.1749↑	0.0238↑
all-MiniLM-L6-v2	Multi-label	0.0057↓	0.0573↓	0.0098↓
P-MiniLM-L6	Multi-label	0.0060↑	0.0586↑	0.0102↑
PM-mpnet-v2	Multi-label	0.0086↓	0.0889↓	0.0149↓
PM-mpnet-v2	Single-label	0.0336↑	0.3531↑	0.0578↑

PM-mpnet-v2 models exhibited performance degradation. The PM-mpnet-v2 model demonstrated significant performance improvement under the single-label training strategy. This dichotomy suggests an architecture-dependent optimization landscape where model capacity interacts with sampling strategy effectiveness.

5.3 Competition Results and Analysis

Our official submission results, presented in Tables 3, demonstrate the performance of the fine-tuned PM-MiniLM-L12-v2 model on both all-test and core-test datasets. It should be noted that due to an operational oversight, the potentially superior fine-tuned PM-mpnet-v2 model was inadvertently excluded from the final submission.

The fine-tuned PM-MiniLM-L12-v2 model achieved significant improvements, with a 28.6% enhancement in average recall on the core-test dataset (0.2252). However, its performance on the all-test dataset (0.1677) was comparatively lower, likely due to the model's exclusive training on the core dataset. This suggests potential domain adaptation challenges when transitioning from core to all datasets.

Metric Analysis Across Top-k Thresholds Figure 2 reveals three key patterns in metric behavior across different top-k values (k = 5, 10, 15, 20):

- **Precision** exhibits a clear negative correlation with k
- **Recall** demonstrates a strong positive correlation with k
- **F1-score** shows a moderate positive trend

These trends suggest that optimal k-value selection should be application-dependent, balancing the trade-off between precision and recall based on specific use case requirements.

Dataset	Avg. Recall	Relative Imp.
Core-Test	0.2252	+28.6%
All-Test	0.1677	-

Table 3: Comparative performance analysis across test datasets (Average Recall@k for k=5–50)

5.4 Improvement

Moreover, we tested the Alibaba-NLP/gte-multilingual-base (Zhang et al., 2024; Chen et al., 2024; Saad-Falcon et al., 2024) model and found that it still performs well before training.(0.0275 on Avg. Precision@k, 0.2508 on Avg. Recall@k, 0.0468 on Avg. F1-score@k) The multi-stage training method of the gte-multilingual-base model provides inspiration for subsequent research on this task. In the evaluation, we found that the accuracy of the model is not high. A two-stage prediction method can be adopted: First, use Sentence-BERT to select the top 500 most relevant Subjects, and then let the large model select the top 50 most relevant Subjects with prompt words. The relevant test code has been submitted to https://github.com/Jiangnaio/SemEval2025Task5.

6 Conclusion

In this study, we constructed training datasets using both multi-label and single-label sample methods, mixed them with negative samples at a 1:1 ratio, and then fine-tuned several pre-trained models based on the Sentence-BERT architecture using contrastive learning. Ultimately, we implemented and validated a solution for the subtask 2 of the SemEval-2025 Task 5 under low GPU memory conditions. Our approach enables efficient and rapid topic retrieval with limited computational resources. The best results of our model submitted for official evaluation achieved an average recall rate of 0.1677 on the all-test dataset and 0.2252 on the core-test dataset (performance demonstrated

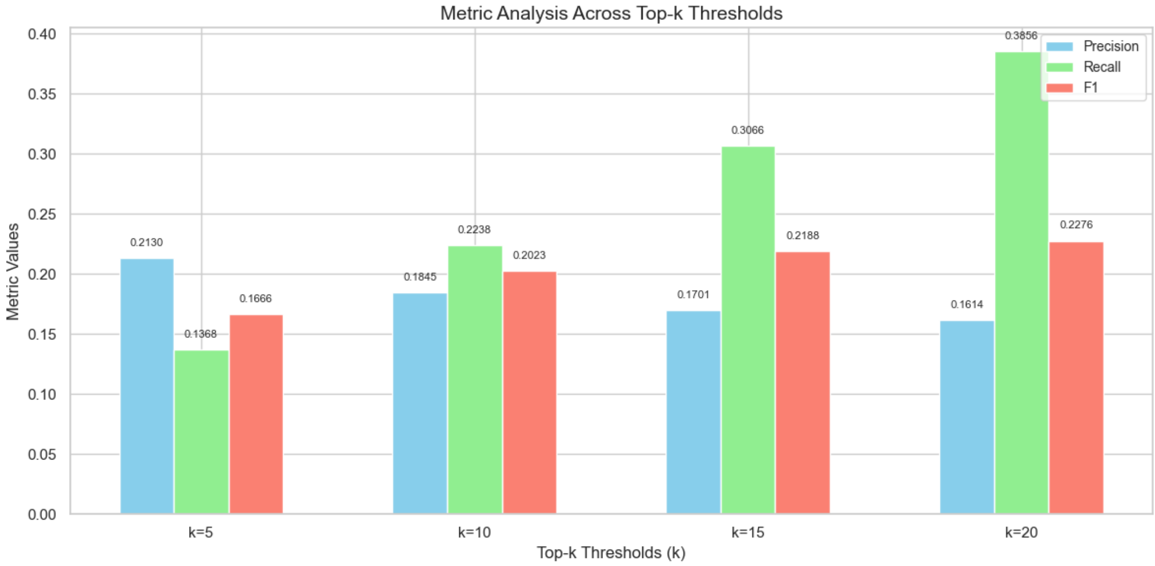

Figure 2: Bar chart of official test results

only in pre-trained models with fewer than 300M parameters).

For future research, bidirectional translation can be employed for data augmentation to address the issue of insufficient sample numbers in certain subjects. During inference, a two-stage approach combined with LLMs can be adopted. Detailed steps for these subsequent tasks are provided in the results and analysis section.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant Nos. 61966038 and 62266051. We would like to thank the anonymous reviewers for their constructive comments.

References

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Daniel Cer, Yinfei Yang, Sheng yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation.

Te-Yu Chi, Yu-Meng Tang, Chia-Wen Lu, Qiu-Xia Zhang, and Jyh-Shing Roger Jang. 2023. Wc-sbert: Zero-shot text classification via sbert with self-training for wikipedia categories.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. 2018. Supervised learning of universal sentence representations from natural language inference data.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

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.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings.

Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tie-Yan Liu. 2019. Representation degeneration problem in training natural language generation models.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *ArXiv*, abs/2104.08821.

Xiang Huang, Hao Peng, Dongcheng Zou, Zhiwei Liu, Jianxin Li, Kay Liu, Jia Wu, Jianlin Su, and Philip S. Yu. 2024. Cosent: Consistent sentence embedding via similarity ranking. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2800–2813.

Hamed Jelodar, Yongli Wang, Chi Yuan, Xia Feng, Xiahui Jiang, Yanchao Li, and Liang Zhao. 2019. Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey. *Multimedia tools and applications*, 78:15169–15211.

M. V. Koroteev. 2021. Bert: A review of applications in natural language processing and understanding. *ArXiv*, abs/2103.11943.

Weijie Li, Jin Wang, Liang-Chih Yu, and Xuejie Zhang. 2025. Topology-of-question-decomposition: Enhancing large language models with information retrieval for knowledge-intensive tasks. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2814–2833, Abu Dhabi, UAE. Association for Computational Linguistics.

Weijie Li, Jin Wang, and Xuejie Zhang. 2024. Promptist: Automated prompt optimization for text-to-image synthesis. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 295–306. Springer.

Xiang Luo, Yanqing Niu, and Boer Zhu. 2022. TCU at SemEval-2022 task 8: A stacking ensemble transformer model for multilingual news article similarity. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1202–1207, Seattle, United States. Association for Computational Linguistics.

Beatrix M. G. Nielsen and Lars Kai Hansen. 2023. Hubness reduction improves sentence-bert semantic spaces.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Jon Saad-Falcon, Daniel Y. Fu, Simran Arora, Neel Guha, and Christopher Ré. 2024. Benchmarking and building long-context retrieval models with loco and m2-bert.

Tiago Fernandes Tavares and Fabio José Ayres. 2025. Multi-label cross-lingual automatic music genre classification from lyrics with sentence bert.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.

Yumeng Wang, Ziran Zhou, and Junjin Wang. 2025. 2-tier simcse: Elevating bert for robust sentence embeddings.

Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, et al. 2024. mgte: Generalized long-context text representation and reranking models for multilingual text retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1393–1412.

Ce Zhou, Qian Li, Chen Li, Jun Yu, Yixin Liu, Guangjing Wang, Kai Zhang, Cheng Ji, Qiben Yan, Lifang He, et al. 2024. A comprehensive survey on pretrained foundation models: A history from bert to chatgpt. *International Journal of Machine Learning and Cybernetics*, pages 1–65.