BAVARIAN JOURNAL OF APPLIED SCIENCES®

# Topical Clustering of Unlabeled Transformer-Encoded Researcher Activity

**Zineddine Bettouche*** **Andreas Fischer***

https://doi.org/10.25929/1rjp-d197

## ABSTRACT

**Transformer models have the ability to understand the meaning of text efficiently through the use of self-attention mechanisms. We investigate the bundled meanings in clusters of transformer-generated embeddings by evaluating the topical clustering accuracy of the unlabeled scientific papers of the DIT publications database. After experimenting with SciBERT and German-BERT, we focus on mBERT as we work with multilingual papers. We create a landscape representation of the scientific fields with active research through the encoding and clustering of research publications. With the absence of topic labels in the data (no ground truth), the clustering metrics cannot evaluate the accuracy of the topical clustering. Therefore, we make use of the coauthorship aspect in the papers to perform a coauthorship analysis in two parts: the investigation of the authors' uniqueness in each cluster and the construction of coauthorship-based social networks. The calculated high uniqueness of authors in the formed clusters and the found homogeneity of topics across the connected components (in social networks) imply an accurate topical clustering of our encodings. Moreover, the constructed social networks indicate the existence of a set of connecting internal authors, whose collaborations with each other formed a large network, holding 74% of all papers in the database.**

*Transformer-Modelle haben die Fähigkeit, die Bedeutung von Texten mithilfe von Self-Attention-Mechanismen effizient zu verstehen. Wir untersuchen die semantische Bedeutung von Clustern, welche sich aus den durch die Transformer generierten Embeddings ergeben. Dabei wird die Treffsicherheit der thematischen Zuordnung ungelabelter wissenschaftlicher Publikationen aus der THD-Publikationsdatenbank bewertet. Nachdem wir mit SciBERT und German-BERT experimentiert haben, konzentrieren wir uns bei der Arbeit mit mehrsprachigen Artikeln auf mBERT. Die dargestellten Cluster der wissenschaftlichen Publikationen ergeben eine durchsuchbare Forschungslandschaft aller mittels Publikationen aktiven Disziplinen der THD. Da in den Daten keine Themenbezeichnungen vorhanden sind (keine Grundwahrheit), können die Clustering-Metriken die Genauigkeit des thematischen Clusterings nicht bewerten. Daher nutzen wir den Aspekt der Koautorenschaft in den Arbeiten, um eine Koautorenschaftsanalyse in zwei Teilen durchzuführen: der Untersuchung der Einzigartigkeit der Autorinnen und Autoren in jedem Cluster und dem Aufbau koautorenschaftsbasierter sozialer Netzwerke. Die berechnete hohe Einzigartigkeit der Autorinnen und Autoren in den gebildeten Clustern und die gefundene Homogenität der Themen über die verbundenen Komponenten (in sozialen Netzwerken) implizieren eine genaue thematische Clusterung unserer Kodierungen. Darüber hinaus weisen die konstruierten sozialen Netzwerke auf die Existenz einer Reihe miteinander verbundener interner Autorinnen und Autoren hin, deren Zusammenarbeit untereinander ein großes Netzwerk bildete, das 74 % aller Beiträge in der Datenbank enthält.*

* Faculty of Computer Science at the Deggendorf Institute of Technology

KEYWORDS

**Topical clustering, document similarity, document encoding, BERT, natural language processing, clustering, k-means, DBSCAN, keyword extraction**

*Thematische Gruppierung, Ähnlichkeit von Dokumenten, Kodierung von Dokumenten, BERT, Computerlinguistik, Clusteranalyse, k-Means, DBSCAN, Schlagwortextraktion*

## 1. Introduction

In recent years, the development of transformer models, such as the Bidirectional Encoder Representations from Transformers (BERT) model or the GPT series responsible for the popular ChatGPT, has revolutionized the field of natural language processing (NLP). These models have achieved stateof-the-art performance on various NLP tasks, including text classification, sentiment analysis, and question-answering.

A key feature of transformer models is their ability to encode the meaning of text efficiently. This allows them to generate contextualized mappings into a multidimensional vector space (embeddings) for each sentence. These embeddings can then be used as input to downstream tasks, such as classification or prediction. A clustering of these vectors is expected to highlight groups of semantically similar publications.

The focus of this article is on the topical clustering of scientific papers in the publications database of the DIT. These papers are published on certain topics, and their transformergenerated encodings reflect their corresponding topics. The goal of the approach is to highlight topical clusters without deliberately labeled data. The topic of each paper is represented by the contextualized transformer-generated embedding (vector). We assume that clustering the paper vectors leads to clusters representing the collective topic of its papers. Our investigation tries to establish that the identified clusters reasonably reflect the active research areas (in terms of published research papers) at the DIT.

The main research questions answered in this paper are:

- Can a fully unsupervised approach provide topical clusters that are semantically coherent and useful to understand the research landscape?
- Regarding multilingual input (which is relevant at any non-English research institution), can a multilingual model such as mBERT perform on par with a specialized model such as SciBERT, while including non-English texts?

In our previous work [1], we established a methodology for calculating the cross-distance between a pair of authors based on the respective encodings of their papers. We utilize this methodology to investigate the topics in the clusters. Initially, we reintroduce the encoding of data using Base-BERT and SciBERT and focus on obtaining a direct distance between any given pair of authors. We also utilize German-trained BERT models to process and investigate the German papers in our data that were cast aside previously. To consolidate all papers into a single landscape, we then employ a multilingual BERT model (mBERT), which provides efficient encoding regardless of language while still offering reasonable clustering performance.

Since topical labels are not pre-assigned in the publications database, the quality of the obtained clusters is not straightforward to measure. Instead, quality is verified in two ways: Author cluster uniqueness and coauthor cluster consistency. A high number of unique authors per cluster, i.e., authors belonging to only that cluster, indicates that authors are clustered in a meaningful way. Investigating the co-author social networks, network size is expected to correlate strongly with the number of clusters covered. The experiments performed verify that assumption.

Surprisingly, we discovered that even in the comparatively small publications database of the DIT, a large connected component is found, covering about 3/4 of all publications.

Still, even in this large component, the topical clustering is clearly recognizable. The results of this paper therefore support the idea of an unsupervised approach for identifying topical clusters of research topics. This is the basis for drawing a comprehensive research landscape of publishing authors at the DIT.

As for the structure of this article, Section II presents the background of the technologies we use, such as transformer models, clustering techniques, and social networks. Section III discusses the previous works that dealt with BERT models, semantic similarity, and the clustering of transformer-generated encodings. Section IV is a data analysis section, in which we analyze the data we use in terms of textual property distributions (i.e., character and token count distributions). Section V presents the methodology of our work. It is similar to the methodology used in our previous paper [1]; however, we discuss the new angle, which we make use of to achieve our aim in this work. Section VI presents the experiments done in this work, their implementation and the rationale behind them. Finally, Section VII concludes the work and sets up possible future developments that could be built on our findings.

## 2. Background

This section introduces the applied techniques, in particular: transformers, clustering techniques, cluster evaluation metrics, keyword extraction, visualization of high-dimensional vector spaces, and social network analysis.

### *A. Transformers*

Transformer models are a class of deep neural networks that have greatly advanced natural language processing (NLP) tasks in recent years. They were introduced in a landmark paper by Vaswani et al. [2]. Traditional NLP models, such as recurrent neural networks (RNNs) and convolutional neural networks (CNNs), have limitations in handling long-range dependencies and contextual information in a sentence or document. Transformer models overcome these limitations by using self-attention, a mechanism that allows the model to focus on the most relevant parts of the input text at each time step, while capturing long-range dependencies between words.

The transformer model consists of an encoder and a decoder, with self-attention as its key component. The encoder takes in the input text and encodes it into a sequence of hidden states, which are then used by the decoder to generate the output sequence. Self-attention is applied to each token in the input sequence to compute a weighted sum of all the tokens, with the weights determined by their similarity to the current token. This enables the model to attend to the most important parts of the input at each time step, while capturing longrange dependencies between words. Transformer models have achieved state-of-the-art performance on a wide range of NLP tasks, including language modeling, machine translation, and text generation, and are widely used in both academia and industry.

In our work, we use models of BERT [3] (Bidirectional Encoder Representations from Transformers) to encode our scientific papers, because they are capable of capturing semantic and contextual information in the text, which is crucial for understanding the research landscape of the papers. The encoded representations generated by transformer models are high-dimensional and dense, which can capture the complex relationships between the phrases in the papers. This makes the clustering process on the paper encodings likely to be dependent on the topics of these papers, which is what our paper attempts to investigate.

### *B. Clustering Techniques*

K-means [4] and DBSCAN [5] are two widely used clustering algorithms in machine learning. K-means is a partitioning algorithm that works by dividing data into K clusters, where K is a pre-defined hyperparameter. The algorithm starts by randomly selecting K points from the data as the initial cluster centroids. It then assigns each point in the dataset to the nearest centroid and updates the centroids to be the mean of the points in the cluster. This process is repeated until the centroids no longer change, indicating convergence. K-means is widely used due to its simplicity, speed, and scalability. However, it has some limitations, including its sensitivity to the initial selection of centroids and the assumption that the data is globular.

DBSCAN, on the other hand, is a density-based clustering algorithm that works by grouping together points that are closely packed together in high-density regions, while also identifying

points that are outliers. The algorithm defines clusters as areas of high density separated by areas of low density. It starts by selecting a random point and finding all the points that are within a pre-defined distance epsilon of that point. It then expands the cluster by recursively finding all the points that are also within epsilon of those points, until the cluster reaches its maximum density. The algorithm then repeats this process for other points in the dataset, assigning them to existing clusters or marking them as outliers. DBSCAN is useful in identifying clusters of arbitrary shape and is less sensitive to the initial parameters than K-means. However, it can be computationally expensive and requires setting two hyperparameters, epsilon, and the minimum number of points required to form a cluster.

In this paper, we cluster the BERT encodings of the papers in our database with K-means and observe the topics of the formed clusters. Density-based clustering is set to be addressed in future work.

*C. Cluster Evaluation Metrics*

Evaluating the quality of research paper clusters is crucial to ensuring that the resulting clusters are meaningful and accurate. We use Silhouette [6], Calinski-Harabasz [7], and Davies-Bouldin [8] metrics to evaluate the quality of the clusters. Silhouette measures how well each data point fits into its assigned cluster compared to other clusters, while Calinski-Harabasz measures the ratio of between-cluster variance to within-cluster variance. Davies-Bouldin measures the average similarity between each cluster and its most similar cluster. These metrics provide a quantitative measure of the quality of the research paper clusters. We record these metrics as we describe the clustering results for future reference. However, as we work with unlabeled data, it is hard to evaluate the clusters with only these metrics. Therefore, we employ social networks built upon coauthorships.

*D. Keyword Extraction with KeyBERT*

KeyBERT [9] is a state-of-the-art keyword extraction algorithm that uses the transformer architecture to extract the most relevant words or phrases from a given piece of text. Specifically, KeyBERT fine-tunes a pre-trained transformer model, on a large corpus of text to create a keyword extraction model. The algorithm works by first embedding the input text using the pre-trained transformer model and then using a cosine similarity function to compare the embedding of each word or phrase in the text to the overall text embedding. The words or phrases with the highest similarity scores are selected as the most relevant keywords for the text. KeyBERT has several advantages over other keyword extraction algorithms, including its ability to capture the context and meaning of words and phrases, its flexibility in handling different types of text, and its speed and efficiency. In this paper, we use KeyBERT to extract keywords from the research papers in each cluster, which allowed us to explore the topics present in each cluster to assess the topical clustering of our data.

*E. Visualization with UMAP*

UMAP [10], or Uniform Manifold Approximation and Projection, is a powerful dimensionality reduction technique that has gained widespread popularity in recent years. UMAP works by constructing a low-dimensional representation of high-dimensional data such that the local structure of the data is preserved as much as possible. Specifically, UMAP constructs a topological representation of the data using a fuzzy simplicial set, which captures the local relationships between points in the high-dimensional space. It then constructs a low-dimensional embedding of the data using a nonlinear optimization algorithm that preserves these relationships to the highest possible extent. The optimization process is guided by a loss function that balances the preservation of local structure with the need to spread out points in lowdimensional space. UMAP has several advantages over other dimensionality reduction techniques, including its ability to preserve both local and global structure, its ability to handle non-linear relationships between variables, and its speed and scalability for large datasets. In this paper, we use UMAP to visualize the clusters formed by our clustering methodology, providing a powerful tool for exploring the relationships between different research papers and their authors.

*F. Social Networks*

Social networks are a valuable resource for understanding the relationships and collaborations between individuals in a particular field of study. The *networkx* library [11] in Python provides an efficient and easy-

to-use tool for constructing and analyzing social networks. The library allows researchers to create graphs and networks, where nodes represent individuals and edges represent relationships between them. By analyzing the structure of these networks, researchers can gain insights into the patterns of collaboration and knowledge transfer in their field of study. In this paper, we use *networkx* to construct coauthorship networks, which provided us with a unique perspective on the relationships between authors in different research paper clusters. This analysis allowed us to identify authors who were unique to each cluster, providing further evidence for the topic-based nature of our clustering methodology.

## 3. Related Work

The use of transformer models for encoding and clustering scientific data has gained considerable attention in recent years. Guo et al. [12] presented an unsupervised clustering method for grouping scientific articles into meaningful clusters based on the encodings generated by transformer models. However, their data was not multilingual, and their work did not include density-based clustering. Similarly, Beltagy et al. [13] introduced SciBERT, a pre-trained transformer model that is specifically designed for scientific text. SciBERT is trained on a large corpus of scientific documents and has been shown to outperform general purpose language models in various downstream tasks such as named entity recognition and relation extraction. Multilingual clustering is another area where transformer models have been applied. In a paper by Artetxe et al. [14] a method was presented for unsupervised multilingual representation learning that can be used for clustering low-resource languages. Their method leverages cross-lingual encodings generated by transformer models to group similar words and phrases across different languages. This approach has the potential to significantly reduce the amount of labeled data required for clustering low-resource languages.

Concerning semantic similarity, Ostendorff et al. [15] found that SciBERT outperformed other models in measuring aspectbased document similarity. Chandrasekaran and Mago [16] noted that recent hybrid methods show promise in measuring semantic similarity. Kades et al. [17] developed methods to address semantic similarity in medical data using BERT. Yang et al. [18] demonstrated the use of transformer-based models in measuring semantic similarity in clinical texts and found that RoBERTa performed the best.

Social network analysis has also been a popular topic in the field of scientific research. Coauthorship networks, in particular, have received much attention due to their ability to reveal patterns of collaboration and knowledge exchange among researchers. In a paper, Newman et al. [19] provided an overview of coauthorship networks and their applications in different fields, including bibliometrics, sociology, and computer science. Meanwhile, Ravasz et al. [20] proposed a method for detecting overlapping and hierarchical community structures in networks. This method


```
{
    "id":"019844ce-e696-0c48-a2ac-1821047639e0",
    "abstractText":"A new facility designed to perf...",
    "title":"Calibration facility for airborne imaging spectrometers",
    "date":"30.06.2009",
    "referenceAuthors":[
        {"person":{"firstname":"P.","lastname":"G[redacted]"},
            "notes":null,"rank":0},
        {"person":{"firstname":"J.","lastname":"F[redacted]"},
            "notes":null,"rank":1},
        {"person":{"firstname":"P[redacted]","lastname":"S[redacted]"},
            "notes":"p[redacted].s[redacted]@th-deg.de","rank":2},
        {"person":{"firstname":"H.","lastname":"S[redacted]"},
            "notes":null,"rank":3}
    ]
}
```


Figure 1. Paper-Object Example

uses a combination of density-based clustering and hierarchical clustering to identify groups of nodes that are tightly connected to each other.

Density-based clustering algorithms have also been proposed for high-dimensional vectors, such as the DBSCAN algorithm introduced by Ester et al. [5]. DBSCAN is particularly effective at identifying clusters of varying shapes and sizes, which makes it a suitable algorithm for clustering highdimensional data such as transformer encodings. On the other hand, centroid-based clustering algorithms, such as K-means, have been widely used in clustering high-dimensional vectors. Bahmani et al. [21] proposed a scalable version of the Kmeans algorithm that is capable of clustering massive datasets efficiently.

Overall, these studies demonstrate the potential of using transformer models for encoding and clustering scientific data, as well as the importance of considering social networks and density-based clustering algorithms in this context. By leveraging the latest advances in machine learning, we can generate clusters of the BERT encodings of our research papers, and perform a topic-based evaluation by means of coauthorship-based social networks.

## 4. Exploratory Data Analysis

This paper uses the same data as our first results paper [1], consisting of 7,548 references. The data include various types of references, but only 1,500 (sic) scientific articles with abstracts are chosen for this investigation. Each selected entry has at least a title, a list of authors, a date, and an abstract, topically unlabeled. An example of an entry is shown in Figure 1, where authors are marked by name and internal authors are identified by their e-mail addresses. We remove entries that have over 512 tokens in their abstract. The BERT models have a limitation of 512 tokens per input (with the exception of SciBERT, which is limited to 768 tokens). If a text item has more tokens, the BERT model will truncate the input, which can result in the loss of valuable information. Figure 2 shows the histogram of the token count per abstract. After the removal of the items with no abstracts (presentations, interviews, etc.) and the items with an abstract token count over 512, we end up with 1,459 items. An overview is shown in Table I.

| Selection criterion | Value |
|---|---|
| Total # publications | 7,548 |
| Publications w. abstracts | 1,500 |
| Abstracts < 512 tokens | 1,459 |

Table 1: Summary values for the publications database

## 5. Methodology

This section presents the general processing approach implemented in this work and the methodology used to calculate the cross-distances

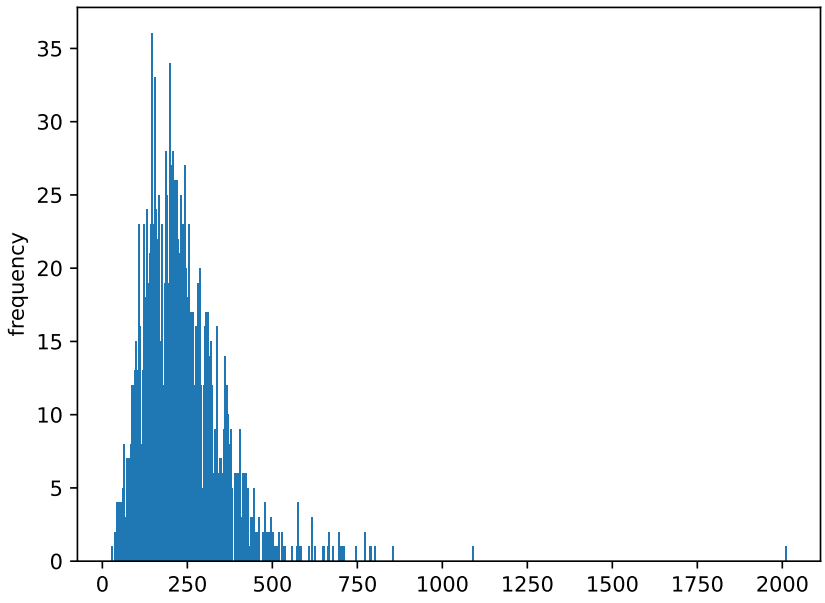


Figure 2. Histogram plot of the tokens count per abstract

between the authors. These cross-distances are the main contribution of our previous work [1]. We reintroduce them here to merge the results of the authors' social networks and the results of the distances between these authors.

*A. General Approach of Processing*

The goal of our implementation is to compare the results of the different BERT models and the different clustering techniques. In parallel, we extract the keywords for each text and use the generated labels (cluster labels) to map these keywords into their respective clusters. This attempts to assign a research field (a topic) for these clusters. We check the accuracy of such assignments by performing a coauthorship analysis afterwards. Figure 3 shows the implementation overview of the processing pipeline. The abstracts are fed to the BERT models as input to generate the high-dimensional encodings (or HD vectors), with each HD vector representing its respective abstract. We cluster the HD vectors using the different clustering techniques (centroid- and density-based) to obtain a set of 1,459 labels. Each label is a natural number from 0 to $n$ (for $n$+1 clusters). We map the HD vectors onto the 2D plane to plot them with label-based coloring. We take the text of each abstract, extract its keywords using KeyBERT, and form clusters of keywords by assigning each set of extracted keywords the respective label of the abstract HD vector.

In terms of centroid-based clustering, we set the clustering process to perform 10 runs, each run with *maximum_iterations* set to 100. At the end of each run, we calculate the sum of the three chosen metrics: Silhouette, Calinski-Harabasz, and Davies-Bouldi. At the end of this iterative process, we take the labels that correspond to the highest sum. We record our results for future reference.

*B. Distances between Authors*

The database we use in our work contains a set of authors and their corresponding research papers.

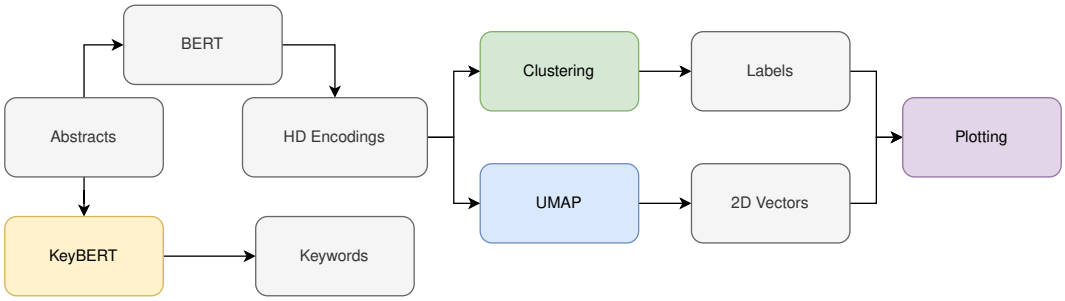


Figure 3. Implementation Overview

We have encoded each paper using BERT models and computed the average distance between the encodings of each pair of authors, which we refer to as the cross-distance. The cross-distance is an indicator of the similarity between two authors in terms of their research topics. Let $P_1$ be the set of papers by author 1 ($A_1$) and $P_2$ likewise be the set of papers by author 2 ($A_2$). Then the distance between authors 1 and 2 is defined as:

$$crossDistance(A_1, A_2) = \frac{\sum_{p_1 \in P_1} \sum_{p_2 \in P_2} \text{dist}(p_1, p_2)}{|P_1| \cdot |P_2|} \quad (1)$$

We have analyzed the self-distance of each author by computing the average distance between the encodings of their own papers. A lower self-distance value reflects the author's precise focus on a specific field. Our previous study [1] has revealed that if two authors have coauthored one or more papers, their cross-distance value is on average lower than the total average distance value, indicating a higher similarity in research topics. Furthermore, we have found that authors with lower self-distance values tend to have a more precise research focus. This demonstrates the effectiveness of the BERT model in encoding research papers to identify similarities between authors in terms of research topics.

We employ the concept of cross-distances to investigate the correlation between the authors' self-distances and their respective number of connections within their social networks.

## 6. Experiments

In this section, the experiments done throughout this work and both the rationale behind them and the results obtained from them are discussed.

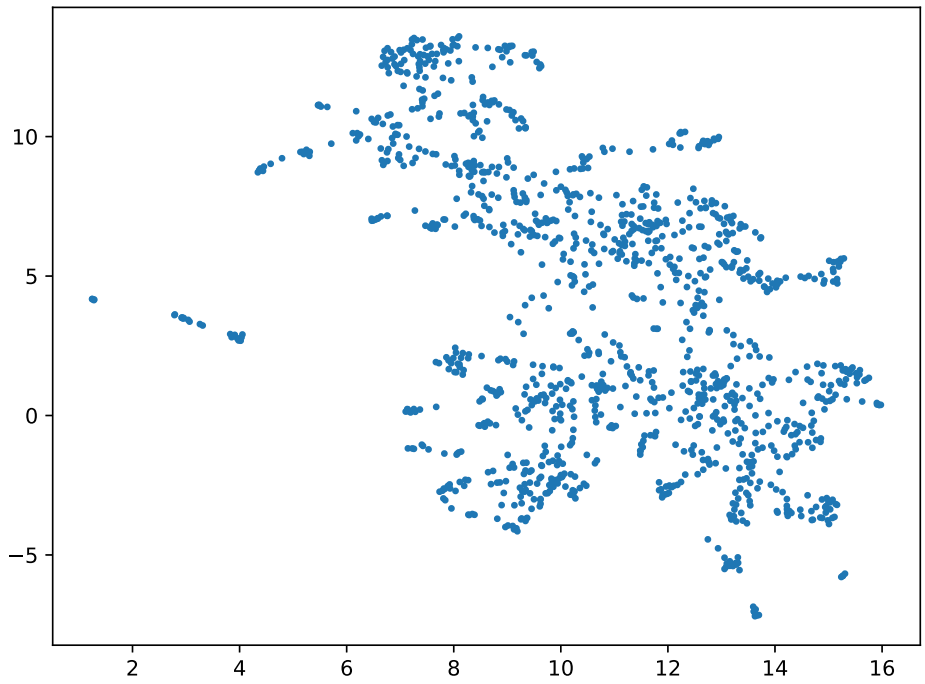


Figure 4. Scatter Plot of UMAP-Reduced SciBERT Encodings

In our previous work [1], the focus was on the Base-BERT model. However, this has two drawbacks: It is not specialized in scientific articles, and it does not cover multilingual input very well. Here, the focus is on investigating and comparing two other models: SciBERT, which is trained specifically on scientific articles, and mBERT, which is trained on multilingual data. A side comparison with BERT models trained on German data is also provided.

*A. Scientific text-oriented Processing: SciBERT*

SciBERT is a modified version of BERT that was trained on scientific text data. The training process for SciBERT was similar to that of BERT, but with some modifications to better handle scientific language. To train SciBERT, the researchers used a large corpus of scientific papers from a variety of fields, including computer science, biology, and physics. The corpus was preprocessed and tokenized in the same way as BERT, using the WordPiece algorithm. The architecture of SciBERT is the same as that of BERT, but the pre-training process was modified to better handle scientific language. The resulting model has been shown to outperform BERT on a range of scientific text-related tasks, including named entity recognition, relation extraction, and sentence classification. We select only the English papers in the database and encode them with SciBERT (our data includes both English and German papers). Figure 4 shows the UMAP-2D vectors of the SciBERT encodings.

We cluster the high-dimensional vectors that are generated by SciBERT using K-means to observe the initial distribution of the English papers. We cluster the high-dimensional vectors instead of the 2D-mapped vectors because the HD vectors contain more information, while the mapped ones are only their projections. The distance between a pair of HD vectors is not in correlation with the distance between the pair of their respective projections. This is also known as the binary stars situation. Figure 5 shows the K-means clusters of the SciBERT encodings (7 clusters). We have noted on the figure the titles of two papers from two clusters randomly selected. By reading these titles, we observe that there might be a similarity of topics in each cluster. For reference, Silhouette, Calinski-Harabasz, and Davies-Bouldin scores are 0.102, 112.266, and 3.067, respectively.

To investigate this, we extract the keywords of the papers in each of these clusters through our KeyBERT-based processing pipeline. Table II displays the major keywords of each cluster. From this table, we observe that despite the similarity between clusters 0 and 6 and the ambiguity of cluster 2, each of the other clusters has a distinct field. However, the data that was fed to SciBERT

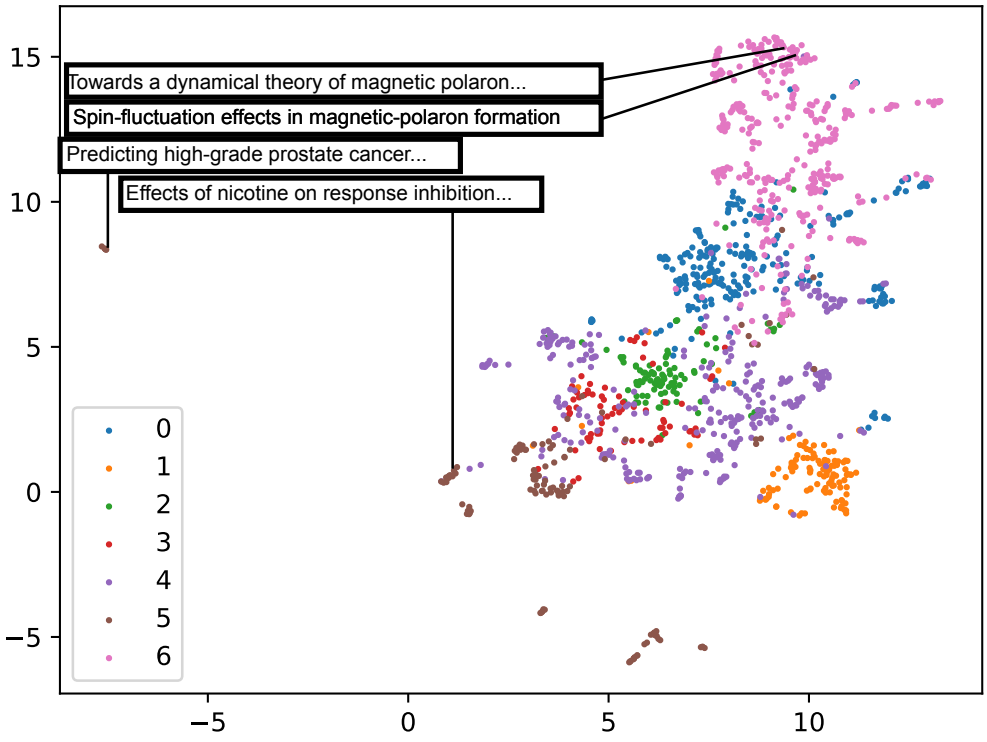


Figure 5. K-means Clusters of SciBERT Encodings (English only)

| cluster | keywords |
|---|---|
| 0 | exciton, quantum, magnetic, electron, polaron |
| 1 | video, classification, recommender, data, 3d |
| 2 | computer tomography, systems, digitalisation, management |
| 3 | optical, polishing, surface laser, machining |
| 4 | melanoma, gene, macrophage, health, biomarker |
| 5 | renewable, solar, photovoltaic, sensor, microgrid |
| 6 | dielectric, plasma, microscopy, nanowire, oxide |

Table 2: KeyBERT-generated Cluster keywords of SciBERT encodings

did not include the German papers. Therefore, we process these papers using German-trained BERT models, as they have been cast out of the processing pipeline so far.

*B. Non-English Data: Processing German Papers*

As previously mentioned, the vectors representing the papers written in German were put into one cluster, as SciBERT has embedded the German texts very similarly to each other, and distinctively from the English texts. The obtained similarity, however, only represents linguistic differences and is not topic-based.

We chose the transfomer models that were trained specifically on German data: German Base-BERT Cased, and German Base-BERT Uncased (DBMDZ).

- *Cased German Base-BERT:* The authors trained on a single cloud TPU v2 with the default settings using Google's Tensorflow code. They trained 30k steps with a sequence length of 512 and

810k steps with a batch size of 1,024 for sequence lengths of 128. While it takes roughly nine days to train, they used news articles, the most recent German Wikipedia dump (6GB of raw txt files), and the OpenLegalData dump (2.4 GB) as training data (3.6 GB). With the help of customized scripts and spacy v2.1, they cleaned the data dumps and utilized the suggested sentencepiece library to build the word piece vocabulary and tensorflow scripts to turn the text into data that could be accessed by BERT in order to construct tensorflow records.

- *Uncased German Base-BERT (DBMDZ):* The work offers another German-language model in addition to the released German BERT model from deepset. A recent Wikipedia dump, the EU Bookshop corpus, Open Subtitles, CommonCrawl, ParaCrawl, and News Crawl make up the model's underlying data. As a result, a datase with 2,350,234,427 tokens and a 16 GB size is produced. The authors employed spacy to separate sentences, and the same preprocessing techniques as those used to train SciBERT (sentence fragment model for vocabulary creation). The model underwent 1.5 M steps of training with a starting sequence length of 512 subwords.

We chose to use both cased and uncased models for the German language because while uppercase nouns may suggest that case is more significant in German than in English, it does not necessarily mean that a cased model will perform better on all tasks. In cases like part-of-speech detection, it is unclear whether the benefits of having a much larger vocabulary from using a cased model outweigh the added complexity. Cased models have separate vocabulary entries for differently-cased words. To observe the potential variations that could occur with different casings, we applied each model to the data and clustered the resulting vectors. Figures 6 and 7 display the outcomes of the cased and uncased models, respectively.

The uncased model in Figure 7 has produced vectors that are closer to each other (in terms of their 2D projection) than the cased model in Figure 6. The latter model apparently makes a sharp distinction between one of the clusters (top left) and the rest. However, the clusters in each graph could not be distinguished to the point of falling into a certain field or topic. As for clustering metrics, the scores of Silhouette, Calinski-Harabasz, and Davies-Bouldin for the cased German BERT are 0.100, 19.480, and 2.815, respectively. Whereas for the uncased German BERT, these respective scores are 0.047, 8.879, and 3.326. Table III shows the keywords of these clusters. From Table III,

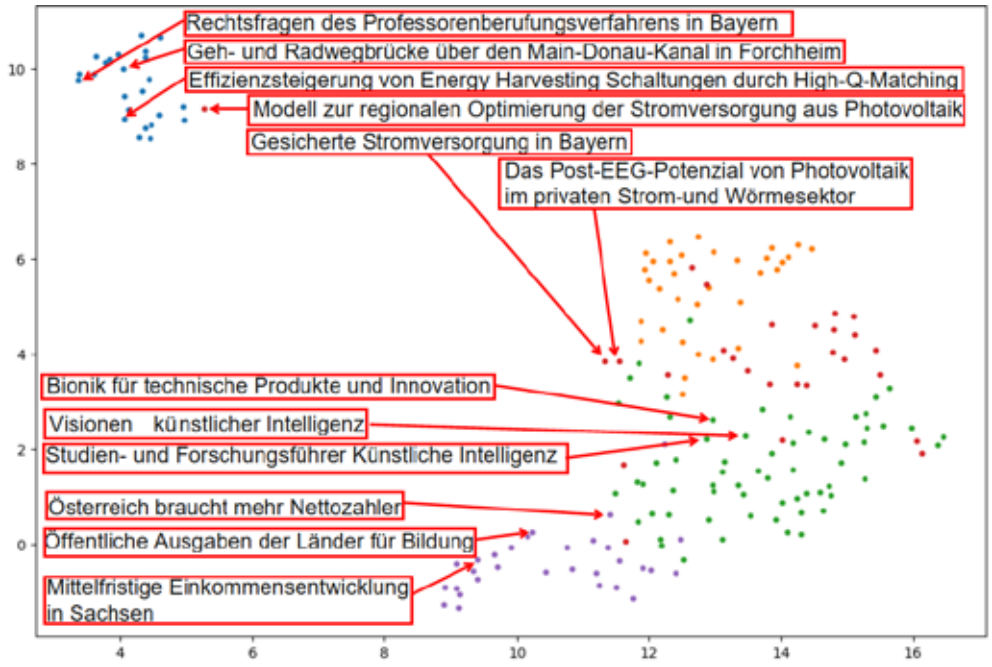


Figure 6. Scatter Plot (Clusters) of UMAP-Reduced Encodings of German Base-BERT Cased

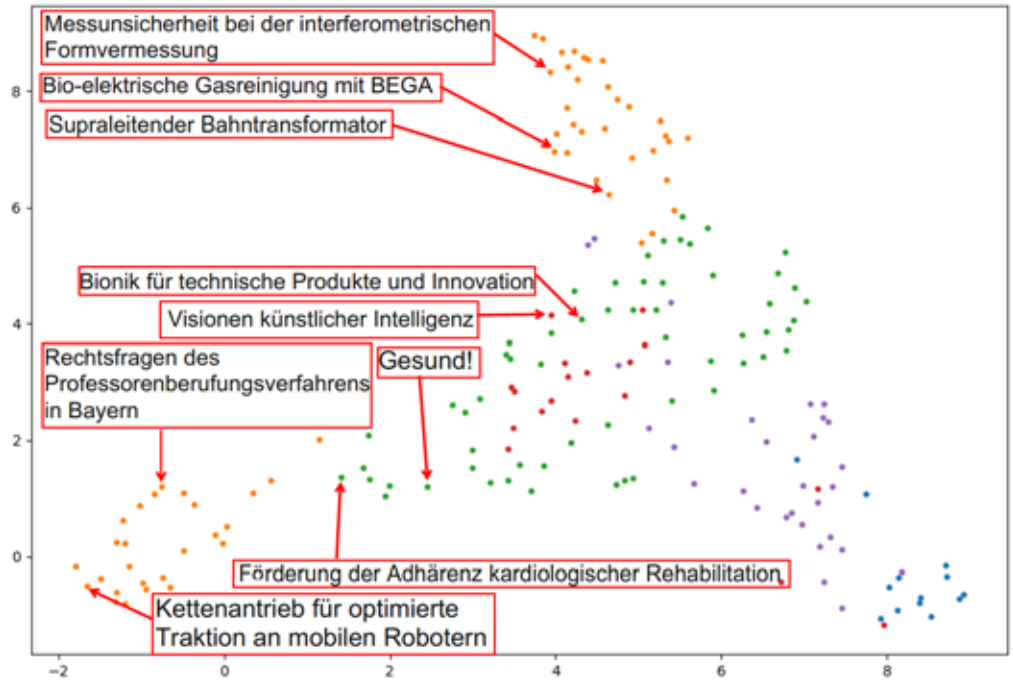


Figure 7. Scatter Plot (Clusters) of UMAP-Reduced Encodings of German Base-BERT Uncased (DBMDZ)

| Cased German BERT | |
|---|---|
| **cluster** | **keywords** |
| 0 | metallkörper, datenanalyse |
| 1 | computertomographie, visualisierung |
| 2 | digitalisierung, innovationsnetzwerke |
| 3 | innovationsmanagement, digitalisierung |
| 4 | beschäftigung, bevölkerung |
| **Uncased German BERT (BDMDZ-BERT)** | |
| **cluster** | **keywords** |
| 0 | beschäftigung, bevölkerung |
| 1 | faserverbundkunststoffe, ethernet |
| 2 | digitalisierung, innovationsmanagement |
| 3 | lernortkooperation, destinationsmanagement |
| 4 | diskriminierung, arbeitseinstellung |

Table 3: Cluster-Keywords Table for German Papers

the exclusive processing of German papers was not exact enough to the point of drawing clear topics, which in turn could represent the landscape of research published in German by the institute.

The uncased model presented in Figure 7 yielded vectors that were more closely situated (based on their 2D projection) than the cased model depicted in Figure 6. The latter model sharply separated one of the clusters (top left) from the others. However, neither of the models allowed for clear differentiation among the clusters with respect to particular fields or topics. In terms of clustering metrics, the Silhouette, Calinski-Harabasz, and Davies-Bouldin scores for the

cased German BERT are 0.100, 19.480, and 2.815, respectively. By contrast, the respective scores for the uncased German BERT are 0.047, 8.879, and 3.326. Table III presents the cluster keywords of these clusters. Overall, our exclusive processing of German papers did not achieve sufficient granularity to reveal distinct research topics that might represent the landscape of German research published by our institute. Therefore, we intend to utilize a multilingual BERT model that can process all of the papers at once to construct the clusters in an appropriate way that spans the whole set of abstracts.

*C. Processing Multilingual Data: mBERT*

Multilingual BERT (mBERT) is a language model developed by Google that can understand and generate text in multiple languages. It is trained on a large corpus of text from 102 different languages, allowing it to effectively model and generate text in diverse linguistic contexts. mBERT uses a transformer-based architecture that employs bidirectional encoding to capture contextual relationships between words in a sentence. This architecture enables it to perform a range of natural language processing (NLP) tasks such as named entity recognition, sentiment analysis, and machine translation. Additionally, mBERT is capable of performing cross-lingual transfer learning, which means that it can transfer knowledge from one language to another and use this to improve the accuracy of its predictions. These features make mBERT a powerful tool for multilingual NLP tasks and have led to its widespread use in academia and industry.

We use mBERT to encode all of the papers in our database. This permits a complete representation of the research landscape in the institute, based on which we generate more fitting clusters. We can then investigate the appropriateness of the formed clusters (in terms of topic) by observing the extracted keywords from each cluster and performing a coauthorship analysis. The mBERT encodings of our data are mapped onto a 2D plane and plotted in Figure 8. The vectors produced by mBERT can be seen to be held on one continent, which initially indicates an appropriate handling of papers regardless of the human language used (as each of our papers is written in either English or German).

We now perform the clustering process on the generated mBERT encodings. The obtained clusters are shown in Figure 9. We have noted on the graph a few random points from each cluster. The initial observation is that the points in each of the groups 1, 2, and 4 fall in one field: computer tomography in group 1 (computer science department), Chinese-German medicine in group 2 (health department), and energy in group 4 (power department). Group 3 is an example of points that appear far on the graph but still fall under one topic (health).

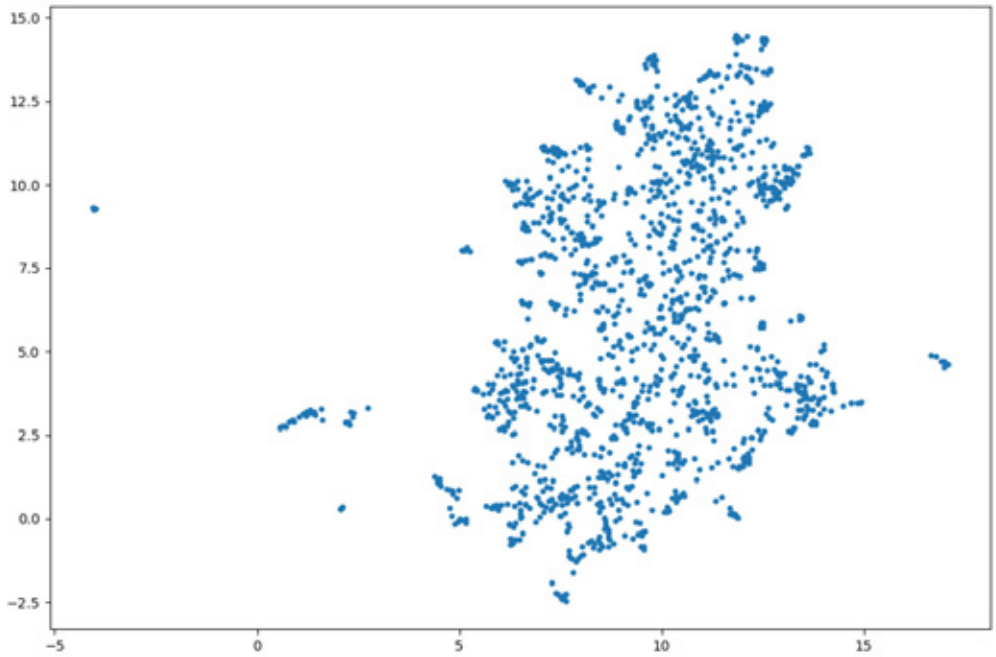


Figure 8. Scatter Plot of UMAP-Reduced mBERT Encodings

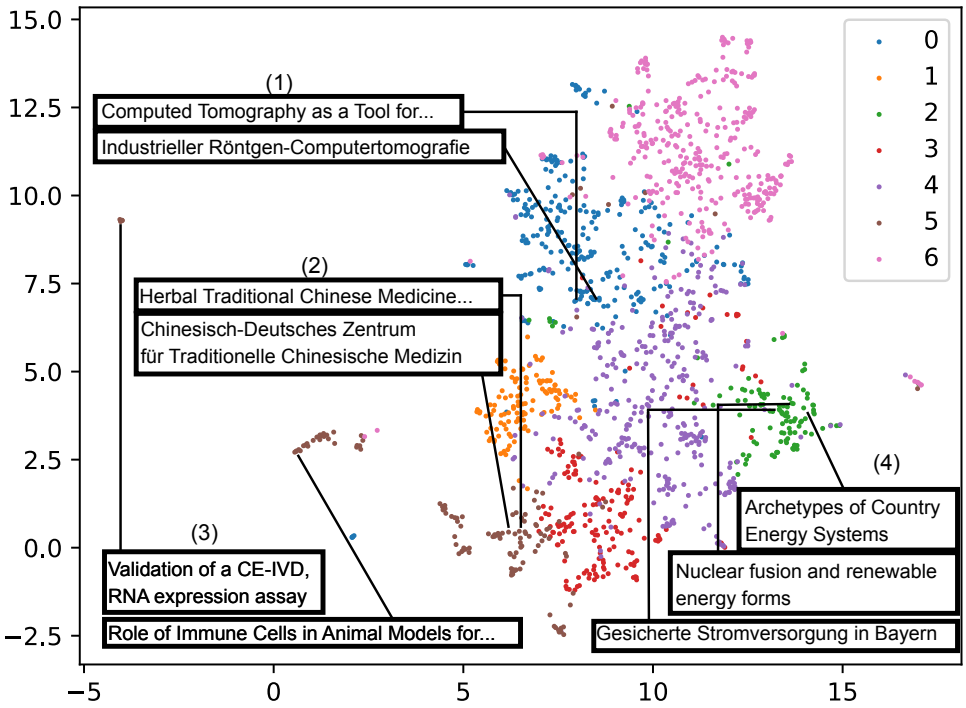


Figure 9. Clusters of mBERT Encodings (multilingual set of papers)

| Multilingual Data | | |
|---|---|---|
| **cluster** | **keywords** | **topic (manual)** |
| 0 | polishing, grinding, welding | Manufacturing |
| 1 | 3dtv, stereoscopic, resolution | Media |
| 2 | renewable, emissions, photovoltaic | Power |
| 3 | tourism, resorts, pension | Economics/Management |
| 4 | classifier, recommender. virtualization | Computer Science |
| 5 | prostate, aerobic, schizophrenia | Health |
| 6 | nanowire, dielectric, semiconductor | Material |

Table 4: Cluster-Keywords Table of Multilingual BERT

The mBERT model has encoded similar papers close to each other regardless of their language (represented by groups 1, 2, and 4 on the graph). The HD-vector clustering appears to be accurate, as each group of papers is held in their own shared cluster, although UMAP has mapped them far from each other at times in 2D (group 3). For reference, the Silhouette, Calinski-Harabasz, and Davies-Bouldin scores are 0.035, 30.369, and 4.468, respectively.

We extract the keywords for each of the formed clusters using KeyBERT to obtain the general topic of each of them. Table IV shows the result of the keyword extraction process. The topics formed from the keyword extraction are the most precise so far (even distinguishing between Media Engineering and Computer Science papers). Having a consistency of keywords across each cluster while spanning all data regardless of human language makes the generated clusters a good reflection of the topics in our database. The observed topics match the departments that are active in research at the DIT. The topics of health, economics, and computer science match their respective departments. The materials topic is in the natural sciences department. Manufacturing is divided

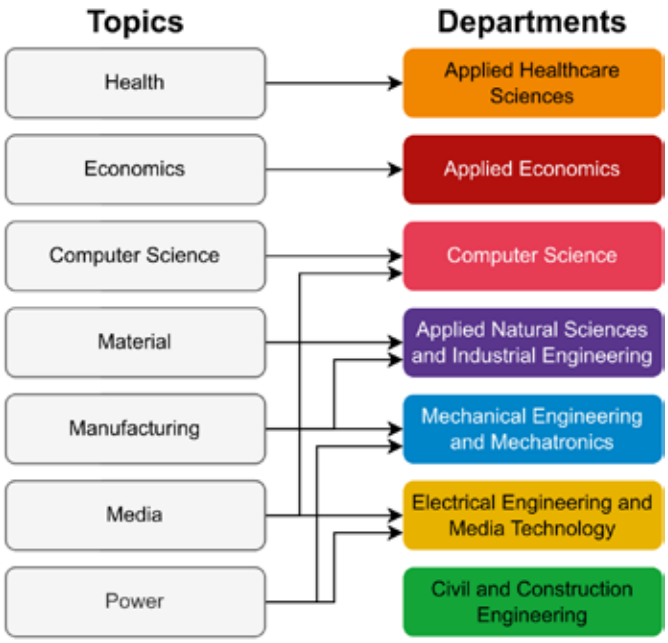


Figure 10. The Observed Cluster Topics in relation to the Departments at the DIT

between the departments of natural sciences and mechanics. Media is also divided between media and computer science departments, and power is divided between electrical engineering and mechanics departments. The department of civil engineering lacks the respective cluster, which is the product of having a smaller set of research papers in comparison to the other departments. This is shown in Figure 10. There exists an 8th department at the DIT, which is the European Campus Rottal-Inn. This department offers a set of different programs, such as industrial engineering and digital health. The papers from this department do not have a common topic but fall into different ones. Therefore, it was topically indistinguishable in the formed clusters.

However, as our dataset is unlabeled, it remains difficult to determine the reliability of the constructed clusters. Therefore, we go beyond the systematic use of cluster metrics by employing the coauthorship aspect of our data to determine the accuracy of this topical clustering.

*D. Authors and Clusters: A Relationship to Investigate*

In Subsection V-B, we have stated the term *self-distance*. This self-distance represents the breadth of research topic for each author (introduced in our previous work [1]). We have observed that even the author with the highest self-distance publishes in one cluster. If authors generally publish their papers in one cluster (one topic), the exclusiveness of authors within a cluster is an indication of its construction accuracy.

Let $L_n$ be the list of clusters $(C_0, C_1, ..., C_k)$ that author $A_n$ is involved in, and $IP(A_n,C_m)$ the Involvement Percentage of author $A_n$ in cluster $C_m$. $IP(A_n,C_m)$ is then defined as:

$$IP(A_n, C_m) = \frac{count(C_m \in L_n)}{|L_n|}$$

The Uniqueness Percentage $UP(C_m)$ of a cluster $C_m$, in terms of how exclusive the authors in the list of its authors $U_m$, is defined as:

$$UP(C_m) = \frac{\sum_{A_n \in U_m} IP(A_n, C_m)}{|U_m|}$$

Figure 11 shows the results of our UP calculations. The lowest average percentage of *uni-clusteric* authors is 80.85% (cluster 3), meaning that most clusters have over 80.85% of unique authors. Although the papers of an author are generally of one cluster, the authors can branch out and collaborate with other researchers in different fields. For example, if *author* A publishes mainly on computer science

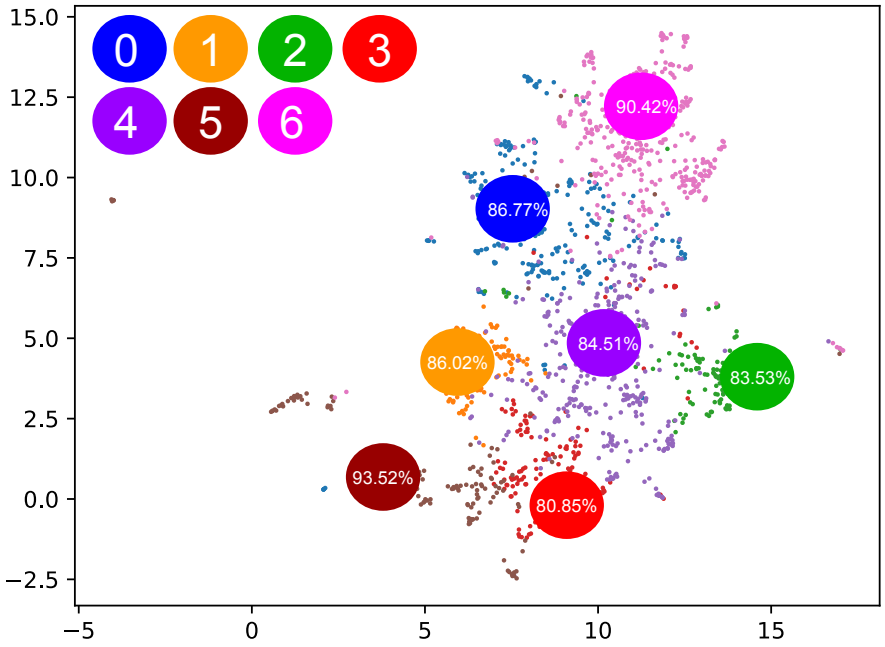


Figure 11. The Uniqueness Percentage of Each of The 7 Clusters in mBERT Encodings


Figure 12. Total Count of Authors in Each Cluster, Along with Count of Shared Authors

topics (cluster 4), but occasionally collaborates with colleagues in the medical sciences field (cluster 5), *author* A is said to be a shared author between clusters 4 and 5. The example of an author being mainly in one cluster and branching out on a few occasions is practical. To visualize the authors shared between cluster pairs, we plot the number of authors inside each cluster and the number of shared authors between them, as shown in Figure 12. We observe a few pairs of clusters that share high numbers of authors, such as clusters 0 and 6 (85 shared authors), clusters 0 and 4 (56 shared authors), and clusters 3 and 4 (61 shared authors). The topics of two clusters can be close enough to overlap, implying authors with papers in both topics, such as clusters 0 and 6 with Industry and Material Engineering or clusters 3 and 4 with Computer Science and Economics (Econ-Informatics being a major sub-department of

Economics). We have observed that our papers contain the general case of authors publishing in one cluster but still coauthoring research with different-field authors on shared topics (or topics that make use of two different fields, such as image processing in medical engineering). The percentage of unique authors (over 80.85%) in each cluster implies an initial indication of an accurate topical clustering of papers. However, to affirm such an indication, we use our data to construct social networks based on coauthorships between researchers. The construction of connected components (research groups) with topical homogeneity would affirm the accuracy of the topical clustering.

*E. Constructing Social Networks of Research Groups*

We construct the social networks that reflect the relationships between the authors. The constructed connected components are expected to represent the research groups, in which the authors take part. The authors are grouped by *coauthorship*, with coauthors as edges to the components. Using the *networkX* library, the connected components of authors are formed. Figures 13 and 14 show the correlation between the edge count and the self-distance of each author and the bar plot of that distribution per author.

Figure 13 indicates that edge counts are exponentially proportional to the self-distances (representing the topic breadth of an author). Therefore, the more topics an author has, the more likely it is that connections will be made. The number of connections in this case is represented by the edge count. In Figure 15 we plotted the edge count in relation to cluster count. It shows that the higher the number of authors in a component, the higher the number of clusters included.

Concerning the node colors in the following graphs, an author with papers strictly falling into one cluster is assigned the color of this cluster. The color gray is assigned to authors having papers in different clusters. We observe the obtained cases in the formed networks:

*1) Single paper with multiple authors:* Figure 16 shows an example for a single paper written by 7 authors.

*2) Close cooperation between authors:* Some groups of authors cooperate very closely. In such a case, we expect a small, fully connected graph where every person cooperates with every other person. Such a case is presented in Figure 17. Three authors work in the same field, publishing three different papers with each other. The group is isolated from other researchers but closely knit within itself.

*3) Close research network:* When different research groups work in the same field, cooperation between them is relatively easier. Figure 18 shows a case of different research leaders (marked with blue) collaborating with each other. When these researchers collaborate, they do not always bring their groups with them. Groups 1, 2, 3, and 4 are not connected. However, groups 4 and 5 are partially connected. The lead researchers of groups 1, 2, and 3 are fully connected to these groups, despite their blindness to each other. The research leaders are all connected to each other, except for the one marked in red. The multiple works (13 papers) of this close research network are all assigned to the same cluster 6 (field of Materials). This indicates accurate topical clustering.

*4) Leader of research group:* A senior researcher can be the leader of multiple research groups. Figure 19 shows an example of one internal research leader (gray node) and the multi-topical research associates. The gray node is connected to every other node in this component, indicating that the leader has worked with every other apparent associate in the graph. At least 7 distinct research groups can be identified. The distinction between two groups is drawn from the nodes of each group connecting only within. The node with the blue mark signifies a vice-leader between groups 1 and 2. A research leader with distinct yet same-field groups indicates that different yet close topics are addressed. For example, if a research leader publishes in both computer vision and natural language processing, the papers produced are going to fall into the computer science field. However, this leader can work with two distinct groups, each focusing on one topic in the field. The research leader in the graph has branched out in two papers (one in computer science and the other in power engineering). All of the other papers are by teams focusing on materials engineering. The occasional interconnections between the same-field groups, along with the absence of connections between the different-field groups, affirm the topic-clustering accuracy.

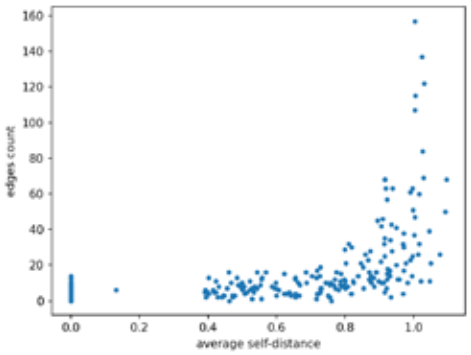


Figure 13. Scatter Plot of Edges Count of Authors in Relation to their Respective Average Self-distance

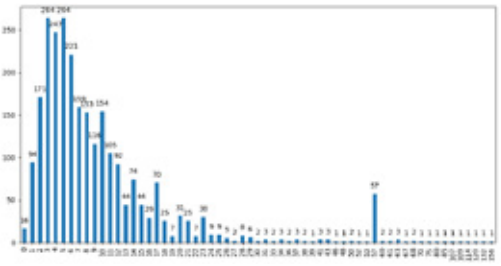

Figure 14. Bar Plot of Edges Count Per Authors (x-axis: number of edges for an author, y-axis: number of authors having n-edges)

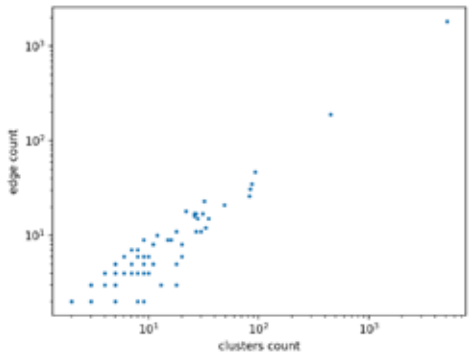


Figure 15. Scatter Plot of Edges Count in Relation to Cluster Count in the formed Social Networks

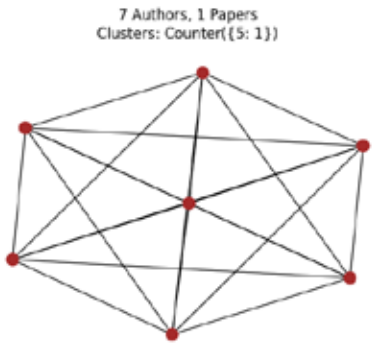


Figure 16. Connected Component For Single Multi-author Paper

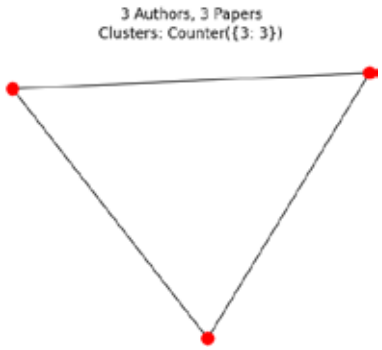


Figure 17. Connected Component For Same Authors Repeatedly

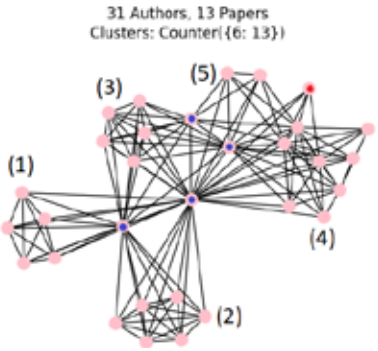


Figure 18. Connected Component For Close Research Network

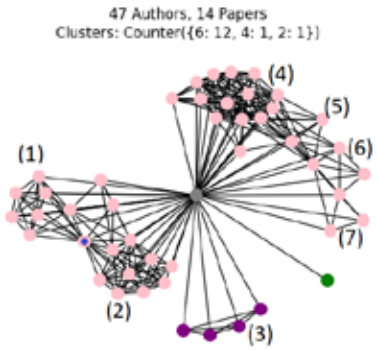


Figure 19. Connected Component For Internal Research Leader

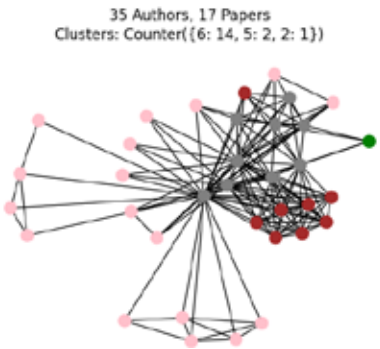


Figure 20. Connected Component For A More Complex Research Network

| clusters | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| paper count | 240 | 115 | 87 | 62 | 252 | 115 | 232 |

Table 5: Cluster-Counter Table of the largest component

*5) Complex research group:* Research groups can take a complex form that is difficult to comprehend. Figure 20 shows an example of such a complex research network. There are 9 authors that have published in different fields (gray nodes). At least 1 out of these gray authors has published in 3 fields: materials, power, and economics. The gray nodes take a central position in the graph, whereas the others are drawn peripherally. These centralized authors are considered to bring research together in a relatively small format on different topics.

*6) The largest component:* Research groups are assumed to be topically isolated within the DIT. However, other than the minority of the isolated networks (such as the ones presented previously), the majority of authors are contained in a large connected component, displayed in Figure 21 (bigger nodes represent the internal authors). This large component is composed of 1,832 authors having 1,103 papers. The cluster counter is recorded in Table V.

In the network, the internal authors are connected closely to their research groups (on both ends). However, they always make a connection with other authors (often internally). This action snowballed to the point of creating such a massive network. The internal authors, having a few topics, collaborate occasionally with each other to form a circle that keeps the research connected within the institute. The collaborations always include at least one multi-clustral author, whose associates are joint-in. These gray authors are the reason for such formation. We deduce the following points from this large structure:

- Research covering different fields can be attractive. Especially with the availability of different yet close fields in the institute of applied sciences. A topic in power engineering can have an industry or material aspect to it. The same applies for other topics in computer science and media engineering. Also, many topics in the health department use technologies developed through research in computer science.
- Other than this, the employment of a generated result in a different topic than its generation-related topic is a valid attempt to extrapolate these results.

Contrary to our expectations regarding the structures of social networks, there exists a set of internal authors whose collaborations with each other form this large network. The ends of this major component represent the research groups of these internal authors. In each of these ends, there exists topic homogeneity. This indicates that the research groups of these internal authors also follow the same topical pattern as the isolated groups. If the collaborations between these internal authors are removed, the large network breaks down into smaller components that follow similar patterns as presented previously, implying an accurate topical clustering of the BERT encodings.

## 7. Conclusion

This article deals with the topical clustering of the scientific papers in the internal publications database of the DIT. The transformer-generated encodings of these papers reflect their corresponding topics. We investigated the topical clustering of such unlabeled data. In our previous work [1], we established a methodology for calculating the cross-distance between a pair of authors based on the respective encodings of their papers. We utilize such a methodology to investigate the topics in the clusters. This previous work focused on the use of Base-BERT and SciBERT and ignored the non-English papers. We reintroduced SciBERT and the centroid-based clustering technique (K-means). We extracted the keywords for each cluster and observed an ambiguity in the keywords of the generated clusters. In parallel, we investigated the non-English papers in our data (German papers) by using both cased and uncased models. We analyzed the minor differences between the two models and extracted their cluster keywords. However, to generate a single landscape for all papers, we employed a multilingual BERT model (mBERT). mBERT was efficient in generating a research landscape that included all papers. Although the texts are

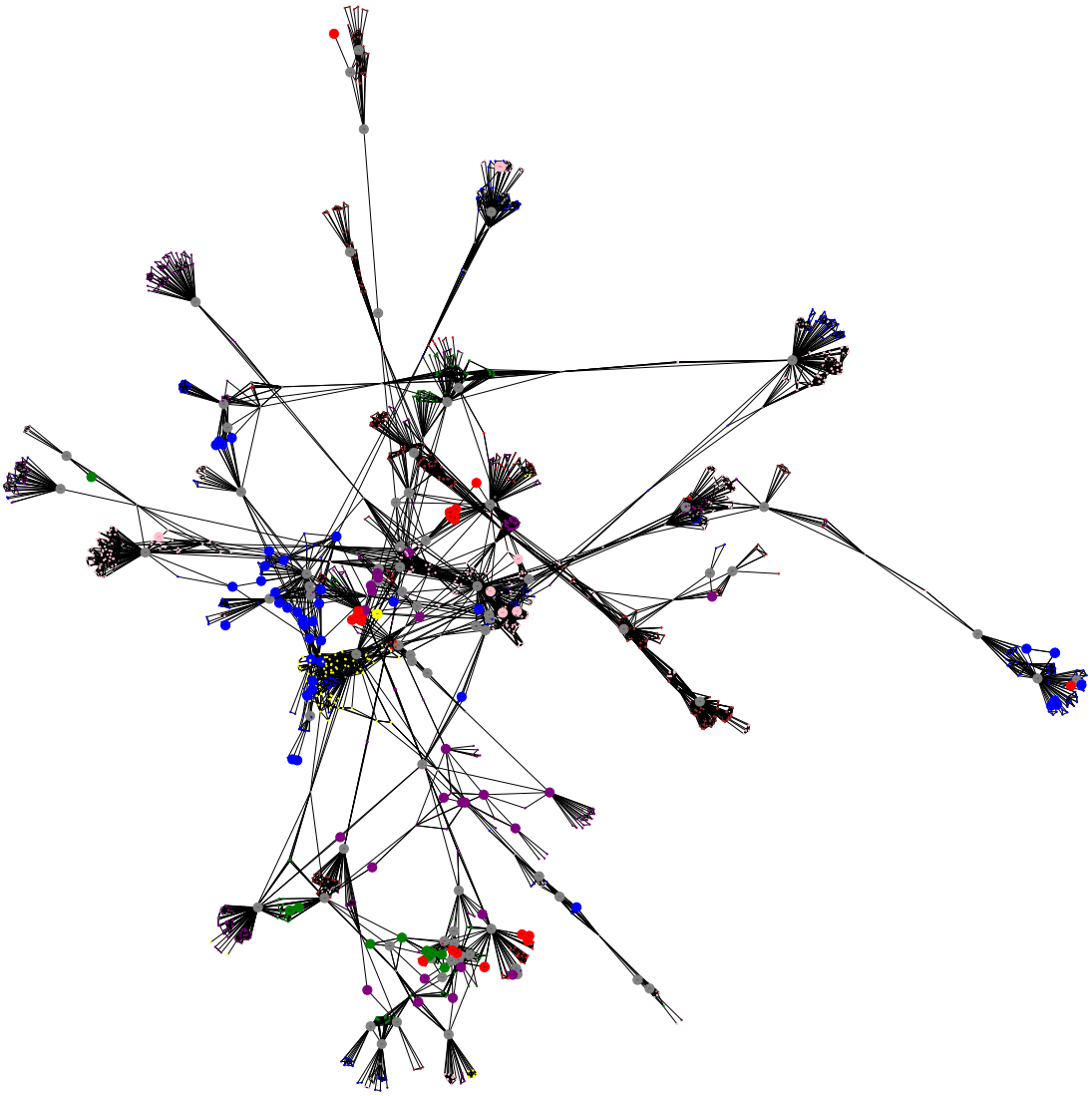

Figure 21. The Largest Connected Component

written in scientific language, mBERT showed acceptable results, as its clusters (and their keywords) matched the research departments of the DIT, despite the fact that mBERT is not trained specifically on scientific data (compared to SciBERT). The keyword-based approach of relating the researchers (the previous approach of relating authors [22]) is more language-dependent than our new transformer-based approach. The model mBERT makes it possible to transcend language barriers. Two papers written in different languages but focusing on similar topics, are encoded into relatively close vectors by mBERT. This ability provides accuracy in the topical clustering of different papers, contextually and regardless of their languages. Our work with mBERT finalizes a major point in the future work section of our previous paper [1]. Due to the absence of labels, the clustering metrics cannot fully affirm the accuracy of the topical clustering. We resorted to keyword extraction and coauthorship analysis, making use of the coauthorship aspect of our textual data. The first part of the analysis involves investigating the uniqueness of the authors in each cluster. Our calculations indicated a high uniqueness percentage of authors in each cluster (over 80%). The second part of the coauthorship analysis is the construction of coauthorship-based social networks. The constructed components contain a large network. This large network holds 74% of internal authors, whose collaborations with each other are key to this large formation. Without these collaborations, the large network decomposes into a set of small components that have a similar structure to the other networks. The construction of coauthorship-based social networks showed

topic homogeneity in the formed components, which represent the research groups at the DIT. Taking all this into account, we conclude that the generated clusters are semantically meaningful.

In our research, we employ pre-trained models without undergoing any fine-tuning. Consequently, the methodology and approach presented here can be applied to analogous publications databases. While the comprehensiveness of the respective outcomes cannot be guaranteed until they are investigated, similar results are to be expected. Still, a limitation of our work is the absence of text for other publications, such as presentations, interviews, or similar. The data used here is limited to research papers only. Although we do not use other types of publication, we fully acknowledge the importance of such contributions. Our decision to focus on only research papers with abstracts stems from the need for the availability of an expressive text for each item, permitting the transformer models to encode it in a comprehensive way.

In future work, we will consider how to incorporate other types of publication into our approach. Moreover, graph neural networks can be employed to predict the missing connections in the co-publication graphs. Enhancing the keyword extraction process leads to an accurate semantic meaning for the clusters formed by K-means in the latent space of publication vectors. The large social network can be used to identify the connecting researchers. We have also performed preliminary experiments with DBSCAN on the HD-vectors that need to be focused in the future.

## Acknowledgement

This paper has received funding from the State of Bavaria in the context of the project SEMIARID, funding no. DIK-2104-0067// DIK0299/01

## References

[1] Z. Bettouche and A. Fischer, “Mapping researcher activity based on publication data by means of transformers,” in Proceedings of the Interdisciplinary Conference on Mechanics, Computers and Electrics (ICMECE 2022), 2022.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” in Advances in Neural Information Processing Systems, 2017, pp. 5998–6008.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[4] J. A. Hartigan and M. A. Wong, “A k-means clustering algorithm,” JSTOR: Applied Statistics, vol. 28, no. 1, pp. 100–108, 1979.

[5] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, “A density-based algorithm for discovering clusters in large spatial databases with noise,” in Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. AAAI Press, 1996, pp. 226–231.

[6] P. J. Rousseeuw, “Silhouettes: a graphical aid to the interpretation and validation of cluster analysis,” Journal of computational and applied mathematics, vol. 20, pp. 53–65, 1987.

[7] T. Calinski and J. Harabasz, “Dendrite method: A non-parametric linkage method for cluster analysis,” Hierarchical methods, vol. 1, pp. 95–113, 1974.

[8] D. L. Davies and D. W. Bouldin, “Cluster separation measure,” IEEE Transactions on Pattern Analysis and Machine Intelligence, no. 2, pp. 224–227, 1979.

[9] P. Molino, E. Giovanelli, K. Kucher, K. De Grave, and N. Moreau, “Keybert: A minimalistic approach for keyword extraction and question answering,” arXiv preprint arXiv:2007.14609, 2020.

[10] L. McInnes, J. Healy, and J. Melville, “Umap: Uniform manifold approximation and projection for dimension reduction,” 2018, comment: Reference implementation available at http://github.com/lmcinnes/umap. [Online]. Available: http://arxiv.org/abs/1802.03426

[11] A. A. Hagberg, P. J. Swart, and D. C. S Chult, "Exploring network structure, dynamics, and function using networkx," Proceedings of the 7th Python in Science Conference (SciPy2008), vol. 11, pp. 11–15, 2008.

[12] Y. Guo, Y. Li, and Y. Zhang, "Unsupervised clustering of transformer embeddings for scientific articles," in Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, 2021, pp. 4022– 4028. [Online]. Available: https://www.aclweb.org/anthology/2021.eaclmain. 315/

[13] I. Beltagy, K. Lo, and A. Cohan, "SciBERT: A pretrained language model for scientific text," in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3615–3620. [Online]. Available: https://aclanthology.org/D19-1371

[14] M. Artetxe, G. Labaka, and E. Agirre, "Unsupervised multilingual representation learning for clustering low-resource languages," in Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018, pp. 1673–1683. [Online]. Available: https://www.aclweb.org/anthology/P18-1158.pdf

[15] M. Ostendorff, T. Ruas, T. Blume, B. Gipp, and G. Rehm, "Aspect-based document similarity for research papers," in Proceedings of the 28th International Conference on Computational Linguistics. Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 6194–6206. [Online]. Available: https://aclanthology.org/2020.coling-main.545

[16] D. Chandrasekaran and V. Mago, "Evolution of semantic similarity—a survey," ACM Comput. Surv., vol. 54, no. 2, feb 2021. [Online]. Available: https://doi.org/10.1145/3440755

[17] K. Kades, J. Sellner, G. Koehler, P. M. Full, T. Y. E. Lai, J. Kleesiek, and K. H. Maier-Hein, "Adapting bidirectional encoder representations from transformers (bert) to assess clinical semantic textual similarity: Algorithm development and validation study," JMIR Med Inform, vol. 9, no. 2, p. e22795, Feb 2021. [Online]. Available: https://medinform.jmir.org/2021/2/e22795

[18] X. Yang, X. He, H. Zhang, Y. Ma, J. Bian, and Y. Wu, "Measurement of semantic textual similarity in clinical texts: Comparison of transformerbased models," JMIR Med Inform, vol. 8, no. 11, p. e19735, Nov 2020. [Online]. Available: http://medinform.jmir.org/2020/11/e19735/

[19] M. Newman, "Co-authorship networks: A review of the literature," Epsrc, UK, 2004, vol. 21, pp. 1–12, 2004.

[20] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A.-L. Barab´asi, "Detecting overlapping and hierarchical community structure in networks," Physical Review E, vol. 69, no. 2, p. 026113, 2004.

[21] B. Bahmani, B. Moseley, A. Vattani, and R. Kumar, "Scalable kmeans++," in Proceedings of the VLDB Endowment, vol. 5, no. 7. VLDB Endowment, 2012, pp. 622–633.

[22] M. Kretschmann, A. Fischer, and B. Elser, "Extracting keywords from publication abstracts for an automated researcher recommendation system," Digitale Welt, vol. 4, pp. 20–25, 01 2020.

## Zineddine Bettouche

Zineddine Bettouche received his Master's degree at the Deggendorf Institute of Technology in the field of Automation Engineering and IT. Since 2022, he has been a research associate at the DIT in the field of artificial intelligence. His current research focuses on utilizing AI, primarily in natural language processing and transformer models, with some interests in computer vision.

*Zineddine Bettouche erhielt seinen Master-Abschluss an der THD im Bereich Automation Engineering und IT. Seit 2022 ist er an der THD als wissenschaftlicher Mitarbeiter im Bereich Künstliche Intelligenz tätig. In seiner derzeitigen Forschung fokussiert er sich auf die Anwendung von Transformer-Modellen im Bereich Natural Language Processing und forscht zusätzlich im Bereich Computer Vision.*

Kontakt / *Contact*
✉ **zineddine.bettouche@th-deg.de**

## Andreas Fischer

Andreas Fischer has been a professor at the Deggendorf Institute of Technology (DIT) since 2017 and its CIO since 2020. Before 2017, he was a postdoc at Karlstad University, Sweden, and a research associate at the University of Passau. He received his diploma (M.Sc.) and PhD from the University of Passau in 2008 and 2017, respectively. He is interested in natural language processing, transformers, and artificial intelligence in general.

*Andreas Fischer ist seit 2017 Professor an der Technischen Hochschule Deggendorf (THD) und seit 2020 Chief Information Officer der THD. Vor 2017 war er Post-Doc an der Karlstad Universität in Schweden und wissenschaftlicher Mitarbeiter an der Universität Passau. Er schloss an der Universität Passau das Diplom 2008 und die Promotion 2017 ab. Er ist an Natural Language Processing, Transformer-Modellen und Künstlicher Intelligenz generell interessiert.*

Kontakt / *Contact*
✉ **andreas.fischer@th-deg.de**