STATISTICS IN TRANSITION new series, March 2026
Vol. 27, No. 1, pp. xxx–xxx, https://doi.org/10.59139/stattrans-2026-xxx
Received – xx.xx.xxxx; accepted – xx.xx.xxxx

Application of BERTopic models in the analysis of Polish research publications in the field of economics and management

Paweł Lula[1]

Abstract

The main objective of the article is to analyze topics from the field of economics and management discussed in the Polish publications from 2000 to 2024. The research process allowed the identification of the main topics and the evaluation of their importance in subsequent years covered by the analysis. The BERTopic model was chosen as the main research method. The paper presents both the theoretical basis of the employed research method and the results of its application to the analysis of the Polish publication achievements registered in the Scopus database. The paper presents a description of topics identified, a specification of the relationship between them and changes in the importance of each topic between 2000 and 2024. All calculations were performed using computer programs prepared in Python language.

Key words: publication achievements, topic modelling, BERTopic method.

1. Introduction

The analysis of issues discussed in Polish publications related to the field of economics and management in the period 2000–2024 was the main goal of the research. Topic modelling, and BERTopic model in particular, was chosen as the main research tool.

Topic modelling belongs to main subareas of the natural language processing. It allows for identification of main issues raised in large collections of documents and for the evaluation of the significance of identified topics. The development of methods of topic modelling and analysis can be observed since the 1990s. A brief overview of the approaches used in this field can be found in Section 2 of the paper. Section 3 presents BERTopic models, while Section 4 discusses methods for assessing topic model's quality. Section 5 presents the results of the analysis of Polish publication achievements in the area of economics and management in the period 2000–2024.

[1] Krakow University of Economics, Poland. E-mail: pawel.lula@uek.krakow.pl. ORCID: https://orcid.org/0000-0003-2057-7299.

2. Topic modelling

Topic modelling allows for the identification and description of main issues discussed in a collection of documents. Having analyzed works on the use of statistical methods for natural language processing, several different approaches to the problem of topic modelling in documents can be identified:

1. Algebraic methods – among which Latent Semantic Analysis (LSA) (Deerwester *et al.*, 1990) is the best known solution. This method is based on frequency matrix representation and allows for the presentation of documents and words in a common base in which dimensions correspond to latent semantic components that can be interpreted as the main issues represented in the corpus. From the computational side, LSA is based on the SVD decomposition of the frequency matrix. Also, non-negative matrix factorization can be used for topics identification (Lee and Seung, 1999).
2. Probabilistic methods – in this approach every topic is described by specifying the distribution over words and every document is represented by the distribution over topics. Latent Dirichlet Allocation (LDA) (Blei, Ng and Jordan, 2003) method is the best known representative of this group of models. LDA can be considered as a generalization of the probabilistic latent semantic analysis (Hofmann, 1999).
3. Transformer-based models. Transformers can be defined as linguistic models that are able to process sequences of tokes (Vaswani *et al.*, 2017). They take into account the semantic aspects of words by using an embedding-based representation. They allow of describing the relationships between words through the use of the attention mechanism. Complex neural networks are used in their construction, where the learning process is carried out based on large corpora of documents. BERTopic technique is one of the most popular approaches belonging to this group of models (Grootendorst, 2022).

3. BERTopic model

BERTopic technique allows for the identification, description and analysis of topics discussed in the collection of documents. This method consists of the following steps:

1. Calculation of sentence embeddings.
2. Dimensionality reduction of embeddings.
3. Identification of topics by clustering of reduced embeddings.
4. Building topic's description.

3.1. Calculation of sentence embeddings

An embedding is a vector representing a given object in the semantic space. The more similar the objects are to each other, the smaller the distance between their embeddings. In natural language processing, embeddings can represent words, sentences, paragraphs or whole documents. Embeddings should present linguistic objects embedded in their context. One of the first researchers to draw attention to the crucial role of context in understanding words was John Rupert Firth (Firth, 1962).

For comprehensive presentation of the process of calculating embeddings of sentences, the architecture of the BERT model should first be presented (Devlin et al., 2019). Taking the transformer architecture as a starting point, it can be concluded that the BERT model performs the functions of the encoder, which determines the numerical representation for tokens comprising the input sequence. BERT is a neural network model which:

- takes as input a sequence of tokens forming two sentences,
- is trained to solve two types of tasks: predicting the missing word in a sentence based on the remaining words, and checking whether two input sentences form a logical sequence,
- uses the attention mechanism to describe the relationships between words forming input sentences,
- is used for the calculation of embeddings - output values of the neural network calculated for a given input word form its embedding.

SBERT (Reimers and Gurevych, 2019) is a version of the BERT model optimized to calculate sentence embeddings. For SBERT model, it is assumed that one sentence is provided as an input and that values of embedding vector for the whole sentence are produced as an output.

3.2. Dimensionality reduction of embeddings

Sentence embeddings calculated with the use of the SBERT model are vectors with several hundred elements. During the current step of the analysis, embeddings are reduced to vectors of a few or a dozen elements. Most often this operation is performed using the UMAP (Uniform Manifold Approximation and Projection for Dimension Reduction) algorithm, which was introduced in (McInnes, Healy and Melville, 2020).

The UMAP algorithm uses weighted graphs to describe the structure of objects in high-dimensional space and the structure of objects' projections in low-dimensional space. The main objective of the method is to determine such a configuration of objects in a low-dimensional space for which the dissimilarity measure between the graphs describing the distribution of objects in each space will be the smallest. In graphs

describing the distribution of objects, edges are created between each node and its n nearest neighbours. The weight assigned to the edge between i-th and j-th vertex defines the probability that a relationship between these vertices exists. All probabilities of link existence between the i-th vertex and the vertices not belonging to its neighborhood are assumed to be zero. The matrices $\mathbf{W}_{n\times n}$ and $\mathbf{V}_{n\times n}$ are the weight matrices of the graphs describing the arrangement of objects in high-dimensional and low-dimensional space. Cross entropy is taken as a measure of the dissimilarity of graphs:

$$H(\mathbf{W},\mathbf{V}) = -\sum_{i,j} w_{ij}\log v_{ij} + \big(1-w_{ij}\big)\log\big(1-v_{ij}\big) \tag{1}$$

Using the UMAP method, the optimization algorithm searches for such a distribution of object projection in low-dimensional space for which $H(\mathbf{W},\mathbf{V})$ takes the smallest value.

To summarize the current section, it may be stated that the UMAP-step transforms sentence embeddings into vectors with several elements, in a way that minimizes the loss of semantic information of sentences.

3.3. Cluster analysis of reduced embeddings

In this step of the analysis sentence embeddings are grouped into clusters with the use of the HDBSCAN method (Hierarchical Density-Based Spatial Clustering of Applications with Noise) (Campello Ricardo J. G. B. and Moulavi, 2013).

In the first step, the HDBSCAN method estimates the probability density function for the analyzed set of points. Next, potential clusters are extracted by finding regions of data space corresponding to every peak of probability density function. The main problem which should be solved is related to the distinction of peaks representing clusters from peaks corresponding to a group of objects forming a part of a larger cluster. This decision is based on the comparison of probability masses of descendant clusters with the probability mass of ancestor cluster reduced by the sum of children masses. If probability masses of descendant clusters are dominant then a current cluster should be split into two new clusters. HDBSCAN splits objects into clusters in a way that maximizes the sum of probability masses of recognized clusters.

The idea presented above is implemented by performing the following steps:

1. Calculation of mutual reachability distances between every pair of embeddings.

   Let us assume that $d(x,y)$ is a distance between x and y points and r_x^n and r_y^n are radii of the smallest circles with centers respectively at point x and y containing n points belonging to the neighborhood of each of these points. Then the mutual reachability distance between x and y point may be defined as:

$$d_{MRD}(x,y) = max\big(d(x,y), r_x^n, r_y^n\big) \tag{2}$$

If points being compared are densely distributed in the space, then $d_{MRD}(x, y)$ is equal to $d(x, y)$. In the case of sparsely distributed points, the $d_{MRD}(x, y)$ is greater than $d(x, y)$.

2. Building a minimum spanning tree (MST).

   Every pair of objects x and y, for witch $d_{MRD}(x, y) > 0$ is linked by an edge to create an undirected graph with $d_{MRD}(x, y)$ as weights. Next, a minimum spanning tree is found with the use of Prim's algorithm (Prim, 1957). This operation is equivalent to dendrogram building with the use of single linkage method and mutual reachability distance.

3. Performing a pruning process.

   Leaves of the MST are combined to form groups containing the required number of objects.

4. Extraction of clusters.

   The main objective of this step is to answer the question whether the probability mass of the descendant clusters is high enough to separate them into separate clusters. Estimation of the probability mass corresponding to a given cluster is performed by analyzing the weights assigned to the edges leading from the node forming a given cluster to the nodes where potential descendant clusters are created.

During this step of analysis sentences are grouped into clusters. Clusters in which the number of elements exceeds the declared threshold value are treated as topics. The remaining sentences are treated as noise.

3.4. Building topic's description

For every extracted topic, its description is built. It has a form of a sequence of words which are crucial to a given topic. A class-based version of the TFIDF schema is used to create topic description (Sparck Jones, 1972). The algorithm is composed of several steps:

1. All sentences assigned to every cluster are merged into separate document,
2. For a set of documents obtained as a result of step 1, a frequency matrix $\mathbf{TF}_{[W\times C]} = [f_{ij}]$, where $i = 1, \dots, W$ indicates a word, and the $j = 1, \dots, C$ represents a cluster, symbol f_{ij} denotes the number of occurrences of the i-th word in the j-th cluster.
3. Weights are calculated with the use of the formula:

$$w_{ij} = f_{ij} \times \log\left(1 + \frac{A}{f_i}\right) \tag{3}$$

where A is an average number of words per class and f_i denotes frequency of the i-th word across all classes.

4. Labels for j-th cluster are created by merging words with highest values of w_{ij}.

4. BERTopic model quality

One of the main methods used to evaluate topic models is coherence measure C_V, which can be defined as the average of consistency coefficients calculated for the n most important words for each topic. The below presentation of C_V is based on (Rijcken, 2023).

The concept of pointwise mutual information (PMI) is a starting point for defining the consistency of words. PMI is a measure of association between two events x and y and can be defined as:

$$\mathrm{pmi}(x;y) = \log\frac{P(x;y)}{P(x)P(y)} \tag{4}$$

PMI compares the probability of the simultaneous occurrence of two events with the probability of their simultaneous occurrence when they are independent. The PMI value can be normalized using the formula:

$$\mathrm{npmi}(x;y) = \frac{\mathrm{pmi}(x;y)}{-\log\big(P(x;y)\big)} \tag{5}$$

where npmi is a normalized (to $[-1;1]$ range) pointwise information and $-\log\big(p(x;y)\big)$ is a self-information (Shannon information) related to the message about simultaneous occurrence of x and y.

Assuming that:

- $\boldsymbol{D} = \{d_1, d_2, \dots, d_{|\boldsymbol{D}|}\}$ is a set of documents,
- $\boldsymbol{d_i} = [w_{i,1}^d, w_{i,2}^d, \dots, w_{i,|d_i|}^d]$ defines the i-th document as a list of words,
- $S(d_i, j, \sigma)$ is a sliding window defined for the i-th document, starting at the j-th position and including σ words,
- $T = \{t_1, t_2, \dots, t_{|T|}\}$ is a set of identified topics,
- $V_k = [v_{k,1}, v_{k,2}, \dots, v_{k,N}]$ defines a list of N words defining the k-th topic.

Next, the matrix of association coefficients between words defining every topic should be created. For topic k the matrix $\mathbf{Q}_k$ has a form:

$$\mathbf{Q}_k = \begin{bmatrix} q_{1,1}^k & \dots & q_{1,N}^k \\ \dots & \dots & \dots \\ q_{N,1}^k & \dots & q_{N,N}^k \end{bmatrix} \tag{6}$$

where:

$$q_{x,y}^k = nmpi\big(v_{k,x}, v_{k,y}\big) \tag{7}$$

For topic k, normalized PMI values are calculated using probabilities of occurrence of words $v_{k,x}$ and $v_{k,y}$ inside the sliding window $S(d_i, j, \sigma)$ moving through all documents in $\boldsymbol{D}$. It may be expressed as:

$$nmpi\left(v_{k,x}, v_{k,y}\right) = \frac{\log\frac{P\left(v_{k,x}, v_{k,y}\right)+\varepsilon}{P\left(v_{k,x}\right)P\left(v_{k,y}\right)}}{-\log\left(P\left(v_{k,x}, v_{k,y}\right)+\varepsilon\right)} \tag{8}$$

where $P\left(v_{k,x}, v_{k,y}\right)$ is defined as:

$$P\left(v_{k,x}, v_{k,y}\right) = \frac{\sum_{a=1}^{|D|}\sum_{b=1}^{|d_a|-\sigma+1} g\left(a,b,\sigma,v_{k,x},v_{k,y}\right)}{\sum_{a=1}^{|D|}\left(|d_a|-\sigma+1\right)} \tag{9}$$

where:

$$g\left(a, b, \sigma, v_{k,x}, v_{k,y}\right) = \begin{cases} 1 & if\ v_{k,x}, v_{k,y} \in S(d_a, b, \sigma) \\ 0 & otherwise \end{cases} \tag{10}$$

and $P\left(v_{k,x}\right)$ is calculated using the formula:

$$P\left(v_{k,x}\right) = \frac{\sum_{a=1}^{|D|}\sum_{b=1}^{|d_a|-\sigma+1} h\left(a,b,\sigma,v_{k,x}\right)}{\sum_{a=1}^{|D|}\left(|d_a|-\sigma+1\right)} \tag{11}$$

where:

$$h\left(a, b, \sigma, v_{k,x}\right) = \begin{cases} 1 & if\ v_{k,x} \in S(d_a, b, \sigma) \\ 0 & otherwise \end{cases} \tag{12}$$

Next, for every topic, vector $\boldsymbol{M}_k$ is calculated:

$$\boldsymbol{M}_k = \left[m_{k,1}, m_{k,2}, \ldots, m_{k,N}\right] = \left[\sum_{j=1}^{N} q_{j,1}^{k}, \sum_{j=1}^{N} q_{j,2}^{k}, \ldots, \sum_{j=1}^{N} q_{j,N}^{k}\right] \tag{13}$$

Elements of $\boldsymbol{M}_k$ are calculated as sums of values located in subsequent columns of $\mathbf{Q}_k$. $\boldsymbol{M}_k$ may be treated as a k-th topic representation.

To calculate the coherence measure for a given set of topic, the $\boldsymbol{C}$ matrix is first calculated.

$$\boldsymbol{C}_{[|T|\times N]} = \begin{bmatrix} sim(\boldsymbol{M}_1, \boldsymbol{q}_1^1) & \ldots & sim(\boldsymbol{M}_1, \boldsymbol{q}_N^1) \\ \ldots & \ldots & \ldots \\ sim\left(\boldsymbol{M}_{|T|}, \boldsymbol{q}_1^{|T|}\right) & \ldots & sim\left(\boldsymbol{M}_{|T|}, \boldsymbol{q}_N^{|T|}\right) \end{bmatrix} \tag{14}$$

where symbols $\boldsymbol{q}_j^k$ represent the j-th row of the $\mathbf{Q}_k$ matrix and $sim(.)$ is a cosine similarity between vectors.

Finally, the coherence measure C_V is calculated as an arithmetic average of elements of the $\boldsymbol{C}$ matrix.

Values of the C_V coefficient belong to the range $[0; 1]$. Higher values indicate higher consistency of topics. When deciding on the number of topics, it is advisable to maximize this indicator.

5. The analysis of Polish research publications in the fields of economics and management

5.1. The scope of the analysis

The dataset included titles and abstracts of research publications published in the period 2000–2024, with at least one Polish author, registered by the Scopus database and assigned to *BUSI* (business), *ECON* (economics) or *DECI* (decision science) areas. The total number of publications which met the above conditions was 36445, but the analysis covered 35626 publications that had a title and an abstract in English.

Basic quantitative indicators describing the whole set of Polish publication achievements (36445 works) are presented in Figure 1.

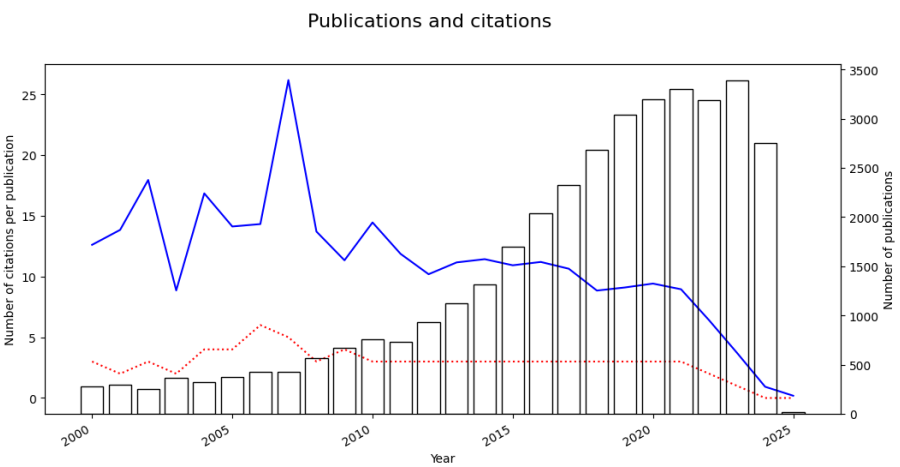

Figure 1. Number of publications (bar plot, right axis), number of citations per publication (average value – blue solid line, left axis; median value – red dotted line, left axis)

Source: own work based on Scopus database.

A rapidly increasing number of publications can be seen by 2021. The number of published papers seems to stabilize in the following years. In contrast, the number of citations per published paper has been decreasing over the past 15 years.

5.2. BERTopic model building and interpretation

The BERTopic model was used for the analysis of titles and abstracts of Polish publications. First, documents were split into tokens with a form of sentences. All tokens with 28 or less letters were removed (these tokens most often contained names of publishing houses or names of affiliated institutions). Finally, in the analysis 283576 sentences were used.

Next, embeddings for sentences were calculated with the use of the SBERT model.

Several BERTopic models were tested and finally the model with 9 topics was chosen. This decision was taken on the basis of the C_V coherence, which, depending on the number of topics, took values shown in Table 1.

Table 1. Values of the C_V coherence for models with different number of topics

Number of topics	C_V
6	0.4228
7	0.4338
8	0.4438
9	0.4823
10	0.4643

Source: own work.

Table 2 shows the number of tokens (sentences) assigned to every topic.

Table 2. Number of sentences assigned to every topic

Topic ID	***Topic -1***	***Topic 0***	***Topic 1***	***Topic 2***	***Topic 3***
Count	133358	108453	18851	13650	3351
Topic ID	***Topic 4***	***Topic 5***	***Topic 6***	***Topic 7***	
Count	2432	1689	1656	136	

Source: own work.

Topic -1 represents all sentences which have been identified as noise and are not related to any of the recognized topics.

In order to interpret each topic, lists of the words most closely related to each topic have been created (Figure 2).

Figure 2. The most important words for identified topics

Source: own work.

Topic 0 covers issues related to economic development and management methods and their implementation in Poland. The key issue addressed under *Topic 1* is decision support methods. *Topic 2* represents issues specific to commodity science. Issues specific to natural environment and regional development are discussed within *Topic 3*. Subjects related to image processing are discussed under *Topic 4*. Health care issues are related to *Topic 5*. Biology and genetics issues are related to *Topic 6*. *Topics 7* is related to mathematics, in particular to geometry.

Issues specific to identified topics are in many cases related to each other. A visualization of the similarity matrix is shown in Figure 3.

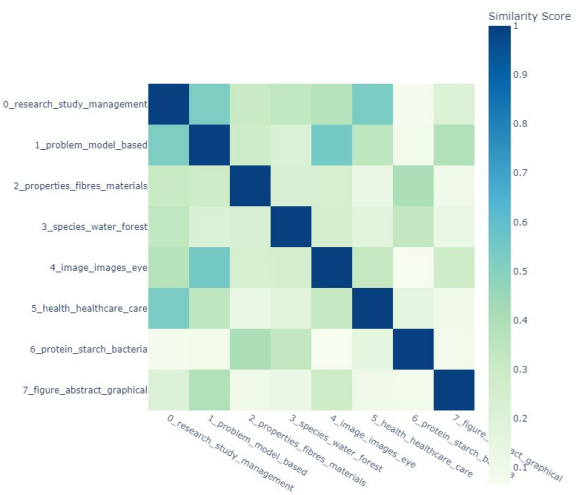

Figure 3. Visualization of the similarity matrix between topics

Source: own work.

A useful tool for analyzing relationships between topics can also be an intertopic distance map presented in Figure 4.

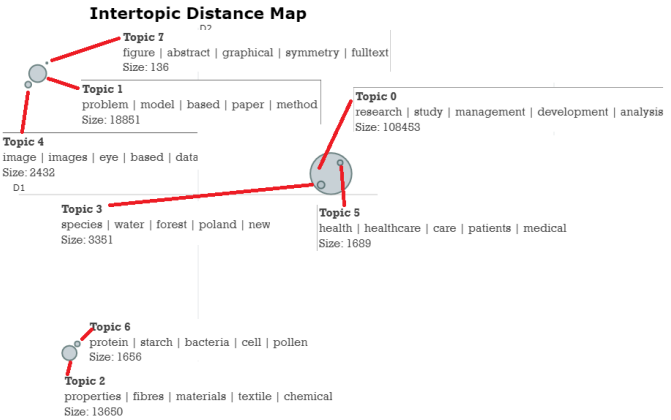

Figure 4. Intertopic distance map for identified topics

Source: own work.

An analysis of Figure 4 indicates that three groups of themes can be identified:

- Group 1: Topic 1, Topic 4, Topic 7.
- Group 2: Topic 0, Topic 3, Topic 5.
- Group 3: Topic 2, Topic 6.

In the next step of the research, a sentence-topic matrix was estimated to determine the importance of every topic in every single sentence. Next, information about topic contribution to every sentence was aggregated at the level of every document. The aggregation was done by calculating geometric average for values relating to sentences that formed a given document. Then, using the same approach, an aggregation of the importance of each topic was carried out for each year included in the scope of analysis. The results are presented in Figure 5.

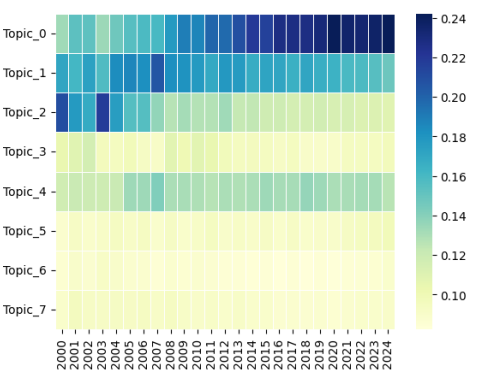

Figure 5. Changes in topics' importance in Polish research works over years

Source: own work.

Analyzing the data presented in Figure 5, it is worth noting that the values presenting the importance of each topic in consecutive years are relative (they add up to unity for each year). At the beginning of the current century, the greatest publication achievements were related to decision support systems and commodity science. Since the second decade of the 21st century, economic development and management issues have played a key role in the publication output. In contrast, the importance of commodity science and decision support systems has been declining. Interest in image processing methods, which fall under the umbrella of multivariate analysis, is also noticeable. The importance of the other topics was rather low.

6. Conclusions

The research carried out allows for formulation of the following conclusions:

1. In quantitative terms, the Polish publication achievements in the field of economics and management has increased significantly since the beginning of this century, although the number of publications stabilized in the last few years. The growth potential seems to be exhausted.
2. The quality of the analyzed publication achievements, measured by the number of citations, has not shown any positive change for the last 15 years.
3. Main topics discussed in Polish publications included: economic development, management methods, decision support solutions, commodity science issues, natural environment and regional development, health care system, biology and genetics and mathematics.
4. Topics related to economic development and management issues gained the most importance in the last two decades.
5. Decision support systems and commodity science issues have lost their relevance.
6. The importance of the quantitative approach remains noticeable and unchanged.
7. The remaining topics have relatively small significance.
8. The use of the BERTopic model has made it possible to analyze large text datasets and aggregate the results.
9. Further research on BERTopic and other topic modelling methods should be considered as necessary.

References

Blei, D., Ng, A. and Jordan, M., (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, (3), pp. 993–1022.

Campello Ricardo J. G. B. and Moulavi, D. and S. J., (2013). Density-Based Clustering Based on Hierarchical Density Estimates, in V.S. and C.L. and M.H. and X.G. Pei Jian and Tseng (ed.) *Advances in Knowledge Discovery and Data Mining*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 160–172.

Deerwester, S. *et al.*, (1990). Indexing by latent semantic analysis. *Journal of the Society for Information Science*, 41(6), pp. 391–407.

Devlin, J. *et al.*, (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Available at: https://arxiv.org/abs/1810.04805.

Firth, J. R., (1962). A synopsis of linguistic theory, 1930–1955, in *Studies in Linguistic Analysis*. Oxford: Blackwell.

Grootendorst, M., (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794* [Preprint].

Hofmann, T., (1999). *Probabilistic Latent Semantic Indexing*. New York: ACM.

Lee, D. D., Seung, H. S., (1999). Learning the parts of objects by nonnegative matrix factorization. *Nature*, 401, pp. 788–791.

McInnes, L., Healy, J. and Melville, J., (2020). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. Available at: https://arxiv.org/abs/1802.03426.

Prim, R. C., (1957). Shortest connection networks and some generalizations. *The Bell System Technical Journal*, 36(6), pp. 1389–1401. Available at: https://doi.org/10.1002/j.1538-7305.1957.tb01515.x.

Reimers, N. and Gurevych, I., (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. Available at: https://arxiv.org/abs/1908.10084.

Rijcken, E., (2023). *CV Topic Coherence Explained. Understanding the metric that correlates the highest with humans*.

Sparck Jones, K., (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1), pp. 11–21.

Vaswani, A. *et al.*, (2017). Attention is all you need, in *Advances in Neural Information Processing Systems*.