Transnational University Limburg is a unique collaboration of two universities in two countries: the University of Hasselt and Maastricht University.

UHASSELT
KNOWLEDGE IN ACTION

www.uhasselt.be
Universiteit Hasselt
Campus Hasselt:
Martelarenlaan 42 | 3500 Hasselt
Campus Diepenbeek:
Agoralaan Gebouw D | 3590 Diepenbeek

UHASSELT
KNOWLEDGE IN ACTION

Maastricht University

Faculty of Sciences
School for Information Technology

Master of Statistics and Data Science

Master's thesis

Metadata Enrichment and Exchange in Research Information Systems

Hasan Görkem Uyanik

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics and Data Science, specialization Data Science

SUPERVISOR :

dr. Amr ALI ELDIN

2023
2024

Maastricht University

Faculty of Sciences
School for Information Technology
Master of Statistics and Data Science

Master's thesis

Metadata Enrichment and Exchange in Research Information Systems

Hasan Görkem Uyanik

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics and Data Science, specialization Data Science

SUPERVISOR :

dr. Amr ALI ELDIN

Abstract

Research Information Systems (RIS) play a crucial role in managing and sharing scientific research data. A common problem is the missing linkages between publications and the projects they belong to. These missing connections limit the ability to track research impact. This study aims to improve techniques used to link projects and publications on the Flanders Research Information Space (FRIS) portal. Two datasets, from FRIS and Dimensions, are analyzed. Labeled Latent Dirichlet Allocation (L-LDA) and BERTopic are used to obtain probability matrices of projects and publications. Then, cosine similarity and Jensen-Shannon divergence (JSD) are used as distance metrics to construct distance matrices from the probability matrices, which along with other engineered features, are input to a logistic Gradient Boosting (GB) model to predict publication-project links. The GB model gives significantly higher accuracies than Support Vector Machine (SVM) or distance matrices alone. However, further analysis revealed the high FRIS accuracies are mainly due to a 'common organization ratio' variable. Excluding this variable dramatically reduces accuracy. The methodology significantly improved FRIS results, while Dimensions accuracies are lower due to lack of certain variables. Using ChatGPT for topic modelling could be explored in future work. The study provides an effective approach for enhancing data quality and research impact assessment in RIS.

Key Words: FRIS, Dimensions, L-LDA, BERTopic, Gradient Boosting, Support Vector Machine, topic modelling, distance matrix, cosine similarity, Jensen-Shannon divergence

---

1 Introduction

Research Information Systems (RIS) play a crucial role in managing and disseminating scientific research data [1]. These systems store and organize information about research projects, publications, researchers, and institutions. However, a common challenge faced by many RIS is the lack of comprehensive linkages between research outputs (such as publications) and the projects that they belong to. This missing connection limits the ability to track research impact and understand the full scope of project outcomes.

The Flanders Research Information Space[1] (FRIS) is an example of a regional RIS that aims to provide a comprehensive view of research activities

[1] `https://researchportal.be/en/about-fris`

in Flanders [2]. While FRIS contains rich metadata about projects and publications, establishing accurate links between these entities remains an ongoing challenge. Improving these connections is vital for enhancing data quality and enabling more effective research assessment and policy-making.

Artificial intelligence (AI) and machine learning techniques offer promising solutions for automating and improving the process of linking research projects to their associated publications [3]. By leveraging natural language processing, topic modeling, and classification algorithms, it may be possible to identify likely connections between projects and publications even when explicit links are not recorded.

To ensure consistency in our analysis, we have carefully extracted relevant features from the FRIS data, including project and publication metadata such as titles, abstracts, discipline codes, authors, and organizations. By developing accurate prediction models, we hope to contribute to improved data quality and more comprehensive research information management in RIS.

Accuracy and completeness of research metadata is of the utmost importance for scientific research. Although the aforementioned data on FRIS portal for the projects is complete and attainable for most of the projects as long as they are public, a significant portion of the publications are missing either abstracts, or discipline codes, or author names, or any combination of them. Additionally, each publication is linked to a project in reality but this information is not on FRIS for some of the projects and publications, and it is crucial for the data quality of FRIS.

The main objective in this study is to improve existing techniques [2] used in informetrics and scientometrics such as L-LDA (Labeled Latent Dirichlet Allocation) [4] and BERTopic (Neural topic modeling with a class-based TF-IDF procedure) [5] for topic modelling, and SVM (Support Vector Machine) [6] and GB (Gradient Boosting) [7] for classification by implementing GB in conjunction with L-LDA and BERTopic to match given unlinked publications to the correct projects. Labeled data for projects and publications are obtained with FRIS's SOAP[2] (Simple Object Access Protocol) API service [2]. L-LDA and BERTopic models are implemented to obtain the two probability distribution matrices for both projects and publications where each row corresponds to a project/publication and sum up to 1, and columns correspond to research disciplines. To construct a distance matrix where rows correspond to projects and columns correspond to publications, cosine similarity [2], [8], [9] and JSD (Jensen-Shannon divergence) [9] are used for

[2] `https://frisr4.researchportal.be/ws`

the probability vectors from the probability distribution matrices. Cosine similarity is one of the two best-known similarity measures [10]. Finally, a logistic GB model is implemented to predict the probabilities of each publication belonging to each project with date criterion, common organisation ratio, common author ratio, and distance as predictors. In addition to a sample of labeled FRIS data, a sample of labeled Dimensions data is analysed for comparison.

2 Literature Review

2.1 Research Motivation

The growing volume of scientific publications and research projects has made it increasingly challenging to maintain accurate and comprehensive linkages in Research Information Systems. Several studies have explored the use of machine learning techniques to address related issues in research information management:

- Altınel and Ganiz [11] provide a comprehensive survey of text classification methods, including Support Vector Machines (SVM), applied to research document classification. While not specifically addressing project-publication linking, their work demonstrates the potential of these techniques for organizing and categorizing research information.
- Jeong et al. [12] applied Latent Dirichlet Allocation (LDA) to analyze research topics and author relationships. Their method of content-based author co-citation analysis shows promise for understanding connections between different research entities, which could be extended to project-publication linking.

These prior works demonstrate the potential of AI techniques in the domain of research information management, but there remains room for improvement in accuracy and scalability, particularly for the specific task of linking projects and publications.

2.2 Research Information Systems

Research Information Systems (RIS) serve as centralized platforms for managing research-related information. As described by Castro and Puuska [1], key features of RIS typically include:

- Project databases with funding, personnel, and outcome tracking

- Publication repositories
- Researcher profiles
- Institutional hierarchies
- Analytics and reporting tools

RIS play a crucial role in supporting the entire research lifecycle, from proposal writing to result dissemination, and in facilitating evidence-based decision making for research management [1].

2.3 FRIS and Dimensions

FRIS is a regional web portal, governed by the Flemish government [2]. FRIS contains metadata of research projects such that there is the project abstract, a summary of the scientific disciplines in which the project is situated, the authors (researchers) and the organisations involved, the start and end dates, and the funding of the project. FRIS also contains metadata of research publications which is similar to metadata of research projects.

To classify projects and publications by research disciplines, FRIS uses the Flemish Research Discipline Standard[3] ("Vlaamse Onderzoeksdiscipline Standaard", abbreviated as VODS, in Dutch) [2], which is described in [13]. The VODS has four hierarchical levels that correspond to different levels of granularity of research disciplines, with 7, 42, 382, and 2493 disciplines at each level with increasing granularity [2]. Pham et al. [2] describes the first two levels as follows: "The first level corresponds to the OECD FORD [14] classification's six scientific fields (natural sciences (01), engineering and technology (02), medical and health sciences (03), agricultural and veterinary sciences (04), social sciences (05), and humanities and arts (06), expanded with one extra discipline to label administrative and technical research personnel (general and logistic services (07)). The second level contains the major disciplinary subjects (for example, mathematical sciences (0101), information and computing sciences (0102), physical sciences (0103), and so on), while the third and fourth levels correspond to more granular subfields." In this study, we are interested in the second level of VODS classification.

Dimensions[4] provides a large collection of linked research data including grants, publications, datasets, clinical trials, patents, and policy documents. Dimensions uses ANZSRC[5] (Australian and New Zealand Standard Research

[3] https://researchportal.be/en/disciplines

[4] https://www.dimensions.ai

[5] https://www.abs.gov.au/statistics/classifications/australian-and-new-zealand-standard-research-classification-anzsrc/2020

Classification) [15] which has three hierarchical levels that correspond to increasing levels of granularity of research disciplines for division, group, and field, with 23, 213, and 1967 disciplines at each level respectively. An example of three disciplines on three different levels: biological sciences (31), ecology (3103), behavioural ecology (310302). In this study, we are interested in the second level of ANZSRC classification.

2.4 NLP and Topic Modeling in Research Information Management

Natural Language Processing (NLP) and topic modeling techniques have been increasingly applied to analyze and organize research information:

- Thijs et al. [16] used noun phrase extraction in combination with hybrid clustering to improve the analysis of research topics in information system research. Their approach demonstrates the potential of NLP techniques for extracting meaningful information from research texts.
- Beltagy et al. [17] introduced SciBERT, a BERT-based model pretrained on scientific text. While not directly applied to project-publication linking, this work shows the potential of advanced language models for understanding and classifying scientific content.
- He et al. [18] applied dynamic topic modeling to track the evolution of research themes in scientific literature, using citation information to enhance the model. This approach could be adapted to understand the relationships between projects and their resulting publications over time.

These applications highlight the potential for advanced NLP and topic modeling methods to extract meaningful insights from research metadata and improve the organization of research information. By building upon these existing works and adapting them to the specific challenge of project-publication linking, we aim to develop more accurate and efficient methods for enhancing the connectivity and usability of Research Information Systems.

3 Data Collection and Preparation

3.1 FRIS

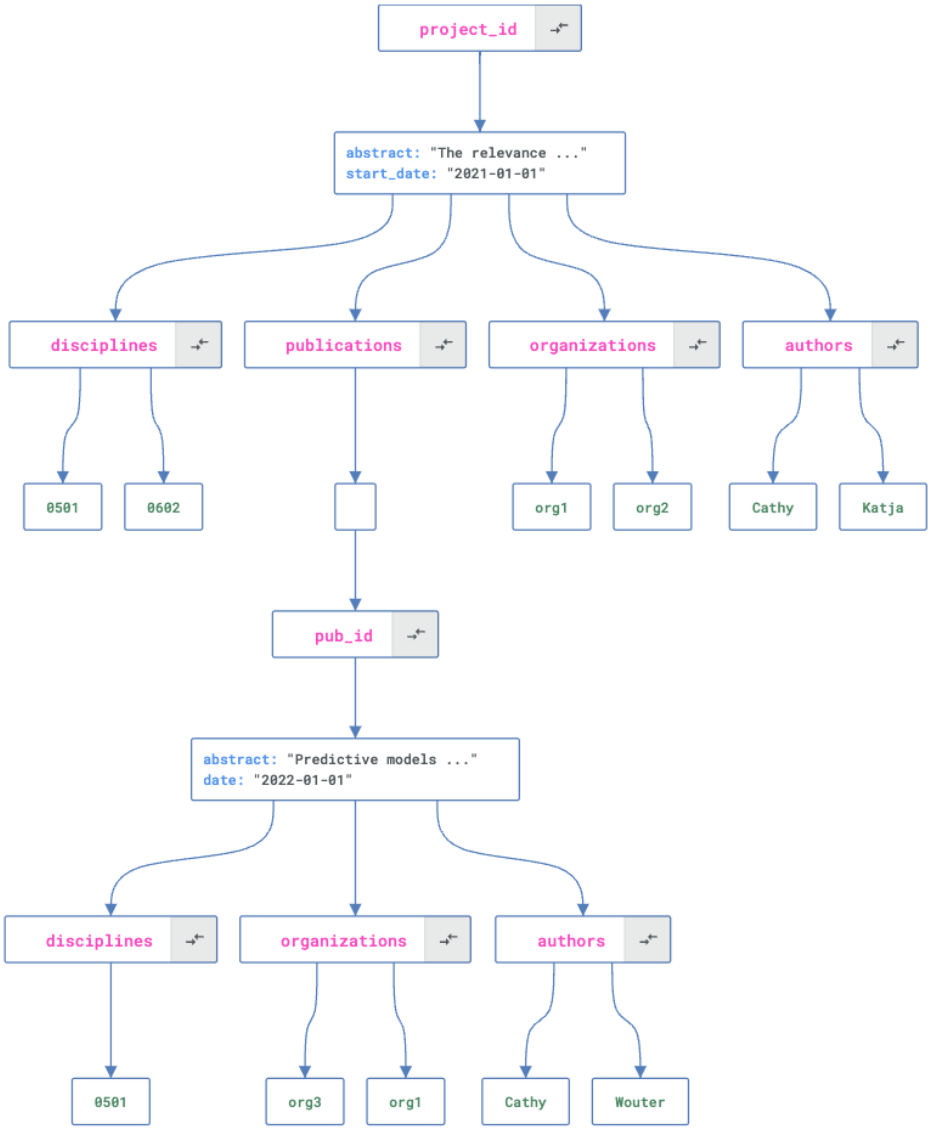

Figure 1: The schematic diagram of the data structure of FRIS dataset

Data is collected through SOAP API service of FRIS [2]. First, from a list of 703 project IDs, corresponding publication IDs are obtained. Then, with all the IDs for projects and publications, the dataset that schematically represented in Figure 1 created. However, there were some caveats worth mentioning during the data retrieval process.

For some projects, it is possible to get the related publications but not the abstract for the particular project because of the restriction of authorisation (Examples: #58 and #68 in the list of project IDs with the titles "H IMPACT A dedicated postdoctoral research and training programme fostering impact development and entrepreneurship" and "SoMe Dem Social media for democracy understanding the causal mechanisms of digital citizenship").

Some projects have no publications which is, of course, somewhat expected and most publications don't have abstracts. But one of the projects was particularly interesting (#8 in the list of project IDs with the title "The H boson gateway to physics beyond the Standard Model"). There is no authorisation problem and the project has 37 publications. However, when a request is sent, sometimes after as much as 20 seconds, the server returns an error message and no publication ID can be received. So, it is concluded that there is a problem in the FRIS server and the personnel in charge of the server should be informed about it. A single project giving an error might not seem like a big problem but without waiting after the error message for at least 5 seconds, it creates a cascading effect and publication IDs cannot be received for the following healthy requests - sometimes for tens of projects. Consequently, waiting 5 seconds after each request takes a long time for all the project IDs.

So, taking all those into account, the dataset is created with IDs, titles, abstracts, keywords, discipline codes, organisation IDs, and author names for the projects and publications. Starting date for projects and publishing date for publications are also retrieved. Projects and publications without abstracts are discarded. Most of the projects have discipline codes and the ones without discipline codes are not included in the final dataset. Organisation IDs and relevant information of dates are present for all the projects and publications. While having no author names is allowed for the projects, it is not permitted for the publications.

Since both cosine similarity and JSD require equal-sized vectors from the same probability distribution to calculate distances [9] it is crucial that the union of unique discipline codes in all of the projects is equal to the union of unique discipline codes in all of the publications. Furthermore, to get equal-sized vectors from the same probability distribution as a result of L-LDA model training and inference it is also crucial that the union of unique discipline codes in all of the projects in the training dataset is equal to the union of unique discipline codes in all of the publications in the training dataset. Even though the unique discipline codes in our training sets don't add up to 41, since we know that those 41 discipline codes are all the codes on the second hierarchical level of VODS with the exclusion of "General

and logistic services" because it is not an actual research discipline [19], we used all of them as labels in the training process. Fortunately, this naturally ensures that the probability vectors are from the same probability distribution. Therefore, a training dataset with 129 projects and 423 publications and a test dataset with 33 projects and 163 publications are obtained. The training and test datasets correspond to 80% and 20% of the projects in the whole dataset respectively.

3.2 Dimensions

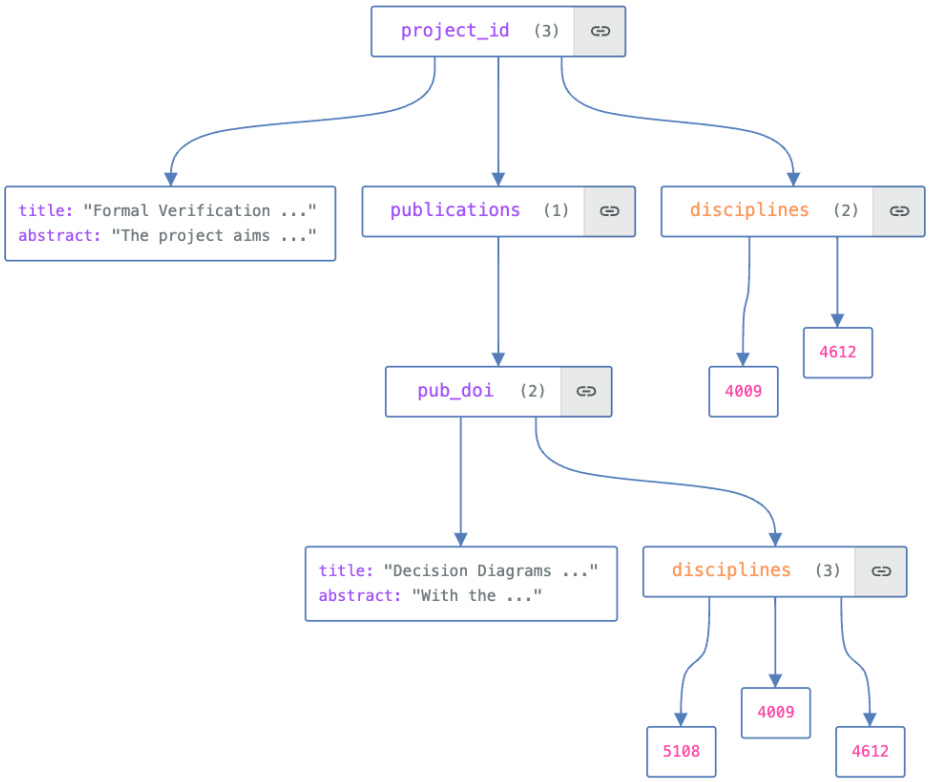

Figure 2: The schematic diagram of the data structure of Dimensions dataset

The data for projects and publications are provided in separate datasets which include IDs, titles, abstracts, and discipline codes. The data is united into a single dataset which is schematically represented in Figure 2. A basic data cleaning is applied to make sure all the data types are correct and the resulting data has 9875 projects and 29181 publications. The dataset has 170 discipline codes out of 213 on the second hierarchical level of ANZSRC. A sample with 225 projects and 635 publications including the same discipline

codes as the original data is created because the number of projects and publications in the original dataset would make the computations extremely heavy. A training dataset with 180 projects and 506 publications and a test dataset with 45 projects and 129 publications are obtained. The training and test datasets correspond to 80% and 20% of the projects in the whole dataset respectively.

4 Methodology

The proposed methodology combines supervised and unsupervised topic modeling techniques with machine learning classifiers to predict links between research projects and publications. The key steps after the data collection and preparation in our approach are presented in this section.

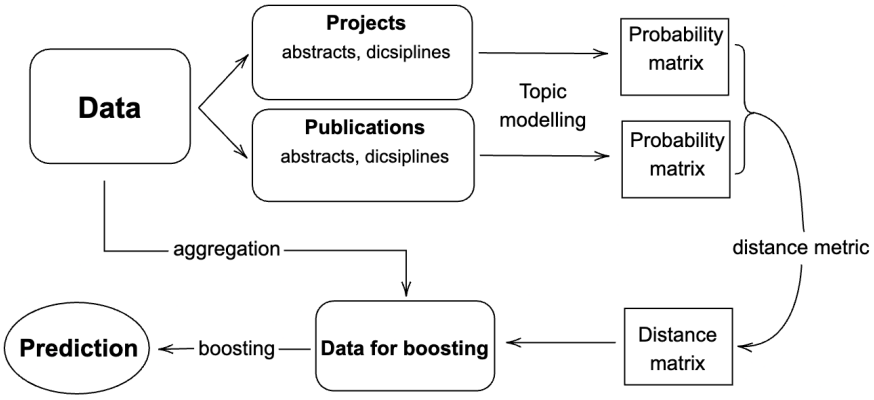

Figure 3: Data processing and modelling

4.1 Topic Modeling

We implement and compare two topic modeling approaches which are L-LDA and BERTopic [4], [5].

4.1.1 Labeled Latent Dirichlet Allocation (L-LDA)

L-LDA is a supervised extension of LDA that incorporates predefined labels (in our case, discipline codes) [4]. The generative process for L-LDA is as follows:

1. For each topic $k \in \{1, \ldots, K\}$:
   (a) Generate $\boldsymbol{\beta}_k = (\beta_{k,1}, \ldots, \beta_{k,V})^T \sim \text{Dir}(\cdot|\boldsymbol{\eta})$
2. For each document d:
   (a) For each topic $k \in \{1, \ldots, K\}$
      i. Generate $\boldsymbol{\Lambda}_k^{(d)} \in \{0, 1\} \sim \text{Bernoulli}(\cdot|\boldsymbol{\Phi}_k)$
   (b) Generate $\boldsymbol{\alpha}^{(d)} = L^{(d)} \times \boldsymbol{\alpha}$
   (c) Generate $\boldsymbol{\theta}^{(d)} = (\theta_{l1}, \ldots, \theta_{lM_d})^T \sim \text{Dir}(\cdot|\boldsymbol{\alpha}^{(d)})$
   (d) For each $i \in \{1, \ldots, N_d\}$:
      i. Generate $z_i \in \{\lambda_1^{(d)}, \ldots, \lambda_{M_d}^{(d)}\} \sim \text{Mult}(\cdot|\boldsymbol{\theta}^{(d)})$
      ii. Generate $w_i \in \{1, \ldots, V\} \sim \text{Mult}(\cdot|\boldsymbol{\beta}_{z_i})$

where $\boldsymbol{\beta}_k$ is a vector consisting of the parameters of the multinomial distribution corresponding to the kth topic, $\boldsymbol{\alpha}$ are the parameters of the Dirichlet topic prior and $\boldsymbol{\eta}$ are the parameters of the word prior, while $\boldsymbol{\Phi}_k$ is the label prior for topic k [4]. The vector of document's labels is defined as $\boldsymbol{\lambda}^{(d)} = \{k|\boldsymbol{\Lambda}_k^{(d)} = 1\}$ which provides the definition of a document-specific label projection matrix $L^{(d)}$ of size $M_d \times K$ for each document d, where $M_d = |\boldsymbol{\lambda}^{(d)}|$, as follows: For each row $i \in \{1, \ldots, M_d\}$ and column $j \in \{1, \ldots, K\}$:

$$L_{ij}^{(d)} = \begin{cases} 1 & \text{if } \lambda_i^{(d)} = j \\ 0 & \text{otherwise.} \end{cases}$$

4.1.2 BERTopic

BERTopic is an unsupervised topic modeling technique that leverages BERT embeddings and a class-based TF-IDF procedure [5]. It allows for modeling of sequentially-organized documents and can capture topic evolution over time.

4.2 Distance Calculation

We compute distances between project and publication topic distributions using cosine similarity and Jensen-Shannon divergence.

4.2.1 Cosine Similarity

When a text document is represented as a vector, the similarity of two documents can be obtained through computing cosine value between the two vectors [9]. The formula of cosine similarity is

$$cos(P,Q) = \frac{P \times Q}{|P| \times |Q|} = \frac{\sum_{i=1}^{n} Freq(w_i|P) \times Freq(w_i|Q)}{\sqrt{\sum_{i=1}^{n} Freq(w_i|P)^2} \times \sqrt{\sum_{i=1}^{n} Freq(w_i|Q)^2}} \quad (1)$$

where P and Q refer to two different documents' vectors, and the components of the two vectors are frequencies of a certain word in the document $Freq(w_i|P)$ and $Freq(w_i|Q)$ [9]. The distance between the vectors is given by $1 - cos(P,Q)$ [2].

4.2.2 Jensen-Shannon Divergence (JSD)

In text categorization, the square root of JSD (Jensen-Shannon divergence) is a measure of distance between two documents [9]. JSD for two documents P and Q is defined as [9], [20]

$$D_{JS}(P||Q) = \frac{1}{2}(\sum_{i}^{n} P(w_i) \log \frac{P(w_i)}{M(w_i)} + \sum_{i}^{n} Q(w_i) \log \frac{Q(w_i)}{M(w_i)}) \quad (2)$$

where w is the word collection of document set $w_1, w_2, ...w_n$, $P(w_i)$ and $Q(w_i)$ refer to the distribution over the word $w_i \in W$ in the documents P and Q, and $M(w_i) = \frac{1}{2}(P(w_i) + Q(w_i))$, $\sum_i^n P(w_i) = 1$, $\sum_i^n Q(w_i) = 1$.

4.3 Feature Engineering

On FRIS dataset, the probability matrices obtained from L-LDA are utilised to construct the distance matrix by measuring the distances between project and publication probability vectors using both cosine similarity and JSD for comparison. To construct the distance matrix from probability matrices obtained by BERTopic, only cosine similarity is used, since for some projects the model predicted zero vectors and this results in undefined Jensen-Shannon divergence values. The reason for this can be seen in equation 2.

In addition to topic-based distances, we engineer the following features:

- Outcome variable
- Date criterion
- Common organization ratio
- Common author ratio

- Common discipline ratio

For each publication, the distance variable for all the projects are already gathered. But it is necessary to further explain the aggregation process. For each publication, the variables for all the projects are obtained as follows: The outcome variable is 1 if the publication belongs to the project and 0 otherwise. The date criterion is 1 if the publication date is later than the starting date of a project and 0 otherwise. For FRIS dataset, the text used for each of the projects and publications for topic modelling is the combination of its title, abstract, and keywords. For Dimensions dataset, the text used for each of the projects and publications for topic modelling is the combination of its title and abstract.

Definition 4.1 (Common Discipline Ratio)**.** Let A be the set of disciplines of a project and B be the set of disciplines of a publication. Then, common discipline ratio (CDR) for the project and the publication is defined as

$$CDR = \frac{|A \cap B|}{|A \cup B|} \tag{3}$$

Common organisation ratio (COR) and common author ratio (CAR) for a publication and a project are defined as the same way as CDR is defined. An example of the data that is used for boosting is given in Table 1. Publication IDs and project IDs are on the table for ease of demonstration; they are not in the actual dataset.

On Dimensions dataset, both cosine similarity and JSD are used to construct the distance matrix from L-LDA. Once again, only cosine similarity is used to construct the distance matrix from BERTopic because of the reasons mentioned before relating to equation 2. The common discipline ratio in the aggregation process in Figure 3 is defined the same way as in definition 4.1.

pub id	pro id	belongs	date	COR	CAR	CDR	distance
1	1	1	1	0.5	0.2	0.3	0.23
	2	0	0	0.0	0.0	0.1	0.46
	3	0	1	0.3	0.0	0.0	0.35
	.	.	.	.	.	.	.
	.	.	.	.	.	.	.
	.	.	.	.	.	.	.
2	1	0	1	0.0	0.5	0.0	0.62
	2	0	1	0.0	0.0	0.1	0.33
	3	1	1	1.0	1.0	0.5	0.25
	.	.	.	.	.	.	.
	.	.	.	.	.	.	.
	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.

Table 1: Data for classification

4.4 Classification

For all the L-LDA models that are implemented for both FRIS and Dimensions datatsets, default Labeled-LDA-Python[6] parameters are used for 10 iterations with Python 3.9. For all the BERTopic models that are implemented for both FRIS and Dimensions datasets, default BERTopic[7] (version: 0.16.2) parameters are used with Python 3.1 by setting the option '$calculate\_probabilities = True$'.

For comparison, the accuracies of the results from distance matrices without further processing is calculated. Also for comparison, instead of GB, SVM is implemented on the same datasets that is used for boosting. The same parameters are used on Dimensions dataset too for consistency.

4.4.1 Support Vector Machine (SVM)

Grid search algorithms are implemented for both L-LDA+SVM and BERTopic+SVM to determine the SVM parameters with the values for $C \in \{0.1, 1, 10, 100, 1000\}$ and $\gamma \in \{1, 0.1, 0.01, 0.001, 0.0001\}$, and radial basis function as the kernel function of SVM. For L-LDA+SVM, the best parameters are found to be

[6] `https://github.com/JoeZJH/Labeled-LDA-Python`

[7] `https://github.com/MaartenGr/BERTopic/`

$C = 100$ and $\gamma = 1$, and for BERTopic+SVM, the best parameters are found to be $C = 100$ and $\gamma = 0.1$.

4.4.2 Gradient Boosting (GB)

The parameters selected for L-LDA+GB on both FRIS and Dimensions datasets are as follows: objective (loss) function is 'binary:logistic', maximum depth of the tree is 3, learning rate is 0.01, number of boosting rounds is 10000, number of early stopping rounds is 1000, and categorical variables are allowed. The parameters selected for BERTopic+GB on both FRIS and Dimensions datasets are the same as the parameters for L-LDA+GB with the exception of number of early stopping rounds being 200.

4.5 Evaluation

It should be noted that, for the next definition, we use the term 'accuracy' loosely. Our definition is similar to the 'correctly predicted discipline percentage' (CPDP) defined by Pham et. al. [2]. If we followed the same naming practice they used, we might have named the following definition 'correctly predicted project percentage' (CPPP) instead of 'accuracy'.

Definition 4.2 (Accuracy)**.** Let m be the number of projects in the whole dataset, n be the number of publications in the test dataset, and $k = 1, 2, ..., m$ be the number of projects predicted where the projects are ordered in a descending way by their predicted probabilities and the first k projects are selected for each of the publications. Then, the accuracy of the model is defined as

$$accuracy = \frac{1}{n} \sum_{i=1}^{n} \frac{|P_i \cap T_i|}{|T_i|} \tag{4}$$

where P_i denotes the set of projects predicted such that $|P_i| = k$ and T_i denotes the set of true projects for the ith publication for $i = 1, 2, ..., n$.

The projects predicted from just distance matrix without further processing are ordered in an ascending way since smaller the distance it is more likely that the publication belongs to the projects. SVM predictions are randomly selected by Python and are not ordered since they are the results of binary classification, and as expected, give the same accuracy measurements for $k \geq 3$ where k is the number of projects predicted. On FRIS data, the projects in the whole dataset are allowed for predictions since some publications in the test dataset belong to more than one project with some of the projects are in the training dataset. On dimensions data, only the projects

in the test dataset are allowed for predictions since it is known that each publication belongs to only one project.

By systematically comparing these methods on the FRIS and Dimensions datasets, it is aimed to identify the most effective approach for predicting project-publication links in Research Information Systems.

5 Results

It is important to keep in mind that the results should not be deemed precise due to the stochastic nature of the algorithms that are implemented for the models. In all of the tables in this section, rows represent the number of projects predicted per publication and columns represent the results solely based on distance matrices, topic modelling + SVM, topic modelling + GB, and GB respectively. '-date' means date criterion variable is excluded, '-CDR' means common discipline ratio (CDR) variable is excluded, and '-date -CDR' means both date criterion and CDR are excluded. As expected, it can be seen that as the number of projects predicted increases the accuracy increases as well.

5.1 FRIS

In figure 4 we can clearly see the feature importances when topic modelling is not involved. From the feature importance plot in Figure 5 (a very similar plot is obtained with JSD) which is obtained after fitting the logistic gradient boosting model, the date criterion variable is considered to be possibly unnecessary. As it can be seen in Table 2 and Table 3, the exclusion of the date criterion variable reduces model performance when either cosine similarity or JSD is used. Other than those results, both cosine similarity and JSD show little to no difference for L-LDA+SVM and L-LDA+GB. For GB, results are consistently worse than L-LDA+GB but shows a similar decrease while the number of projects predicted decreases or the date criterion is excluded.

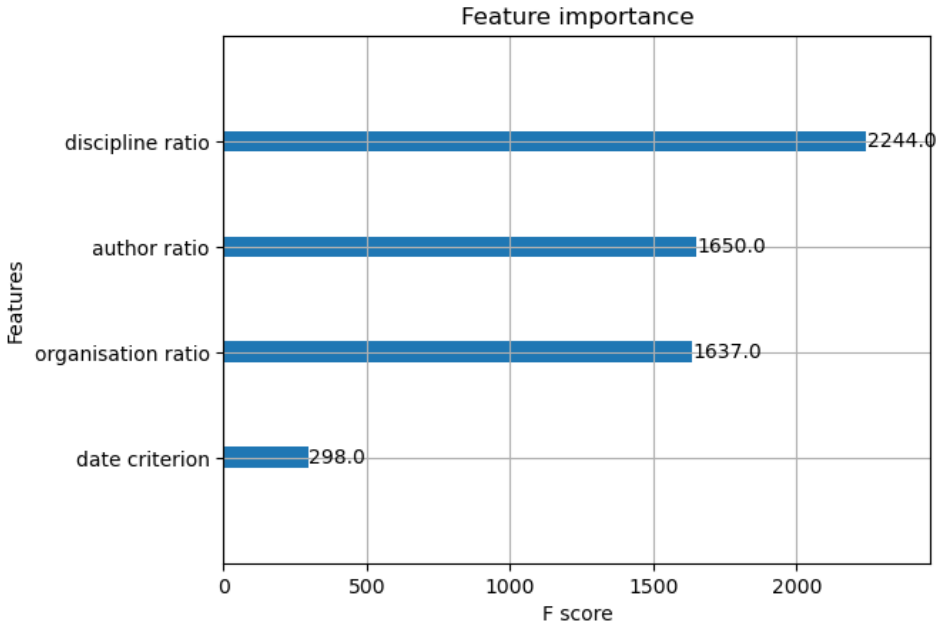

Figure 4: Feature importance plot of GB on FRIS data

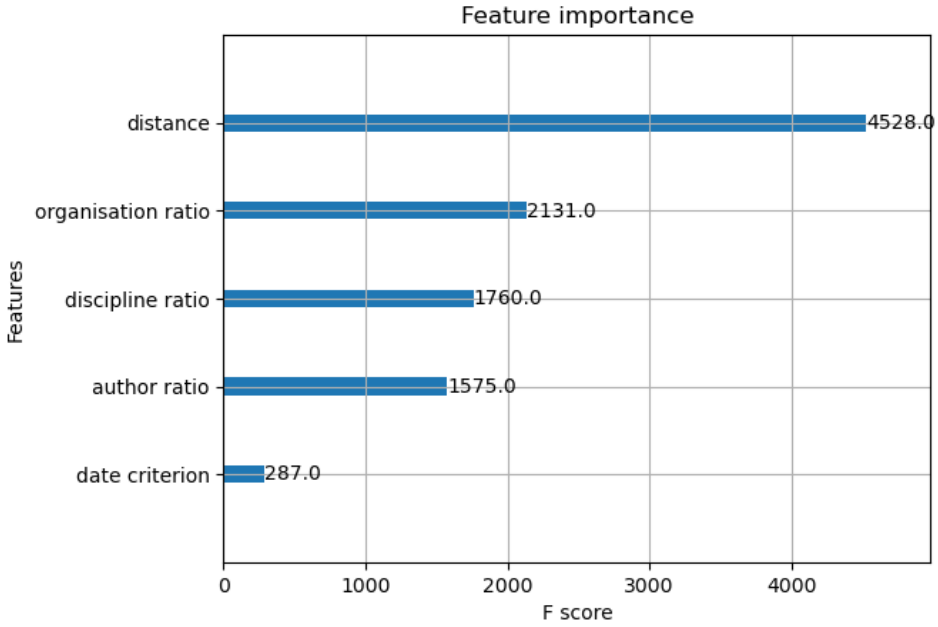

Figure 5: Feature importance plot of L-LDA+GB on FRIS data (cosine similarity)

# of projects	L-LDA	L-LDA+SVM	L-LDA+GB	GB
3	21.0%	24.4%	97.3%	91.8%
2	15.1%	21.7%	94.5%	88.1%
1	11.1%	19.0%	85.2%	70.0%
1 (-date)	-	-	78.1%	65.0%

Table 2: Accuracies of different L-LDA models on FRIS data (cosine similarity)

# of projects	L-LDA	L-LDA+SVM	L-LDA+GB	GB
3	20.4%	24.5%	97.6%	91.8%
2	16.7%	22.4%	94.5%	88.1%
1	9.9%	19.4%	85.8%	70.0%
1 (-date)	-	-	72.7%	65.0%

Table 3: Accuracies of different L-LDA models on FRIS data (JS divergence)

From the feature importance plot in Figure 6 which is obtained after fitting the logistic gradient boosting model, the date criterion variable is considered to be possibly unnecessary. As it can be seen in Table 4, the exclusion of the date criterion variable has an effect on the accuracy. BERTopic+GB shows a similar accuracy to L-LDA+GB, but for a single prediction it does not perform as well as L-LDA+GB. The exclusion of the date criterion or both the date criterion and the common discipline ratio have an effect on the accuracies calculated with each exclusion resulting in a decrease in accuracy. The results for BERTopic (results obtained only from the distance matrix) are worse than those of L-LDA and BERTopic+SVM are better than L-LDA+SVM with cosine similarity.

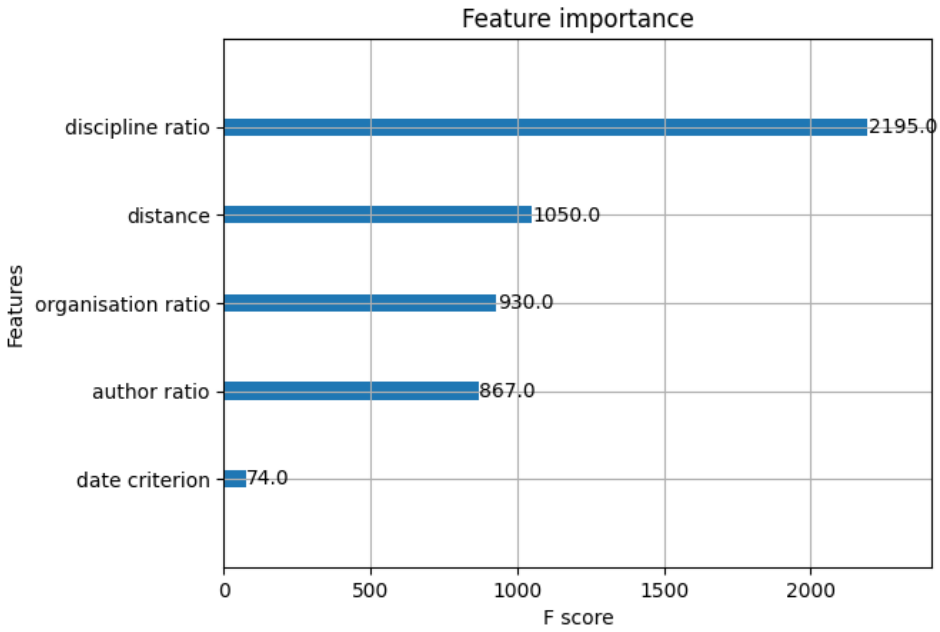

Figure 6: Feature importance plot of BERTopic+GB on FRIS data (cosine similarity)

# of projects	BT	BT+SVM	BT+GB	GB
3	5.5%	44.3%	98.2%	91.8%
2	3.7%	41.2%	92.7%	88.1%
1	3.7%	30.9%	79.8%	70.0%
1 (-date)	-	-	76.9%	65.0%
1 (-date -CDR)	-	-	74.4%	58.2%

Table 4: Accuracies of different BERTopic models on FRIS data (cosine similarity)

5.2 Dimensions

GB results show when topic modelling is now involved. From the feature importance plot in Figure 7 (a very similar plot is obtained with JSD) which is obtained after fitting the logistic gradient boosting model, the common discipline ratio variable is considered to be possibly unnecessary but the results show otherwise. As it can be seen in Table 5 and Table 6, the exclusion of the common discipline ratio variable reduced the accuracy significantly both when cosine similarity and JSD are used. Both cosine similarity and

JSD show no difference for L-LDA+SVM. L-LDA+GB gives better accuracies with JSD for all the predictions. When L-LDA is used for predictions the choice of the distance metric has little to no effect on the accuracies calculated. For GB, results are consistently worse than L-LDA+GB but shows a similar decrease while the number of projects predicted decreases.

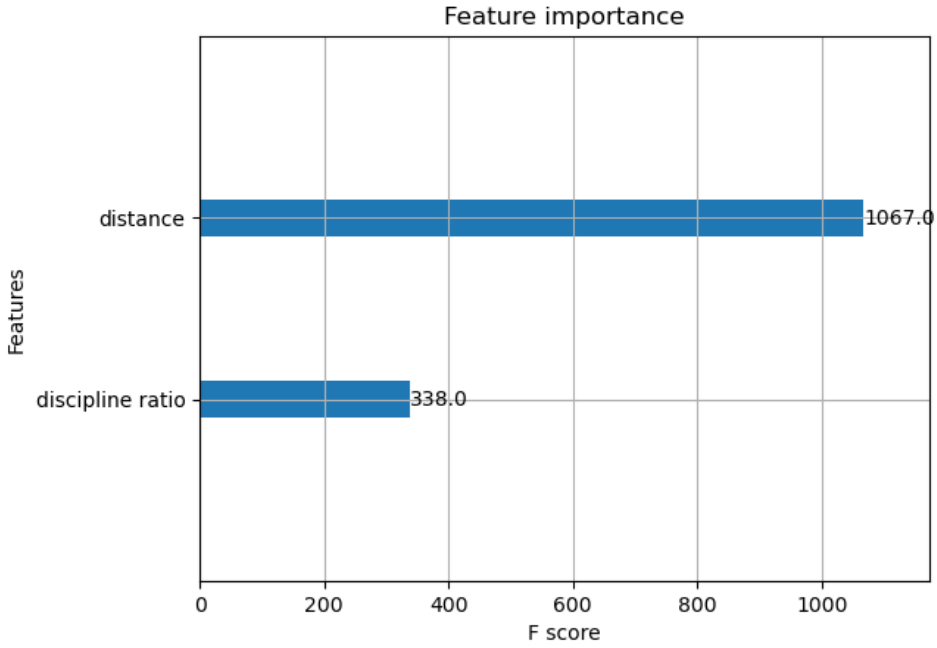

Figure 7: Feature importance plot of L-LDA+GB on Dimensions data (cosine similarity)

# of projects	L-LDA	L-LDA+SVM	L-LDA+GB	GB
3	3.1%	1.6%	38.8%	38.0%
2	1.6%	1.6%	28.7%	31.8%
1	1.6%	1.6%	15.5%	20.9%
1 (-CDR)	-	-	3.1%	-

Table 5: Accuracies of different L-LDA models on Dimensions data (cosine similarity)

# of projects	L-LDA	L-LDA+SVM	L-LDA+GB	GB
3	2.3%	1.6%	45.0%	38.0%
2	1.6%	1.6%	34.9%	31.8%
1	0.0%	1.6%	22.5%	20.9%
1 (-CDR)	-	-	3.9%	-

Table 6: Accuracies of different L-LDA models on Dimensions data (JS divergence)

From the feature importance plot in Figure 8 which is obtained after fitting the logistic gradient boosting model, the common discipline ratio variable is considered to be possibly unnecessary but the results show otherwise. As it can be seen in Table 7, the exclusion of the common discipline ratio variable reduced the accuracy significantly. BERTopic+SVM results are the same as L-LDA+SVM results. BERTopic+GB results are better than those of L-LDA+GB. L-LDA performs better than BERTopic.

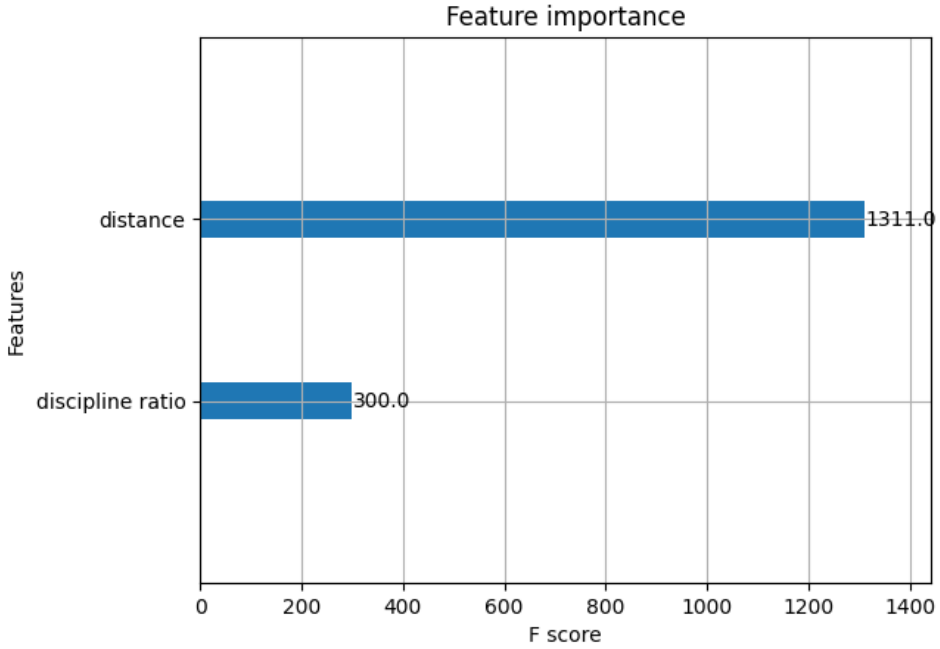

Figure 8: Feature importance plot of BERTopic+GB on Dimensions data (cosine similarity)

# of projects	BT	BT+SVM	BT+GB	GB
3	0.0%	1.6%	50.4%	38.0%
2	0.0%	1.6%	41.9%	31.8%
1	0.0%	1.6%	30.2%	20.9%
1 (-CDR)	-	-	14.0%	-

Table 7: Accuracies of different BERTopic models on Dimensions data (cosine similarity)

6 Discussion

Even though the expectation is L-LDA as a supervised topic modelling method [4] performing better than BERTopic which is an unsupervised topic modelling method [5] and this is the case for FRIS dataset, for Dimensions dataset, both topic modelling methods give very similar results. On the other hand, on FRIS dataset, the difference between the performances of L-LDA and BERTopic is not much especially when the number of projects predicted per publication is greater than 1. This is important because it means that when discipline codes for a publication are not present we can still predict the projects with sufficient accuracy.

Accuracies calculated on FRIS dataset are more than twice of those on Dimensions dataset. This indicates the importance of common organisation ratio (COR) or common author ratio (CAR) variables. To see the impact of COR and CAR variables, different L-LDA+GB models on FRIS dataset are fitted using cosine similarity. The results for these models can be seen in Table 8 with columns representing the exclusion of COR, the exclusion of CAR, and the exclusion of both COR and CAR respectively. The results in Table 8 clearly shows the importance of COR and CAR variables but especially the immense impact of COR. Fortunately, on FRIS portal, organisation IDs are present for most of the projects and publications.

# of projects	-COR	-CAR	-COR -CAR
3	33.1%	93.2%	10.5%
2	27.0%	91.0%	6.2%
1	21.8%	80.5%	1.9%

Table 8: Accuracies of different L-LDA+GB models on FRIS data (cosine similarity)

To further analyse the impact of common organisation ratio (COR) result from BERTopic+GB and L-LDA+GB where only the COR and distance variables are used are presented in Table 9. This shows that we can predict the projects for a given publication with a sufficient accuracy even when the only information we have for the publication and the projects are abstracts and organisation IDs.

# of projects	**BERTopic+GB**	**L-LDA+GB**
3	93.4%	92.6%
2	89.5%	90.4%
1	71.1%	80.9%

Table 9: Accuracies of BERTopic+GB and L-LDA+GB models on FRIS data (cosine similarity) with only predictors being COR and distance variables

A new and promising alternative for unsupervised topic modelling is ChatGPT API [21]. However, the time required for prompt engineering and the overall cost would make it a challenging contribution for the scope of this study. We are planning to include ChatGPT in the future studies as another alternative for comparison.

7 Conclusion

Logistic Gradient Boosting in combination with supervised and unsupervised topic modelling techniques as described throughout this paper, gives results with high accuracy measurements even with a limited number of variables as discussed in the previous section. This study will conceivably reduce the number of disconnected research projects and publications on FRIS portal. It is expected that further improvements can be achieved with larger datasets and more computation power.

Acknowledgments

I would like to express my sincere gratitude to Dr. Hoàng-Son PHAM and my supervisor, Dr. Amr Ali-Eldin for their invaluable assistance and guidance throughout this research. They not only helped me with understanding the models, data, and the research field in general, but also provided some of the necessary datasets and took the time to explain them in detail. Their support and insights have been instrumental in the completion of this work.

References

[1] P. de Castro and H.-M. Puuska, "Research information management systems: Covering the whole research lifecycle," in *EUNIS 2023 Annual Conference*, Paris, France: EUNIS, Jun. 2023. [Online]. Available: `http://hdl.handle.net/11366/2471`.

[2] H.-S. Pham, H. Poelmans, and A. Ali-Eldin, "A metadata-based approach for research discipline prediction using machine learning techniques and distance metrics," *IEEE Access*, vol. 11, pp. 61 995–62 012, 2023.

[3] M. Rivest, E. Vignola-Gagné, and É. Archambault, "Article-level classification of scientific publications: A comparison of deep learning, direct citation and bibliographic coupling," *PLoS One*, vol. 16, no. 5, Art. no. e0251493, May 2021.

[4] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning, "Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora," in *EMNLP '09: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, N. Eight Street, Stroudsburg, PA, 18360, United States: Association for Computational Linguistics, Aug. 2009, pp. 248–256.

[5] M. Grootendorst, *Bertopic: Neural topic modeling with a class-based tf-idf procedure*, arXiv:2203.05794 [cs.CL], 2022.

[6] K. P. Bennett and C. Campbell, "Support vector machines: Hype or hallelujah?" *ACM SIGKDD Explorations Newsletter*, vol. 2, no. 2, pp. 1–13, 2000.

[7] L. Mason, J. Baxter, P. Bartlett, and M. Frean, "Boosting algorithms as gradient descent," in *NIPS'99: Proceedings of the 12th International Conference on Neural Information Processing Systems*, 55 Hayward St., Cambridge, MA, United States: MIT Press, Nov. 1999, pp. 512–518.

[8] G. Salton and C. Buckley, "Dynamic topic models," *Information Processing and Management*, vol. 24, no. 5, pp. 513–523, 1988.

[9] X. Li, H. Jia, and L. Huang, "Investigating the performance of cosine value and jensen-shannon divergence in the knn algorithm," *Advanced Materials Research*, vol. 532-533, pp. 1455–1459, 2012.

[10] B. Vancraeynest, H.-S. Pham, and A. Ali-Eldin, "A new approach to computing the distances between research disciplines based on researcher collaborations and similarity measurement techniques," *Journal of Informetrics*, vol. 18, no. 3, 101527, 2024.

[11] B. Altınel and M. C. Ganiz, "Semantic text classification: A survey of past and recent advances," *Information Processing Management*, vol. 54, no. 6, pp. 1129–1153, 2018.

[12] Y. K. Jeong, M. Song, and Y. Ding, "Content-based author co-citation analysis," *Journal of Informetrics*, vol. 8, no. 1, pp. 197–211, 2014.

[13] S. Vancauwenbergh and H. Poelmans, "The flemish research dsicipline classification standard: A practical approach," *Knowledge Organization*, vol. 46(5), pp. 354–363, 2019.

[14] *Frascati Manual 2015*. OECD, Paris, France, Oct. 2015.

[15] *Australian and New Zealand Standard Research Classification (ANZSRC)*. Australian Bureau of Statistics, Australia, Jun. 2020.

[16] B. Thijs, W. Glänzel, and M. S. Meyer, "Using noun phrases extraction for the improvement of hybrid clustering with text- and citation-based components. the example of "information systems research"," in *Proceedings of the Workshop Mining Scientific Papers: Computational Linguistics and Bibliometrics*, vol. 1384, Jun. 2015, pp. 28–33.

[17] 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.

[18] Q. He, B. Chen, J. Pei, B. Qiu, P. Mitra, and L. Giles, "Detecting topic evolution in scientific literature: How can citations help?" In *CIKM '09: Proceedings of the 18th ACM conference on Information and knowledge management*, New York, NY, United States: Association for Computing Machinery, Nov. 2009, pp. 957–966.

[19] H.-S. Pham, B. Vancraeynest, H. Poelmans, S. Vancauwenbergh, and A. Ali-Eldin, "Identifying interdisciplinary research in research projects," *Scientometrics*, vol. 128, no. 10, pp. 5521–5544, 2023.

[20] D. Carmel, E. Yom-Tov, A. Darlow, and D. Pelleg, "What makes a query difficult?" In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, New York, NY, United States: Association for Computing Machinery, Aug. 2006, pp. 390–397.

[21] O. Gandouet, M. Belbahri, A. Jezequel, and Y. Bodjov, *Distilled chatgpt topic and sentiment modeling with applications in finance*, arXiv: 2403.02185 [cs.LG], 2024.

[22] D. M. Blei and J. D. Lafferty, "Dynamic topic models," in *ICML '06: Proceedings of the 23rd international conference on Machine learning*, New York, NY, United States: Association for Computing Machinery, Jun. 2006, pp. 113–120.

Appendix - Data and Python code

All data and code used in this study are freely accessible at:
`https://github.com/guyanik/FRIS`