IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 25 May 2023, accepted 15 June 2023, date of publication 20 June 2023, date of current version 23 June 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3287935

RESEARCH ARTICLE

A Metadata-Based Approach for Research Discipline Prediction Using Machine Learning Techniques and Distance Metrics

HOANG-SON PHAM[1,2], **HANNE POELMANS**[1,2,3], **AND AMR ALI-ELDIN**[1,2,4], **(Senior Member, IEEE)**

[1]Centre for Research and Development Monitoring (ECOOM-UHasselt), 3500 Hasselt, Belgium
[2]Data Science Institute, Hasselt University, 3500 Hasselt, Belgium
[3]Directorate Research, Library, International Office, Hasselt University, 3500 Hasselt, Belgium
[4]Computer Engineering and Control Systems Department, Faculty of Engineering, Mansoura University, Mansoura 35516, Egypt

Corresponding authors: Hoang-Son Pham (hoangson.pham@uhasselt.be) and Amr Ali-Eldin (amr.alieldin@uhasselt.be)

This work was supported by the Centre for Research and Development Monitoring (ECOOM-UHasselt), Belgium through the Interdisciplinarity and Impact Project.

ABSTRACT Forecasting research disciplines associated with research projects is a significant challenge in research information systems. It can reduce the administrative effort involved in entering research project-related metadata, eliminate human errors, and enhance the quality of research project metadata. It also enables the calculation of the degree of interdisciplinarity of these projects. However, predicting scientific research disciplines and measuring interdisciplinarity in a research endeavor remain difficult. In this paper, we propose a framework for predicting the research disciplines associated with a research project and measuring the degree of interdisciplinarity based on associated metadata to address these issues. The proposed framework consists of several components to improve the performance of research disciplines prediction and interdisciplinarity measurement systems. These include a feature extraction component that utilizes a topic model to extract the most appropriate features. Further, the framework proposes a discipline encoding component that applies a data mapping strategy to lower the dimensionality of the output variables. Furthermore, a distance matrix creation component is proposed to recommend the most appropriate research disciplines and compute interdisciplinarity associated with research projects. We implemented the suggested framework on two separate research information systems databases for research projects, Dimensions and the Flemish Research Information Space. Experimental results demonstrate that the proposed framework predicts the research disciplines associated with research projects more accurately than related work.

INDEX TERMS Metadata, research information systems (RIS), research disciplines prediction, interdisciplinarity, machine learning, distance metrics.

I. INTRODUCTION

There exists a large amount of research project metadata available in many research information systems (RIS) databases. These are rich resources for scientists as well as for policy makers [1]. For efficient searching and analyzing, research projects are categorized by a specific research classification schema which often consists of a list of research disciplines [2], [3].

For example, the Flanders Research Information Space[1] (FRIS) is a regional web portal, governed by the Flemish government. Almost 40 data sources in Flanders contribute information on (partially) publicly financed research (e.g., researchers, research institutes, projects, and publications) to the FRIS site. It empowers the Flemish government to create

The associate editor coordinating the review of this manuscript and approving it for publication was Xianzhi Wang.

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

reports, analyses, and statistics for policy-making and trend monitoring. Each object in FRIS is assigned one or more research disciplines which is obligatory for any object added to the FRIS database since 1 January 2019. To label research objects with research disciplines, FRIS makes use of the Flemish Research Discipline Standard[2] (''Vlaamse Onderzoeksdiscipline Standaard'', abbreviated VODS, in Dutch), which has been described in [3]. The VODS has four hierarchical levels that mirror research disciplines at different levels of granularity, with 7, 42, 382, and 2493 disciplines, respectively. The first level corresponds to the OECD FORD [4] 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. The majority of objects in FRIS are assigned a level 4 discipline.

Assigning disciplines to projects is often done manually by researchers or administrative staff. Besides being a time-consuming task, it is rather subjective and may incur human errors. As a consequence, incorrectly assigning disciplines or missing disciplines may occur. Therefore, for data quality improvement and administrative burden reduction, automatically predicting disciplines related to research projects can be seen as an essential task for research information systems. Besides, it can support other activities such as grant competition management, and interdisciplinarity evaluation by funding agencies.

Predicting multiple disciplines related to a research project, however, remains challenging [1], [5]. The first challenge is related to the quality of labeled data. For example, since research disciplines are often manually assigned to the research projects, the data might have been incorrectly labeled or with missing labels. Further, the abstracts of research projects are often short and do not contain rich information for training machine learning models. Directly applying a machine learning (ML) classification model on these databases may be accompanied by poor accuracy [6], [7]. Another difficulty in multiple discipline classification comes from the ML techniques themselves since multi-label classification is more complicated than multi-class one. In practice, various studies [1], [5] attempted to classify scientific publications by fields of study (research disciplines). However, their experimental results revealed that the performance of classical ML, as well as deep learning algorithms, are not satisfactory, even for single-label classification [5].

Interdisciplinarity is another piece of valuable information for researchers or policymakers to evaluate the degree of knowledge integrated within the research document [8]. Interdisciplinary research (IDR) is crucial to address the complex problems that our society is confronted with like the COVID-19 pandemic and global warming. Governments, funders, and research institutions each have their own activities to stimulate interdisciplinary research. It is however still unclear whether these activities are effective in stimulating interdisciplinary research because it is not known how interdisciplinary research activities can be recognized [9]. In the context of IDR measurement, various methods have been proposed; however, there is no consensus about the validity as well as the results of these methods [10], [11], [12]. According to a study by [13], the choice of data, the methodology, and the indicators can produce seriously inconsistent and even contradictory outcomes.

Most of the related work applied traditional ML classification algorithms [1], [5], while some used deep machine learning to predict fields of study related to the research document [1], [14]. Within these approaches, the percentage of correctly predicted disciplines of the project relies on the performance of the classifier. However, to the best of our knowledge, the performance of these algorithms is not satisfactory in practice.

To overcome the above-mentioned limitations, we propose a framework, with a number of components, to help improve the performance of the research discipline prediction and interdisciplinarity measurement systems. The main contribution of our proposed framework is the introduction of the following three components in the context of research discipline prediction and interdisciplinarity assessment research information systems: 1) A feature extraction component which runs an unsupervised topic model to extract the most appropriate features. 2) A discipline encoding component which provides a mapping technique to reduce the dimensionality of output variables. 3) A distance matrix creation component which generates a distance matrix based on a supervised topic modeling approach in order to provide disciplines relevant to projects. This matrix allows us to determine additional disciplines that are close to the predicted disciplines. The matrix also plays an important role in interdisciplinarity calculation. We evaluate the proposed framework on two RIS databases: Flanders Research Information Space (FRIS) [15], and Dimensions - a largely linked research information dataset [16]. The results show that the performance of the proposed framework with the three proposed components could achieve higher performance than related work.

The rest of the paper is organized as follows. Section II briefly reviews related work. Section III presents the proposed framework. Section IV presents experimental and comparison results. Section V concludes the paper and proposes future work.

II. LITERATURE REVIEW

This work is close to approaches that aim to automatically classify research data by fields of study or research

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

disciplines. In this section, we highlight the most relevant studies in this domain.

The frequently used methods in bibliometrics to classify research documents are bibliographic coupling, co-citation, and direct citation [17]. These techniques are mostly based on citation network analysis. Several clustering techniques can be used to categorize research disciplines connected with documents. This is a bottom-up approach that requires linking the clusters obtained to categories. The main disadvantage of these approaches is that determining optimal levels of cluster aggregation is a difficult task [17].

More recently, machine learning and deep learning techniques have been applied to classify research documents. These techniques could advantageously be used in bibliometrics. Evykens et al. [5] have applied classical supervised machine learning algorithms such as Multinomial Naïve Bayes [18], Support Vector Machine [19], Random Forest Classifier [20], and Gradient Boosting [21] to classify publications. The experimental assessment of these machine learning techniques on titles and abstracts of publications showed that the models did not work as expected with the unseen data. They limited the problem as a multiclass classification problem, i.e, each publication was assigned only one label in a set of 77 labels. As a result, the best-reported F1-score was slightly over 80% for Gradient Boosting models. Other models performed relatively poor when compared to Gradient Boosting.

Rives et al. [1] used deep learning, i.e., a modified character-based convolutional deep neural network to classify articles based on abstracts. They tested the model on a dataset with more than 40 million scientific articles. They also compared the performance of the deep learning model to bibliographic coupling, co-citation, and direct citation. The results showed that the performance of the deep learning approach was equivalent to the graph-based bibliometric approaches. In particular, the precision of deep learning was 57%. It was slightly better than the precision of bibliographic coupling and direct citation which were 41% and 53%, respectively.

Weber et al. [14] employed different machine learning techniques such as Decision Tree Classifier, Random Forest Classifier, and deep learning techniques such as multilayer perceptron neural networks, and recurrent neural networks to classify research documents. They meticulously produced training and testing datasets to evaluate the performance of the models. To avoid the imbalanced data problem, these datasets were cleaned and processed. In particular, the data was compiled into a set of 613,585 records with 20 general fields of study. According to their results, the multilayer perceptron model performed the best, followed by long short-term memory models. The classical machine learning techniques did not perform well on these datasets.

Regarding IDR measurement, there is extensive literature on measuring IDR in general [9]. A typical method for evaluating IDR of research activity is to assess the diversity of disciplines associated with it [10]. In theory, diversity measurement takes into account three factors: balance, variety, and disparity. Regarding IDR calculation, balance is the distribution of disciplines, variety is the number of disciplines, and disparity is the degree of dissimilarity of the disciplines. The most popular method is the citation-based approach which relies on the subject classification of the publication's references. The degree of IDR of a publication is evaluated by the diversity of the categories to which the references belong [12], [22]. Another approach relies on the professional skills of the authors who participated in the research activity. The degree of IDR of research activity is measured by the diversity of the authors' disciplines [12], [23]. Within these methods, the calculation of disparity requires a predefined category similarity matrix which depends on the discipline classification systems. This is the main limitation of these approaches since they depend on the way bibliographic databases assign documents to their subject classification schemes [9]. In addition, any change in the classification systems could significantly affect the results.

Compared to the rich literature of studies on measuring IDR based on citation analysis, there is only a few studies exploring IDR with text-based methods. Typical approaches in this research direction are keyword analysis and topic modeling [24], [25], [26], [27]. Despite these promising approaches in the effective prediction of IDR, text-based approaches need a certain amount of high-quality text which is not always available in many databases.

There are some major differences between the work proposed in this paper and related work. Firstly, the databases used in this work are research project repositories which can be smaller than those used by the related work. However, we assume that this does not have an impact on the efficacy of the proposed framework. Secondly, the number of research disciplines analyzed in this work is larger. Additionally, in order to improve the performance of the classification models, we propose the use of topic modeling to create a feature matrix. The feature matrix is fed into the ML classification model. The rationale for this is that, as will be explained later, the feature matrix produced by the unsupervised topic model is more representative of the projects when fed to the classifier as input data rather than typical project input data. Moreover, because the number of topics is generally much fewer than the number of terms in the data, the time required to train the model is reduced. Further, we propose a data mapping technique to reduce the dimensionality of the classifier outputs which helps improve the performance. Finally, to boost research discipline prediction systems performance, we combine the ML classifier with a pre-computed distance matrix, which can be considered novel for this work as we are not aware of any previous studies that have combined a research discipline prediction classifier with a distance matrix to improve prediction systems.

TABLE 1. List of notations used in the paper.

Notation	Description
C	input project data
T	a set textual description of projects
V	a set of discipline codes used by a database
K	number of unique disciplines in data
M	distance matrix
S	similarity matrix
t	a textual description of a project, $t \in T$
v	a set of disciplines associated with a project, $v \subseteq V$
t'	a textual description of a project, $t' \notin T$
v'	predicted disciplines associated with a project
$TF-IDF$	term frequency and inverse document frequency
$CPDP$	correctly predicted discipline percentage
$CPDP\_D$	$CPDP$ with the distance matrix
mm	number of original disciplines per project
nn	number of newly encoded disciplines per project
N	number of projects
$V1$	number of unique disciplines
$V1'$	discipline encoding table

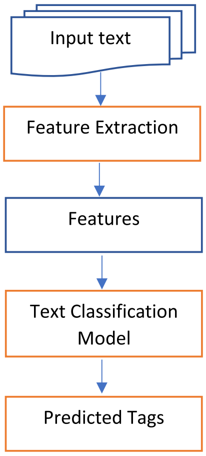


FIGURE 1. Schema of a generic text classification model.

III. PROPOSED FRAMEWORK

A. SYSTEM MODEL

In order to easily read the following sections, we present a list of notations used throughout the paper in Table 1.

We assume that a research information system database, stores projects' metadata, minimally consisting of titles, abstracts, keywords, and disciplines. Each project has a list of disciplines. Each discipline has a unique code identifier. For example, in FRIS, `0101` represents discipline `Mathematical sciences` and `0106` represents discipline `Biological sciences`. The input project metadata is defined as $C = (T, V)$; T is a set of textual descriptions which is a combination of titles, keywords, and abstracts. $V = \{v_1, v_2, \ldots, v_N\}$ is the set of discipline codes in the data. A project $p \in C$ is a pair of (t, v) where $t \in T$ and $v \subseteq V$. Given an unseen project $p' = (t', v')$ with $v' = \emptyset$, the first objective is to predict a set of disciplines $v' \subseteq V$, associated with t'. In order to predict disciplines related to project p', we propose to apply text classification algorithms. The generic framework of text classification is illustrated by Fig. 1. The input data, T, are vectorized in order to create features. Text vectorization is a process through which text data are converted into numerical data. According to a review of feature extraction methods on machine learning by Suhaidi et al. [28], there are various approaches to create features such as Bag of Words, Binary Term Frequency, Term Frequency, and Term Frequency-Inverse Document Frequency (TF-IDF) or using embedding approaches such as Doc2Vec [29]. Within these approaches, TF-IDF and Doc2Vec are widely-used methods since they calculate how relevant a word, in a series, or corpus, to a text is. In addition, since not all machine learning algorithms can deal with categorical data, the input labels, V, need to be converted into numerical data. One-hot encoding [30] is a straightforward method for this purpose. In this paper, we propose a different discipline encoding mechanism which will be introduced later, to deal with this issue which is more suitable for the context of this framework. After creating the features and encoding labels, a text classification model can be used to create the trained classification model. The trained model then is used to predict the disciplines of an unseen project.

In order to enhance the performance of the ML classifier, we integrate it with a pre-calculated distance matrix, enabling us to identify related disciplines in proximity to the predicted ones. The distance matrix, denoted by M, is a KxK dimension matrix where each cell m_{ij} is a value representing the difference between two disciplines i and j. In the distance matrix, $m_{ij} = 0$ (if $i = j$) and $m_{ij} = m_{ji}$ (if $i \neq j$). A distance matrix can be transformed from a similarity matrix, denoted by S, which is a matrix representing the similarity between disciplines. In order to calculate the similarity between two disciplines, we can examine how often they co-occur. For example, we can count the number of co-citations of two disciplines based on publications [9], count the number of co-occurrences of disciplines in projects, or consider co-occurrences of keywords of two disciplines [25]. In this work, we create the distance matrix using the cosine similarity approach [31].

IDR is a research mode that involves two or more research disciplines. To measure IDR, most of the studies focus on analyzing the research disciplines associated with the project [9]. The interdisciplinarity of a project p can be calculated by a measure such as Simpson index [32], Shannon entropy [33], or Rao-Stirling diversity [34]. Common indicators for measuring IDR can be found in the study of Wang et al. [13]. Given a project p' without assigned disciplines, the objective is to effectively predict disciplines and interdisciplinarity score related to p' based on its textual description.

B. FRAMEWORK

In this section, we present the specifications of the framework components (see Fig. 2).

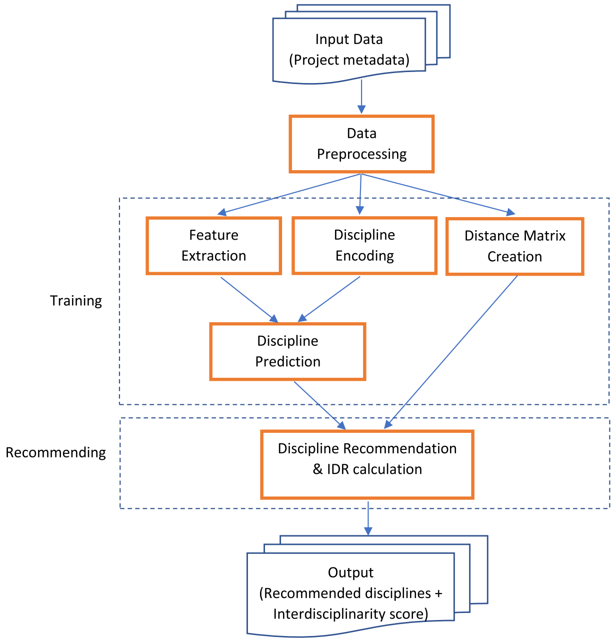


FIGURE 2. **The schematic diagram of the framework.**

1) DATA PREPROCESSING

Preprocessing is an essential step in text classification. This step helps prepare data for the classification model according to the proposed system model. Appropriate data preprocessing contributes to the efficient processing of data in the framework. The preprocessing presented here is done on two different research information systems (RIS) databases and can be reused on any other similar RIS database. This component performs various steps such as data extraction and data cleaning in order to prepare the data for the ML models.

a: DATA EXTRACTION

Research information systems usually provide a set of open APIs or web services interfaces to open up their data for external access. They implement different data structures, for example, Dimensions returns a python data frame object for a request,[3] whereas FRIS returns an XML file with Common European Research Information Format (CERIF).[4] In order to have a specific input, this component is designed to query data from a database and create a python data frame object for the data preprocessing step. Particularly, after collecting project data, we extract titles, keywords, abstracts, and disciplines and store them in a data frame.

b: DATA CLEANING

The collected metadata may contain incorrect text format, empty or too short text, etc. In order to provide good-quality data for the models, the text should be cleaned first. In this study, we employed a conventional method for data cleaning, which typically involves common procedures such as converting text to lowercase, eliminating punctuation marks, and retaining solely English language text.

As mentioned, each project is assigned one or more disciplines from a set of disciplines. The frequency of project disciplines usage can vary widely. As a result, the distribution of disciplines is not balanced. This can have a significant impact on the performance of a classification model. For example, in FRIS data, certain disciplines such as `Biological sciences`, `Computer sciences` occur with high frequency, whereas others such as `Social and economic geography`, `General and logistic services` occur only a few times. The significant difference in occurrences of disciplines in data leads to an imbalanced data problem in the classification model. This step aims to reduce the imbalanced data problem by excluding projects that contain low frequent disciplines. In order to do that, we first identify these disciplines with low frequency then associated projects are removed from the dataset. Improving the prediction of these low frequent disciplines is not in the scope of this work and will be considered in future work.

2) DISCIPLINES ENCODING

Since not all traditional ML classification algorithms can deal with categorical data, in this work, we propose converting categorical data into numerical values. One-hot encoding technique is often used for this purpose [35]. To encode categorical data, this approach performs two steps: 1) For each category, it creates a column, so the number of columns is equal to the number of categories. 2) For each column, it puts '0' for others and '1' as an indicator for the appropriate column. This method is preferable but has some disadvantages. Firstly, it can lead to increased dimensionality since a column is created for each category. This can make the model more complex and slow to train [36]. Secondly, it can lead to over-fitting, especially if there are many categories in the variable and the sample size is relatively small.

The discipline encoding component performs a data mapping technique in order to convert categorical data into numerical data that helps accelerate the performance of the classifier. In particular, instead of using one-hot encoding as the usual machine learning classification approach, we map discipline labels to numerical values. It is noticed that the one-hot encoding creates K output variables. This number is usually large. For example, the Dimensions dataset has 213 labels, whereas the FRIS dataset has 42 labels.

In order to transform discipline labels into numerical values, we first create a coding table where each discipline label corresponds to an integer number. Based on this coding table, we then transform discipline labels in each project into integer values. For example, Table 2 shows a coding table generated from FRIS data. With this coding table, we can transform the disciplines of the projects into n new discipline codes. Each one corresponds to a discipline label of the project.

[3] https://api-lab.dimensions.ai/cookbooks/1-getting-started/3-Working-with-dataframes.html

[4] https://eurocris.org/services/main-features-cerif

Algorithm 1 Discipline Encoding

```
Input: $V$, $V1'$, $n$,
Output: $V'$.
for $v$ in $V$ do
    if $|v| < n$ then
        $v \leftarrow$ add 'NULL' label
    else
        $v \leftarrow$ keep the first $n$ labels in $v$
    end if
end for
for each $v$ in $V$ do
    for each $vv$ in $v$ do
        $vv \leftarrow V1'(vv)$ // adding the corresponding code from the coding table
    end for
    $V' \leftarrow v$
end for
return $V'$
```

TABLE 2. Example of the discipline code table.

Discipline label	Encoded number
'0101 Mathematical sciences'	0
'0102 Information and computing sciences'	1
'0103 Physical sciences'	2
...	...

The discipline encoding process is illustrated in Algorithm 1. The number of labels, n, is defined through an analysis of the distribution of disciplines over the projects in dataset. Usually, projects disciplines set v has more than two disciplines and less than the total number of unique disciplines in the dataset. For each project, if the number of entered disciplines is larger than n, then only the first n disciplines are encoded. The rest of the disciplines are ignored. Assuming that the most relevant disciplines for a project are usually entered first. If the project has fewer disciplines than n then we add a NULL label to that project such that the number of labels is equal to n.

In this algorithm, we transform v with mm disciplines into n disciplines where $n \leq mm$. Each discipline (vv), in v, will have a possible integer number in a range of 0 to $K - 1$. Suppose a project "i" has two disciplines: `0101 Mathematical sciences` and `0102 Information and computing sciences`, and a project "j" has two disciplines: `0101 Mathematical sciences` and `0103 Physical sciences` the encoded integer labels of these two projects is shown in Table 3.

3) FEATURE EXTRACTION

This component adopts an unsupervised topic modeling approach to extract topic probability distribution values over projects which are afterward used as features for the discipline prediction component. LDA [37], Top2Vec [38], BERTopic [39] are examples of unsupervised topic models.

TABLE 3. Example of discipline encoded table.

Project	Discipline_1	Discipline_2
i	0	1
j	0	2
...	...	...

Algorithm 2 Feature Extraction

```
Input: $T$. project data
Output: $F$. feature matrix
    {step 1: data preprocessing, e.g., data cleaning, stemming, stopword removing}
    $T' \longleftarrow$ preprocessing($T$)
    {step 2: Bag of word calculation}
    $F' \longleftarrow BoW(T')$
    {step 3: unsupervised topic model training}
    $model \longleftarrow$ train an unsupervised topic model on $F'$
    {step 4: get project-topic probability distribution matrix from $model$}
    $F \longleftarrow get - project - topic(model)$
    return $F$
```

Unsupervised topic modeling is a probabilistic model for discovering the topics that occur in a collection of documents. The unsupervised topic model excels at feature reduction and can be employed as a preprocessing step for other models, such as machine learning algorithms.

The feature extraction process is illustrated in Algorithm 2. The algorithm takes a set of project data as the input. After preprocessing and calculating bags of words, the unsupervised topic model is able to analyze the data to provide two outputs: a project-topic probability matrix that represents topics probability distribution over projects, and a topic-term probability matrix that represents words probability distribution over topics. In the project-topic probability distribution matrix, each row represents a project, and each column represents a topic. The values in the matrix represent the probability that a particular topic is present in a given project. The sum of the probabilities across all topics for a given project will be equal to one since a project must belong to one or more topics. In the topic-term probability distribution matrix, each row represents a topic, and each column represents a term (word). The values in the matrix represent the probability that a particular term belongs to a given topic. The sum of the probabilities across all terms for a given topic will be equal to one since a mixture of different terms must completely represent a topic.

In this framework, we use the output of the unsupervised topic model, i.e., project-topic probability distribution as feature vectors to input to the machine learning classification algorithm used in the discipline prediction component. The reason for using the project-topic probability matrix is that it provides a more representative representation of the projects. Additionally, we expect that using this matrix as input to the classification model will improve its performance. Another advantage is that the number of topics is typically much smaller than the number of terms in the data, resulting in faster training of the classification model using the topic probability distribution.

4) DISTANCE MATRIX CREATION

Different from the unsupervised topic models, a supervised topic model is able to constrain the topic model to use only those topics that correspond to a project's (observed) label set. For example, the L-LDA topic model [40], a modification of LDA, incorporates supervision by simply constraining the topic model to use only those topics that correspond to a project's (observed) label set. In practice, each project is assigned one or more discipline codes from a set of disciplines. Considering this set of disciplines to be labels of the data, we can apply the supervised topic model to discover the probability distribution of disciplines over projects and the probability distribution of words over disciplines which is known as the discipline-term probability distribution matrix.

The process of generating a discipline-term matrix using L-LDA follows the same procedure as depicted in Algorithm 2, with the exception that L-LDA requires labeled input data (in this case, the labels are disciplines). The resulting output of L-LDA consists of two matrices: discipline-term probability and project-topic probability. Instead of returning the project-topic matrix, the discipline-term one is returned.

In this framework, the supervised topic model is used for calculating the discipline distance matrix which presents the distances between disciplines in the data. It plays an important role in the recommendation of disciplines and in IDR calculation.

As mentioned, the supervised topic model can discover the probability distribution of terms over disciplines. In other words, each discipline is represented by a vector of terms probability distribution values. We assume that the two disciplines are similar if their representative vectors are within a distance threshold. To calculate the distance between two vectors, we apply the cosine distance measure [31] which is a widely utilized similarity metric in machine learning, data mining, and natural language processing. Its primary advantage is its independence on the dimensionality of the vectors being compared, allowing it to be applied to high-dimensional data. Further, it is insensitive to the magnitude of the data, enabling vectors of different magnitudes to still have a high cosine similarity score. Suppose two disciplines $d_i = \{a_1, a_2, .., a_{nt}\}$ and $d_j = \{b_1, b_2, .., b_{nt}\}$, the distance between d_i and d_j is calculated as follows:

$$\begin{aligned} distance(d_i, d_j) &= 1 - cosine\_similarity(d_i, d_j) \\ &= 1 - \frac{\sum_{i=1}^{nt} a_l b_l}{\sqrt{\sum_{l=1}^{nt} a_l^2}\sqrt{\sum_{l=1}^{nt} b_l^2}}, \end{aligned} \tag{1}$$

with nt is the total number of terms.

The procedure to calculate the discipline distance matrix based on the discipline-term matrix is illustrated by Algorithm 3. The input is a discipline-term matrix, D, which represents the probability distribution of terms in a set of disciplines. The algorithm first iterates through all pairs of rows in D and calculates the similarity score between them using the cosine similarity measure. The resulting distance (1-cosine similarity) is then added to the distance matrix, M.

Algorithm 3 Distance Matrix Calculation

```
Input: $D$. discipline-term matrix
Output: $M$. discipline distance matrix
for each pair of $d_i, d_j \in D$ do
    $cs \longleftarrow cosine\_similarity(d_i, d_j)$ {calculate cosine similarity of $d\_i$ and $d\_j$}
    $M[i, j] \longleftarrow 1 - cs$ // add distance to M
end for
return $M$
```

5) DISCIPLINE PREDICTION CLASSIFICATION MODEL

This component is designed to train a multi-label classification model. This classifier is afterward used to predict the disciplines related to an unseen project. In theory, any multi-label classification model can be applied to build a classifier. To evaluate the proposed approach, we employed various classification models, ranging from simple ones like K Nearest Neighbor (kNN) [41], Decision Trees (DT) [42] and Logistic Regression (LR) [43] to more robust ones such as Support Vector Machine (SVM) [19], Random Forest (RF) [20], Gradient Boosting (GB) [21] and Extra Trees (ET) [44]. As we understand it, Random Forest, Extra Trees, and Gradient Boosting are all ensemble learning techniques. Specifically, Random Forest and Extra Trees are known as Bagging Ensemble Learning, while Gradient Boosting Machine is known as Boosting Ensemble Learning. These are well-known algorithms, therefore, we do not present them in detail here. These above mentioned algorithms are adopted due to their simplicity and powerfulness. Each algorithm has its advantages. For example, kNN is a robust classifier that is often used as a benchmark for more complex classifiers. RF is also a widely-used algorithm. It works well with both categorical and numerical data. No scaling or transformation of variables is usually necessary. Last and not least, SVM is one of the most powerful prediction methods.

a: EVALUATION METRIC

In traditional classification models, model accuracy is calculated as follows:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{2}$$

where TP = True positive; FP = False positive; TN = True negative; FN = False negative. Since, in the proposed model, the labels are encoded as numbers, some factors such as TN, FP, and FN are not applicable. As a result, this accuracy measure can not be applied to evaluate the performance of the classification model. For example, given two projects with true and predicted labels as follows: y_test = [[6, 7], [10, 9], y_pred = [[6, 8], [9, 10]]. If we compare disciplines one by one in y_test and y_pred, the true positive of the first project is 50% and the true positive of the second project is 0%. This calculation is however not suitable in the context of this framework because we need to compare the true and predicted labels without considering the order of the labels. With this purpose, we propose a metric to count the true positive values in a way suitable for the context of this framework. We define a metric called the correctly predicted discipline percentage (*CPDP*). *CPDP* of a project is the ratio of the number of correctly predicted disciplines to the number of true disciplines. The *CPDP* of the model is calculated as follows:

$$CPDP = \frac{1}{N}\sum_{i=1}^{N}\frac{\#correctly\_predicted\_disciplines}{\#true\_disciplines}, \quad (3)$$

with N as the number of projects.

Since in this study, we use the distance matrix to improve the performance of the ML classification model, we propose a metric, denoted by *CPDP_D*, that takes into account the distance matrix to measure the performance of the ML classification model (see Algorithm 4).

The algorithm takes in four inputs: *y_test*, *y_pred*, *M*, and *distance_threshold*. *y_test* and *y_pred* are lists of lists, where each inner list represents a sample and contains the true (*y_test*) and predicted (*y_pred*) disciplines for that sample. *M* is a square matrix that contains the distances between all pairs of disciplines. *distance_threshold* is a float number indicating the minimal value of distance for two disciplines to be nominated similar. For each sample in *y_test* and *y_pred*, the algorithm checks each predicted discipline. If it is predicted correctly, we increase the number of correctly predicted disciplines (*CPD*) by one. Otherwise, it looks for the closest discipline to the incorrectly predicted one. If the distance between the closest discipline and the incorrectly predicted one is less than or equal to *distance_threshold* then we can use this closest discipline to compare to the true disciplines. If the closest discipline is found in the true disciplines, *CPD* is increased by one. After counting *CPD* for each sample, the algorithm calculates the average number of *CPD* per predicted discipline for each sample and then calculates the *CPDP* per predicted discipline across all samples.

6) DISCIPLINE RECOMMENDATION & IDR CALCULATION

This component aims to recommend the most appropriate disciplines related to an unseen project based on its textual description.

a: DISCIPLINE RECOMMENDATION

The discipline recommendation component is illustrated by Algorithm 5. It includes the following steps:

1) The trained unsupervised topic model is loaded to discover the topic probability distribution in the input.
2) The trained classifier is used to predict the disciplines of the project based on its topic probability distributions. The classifier outputs a list of predicted disciplines.
3) The distance matrix is used to find the closest disciplines with predicted disciplines. For the sake of simplicity, this step outputs the two closest disciplines for each predicted discipline.
4) v_p and v_c are combined in order to find a set of recommended disciplines. This step will display recommended disciplines to the users.

Algorithm 4 CPDP_D Calculation

```
Input: $y\_test$. list of true labels,
       $y\_pred$. list of predicted labels
       $M$. distance matrix
       $distance\_threshold$. distance threshold.
Output: $CPDP\_D$. correctly predicted discipline percentage.
sum_CPD ⟵ 0
for each sample, s do
    y_test_s ⟵ y_test[s]
    y_pred_s ⟵ y_pred[s]
    for each discipline l in y_pred_s do
        if l is in y_test_s then
            CPD+ = 1
            remove l from y_test_s
        else
            distances ⟵ get distances between l and others from M
            closest_discipline ⟵ get the discipline that has the smallest distance
            if distance of closest_discipline ≤ distance_threshold then
                if closest_discipline in the y_test_s then
                    CPD+ = 1
                    remove closest_discipline from y_test_s
                end if
            end if
        end if
        sum_CPD+ = CPD / by size of y_pred_s
    end for
end for
CPDP_D ⟵ sum_CPD / number of sample
return CPDP_D
```

b: IDR CALCULATION

Given a set of disciplines provided by the discipline recommendation component, we can calculate the diversity of disciplines. In particular, the diversity of disciplines, denoted *idr*, is calculated based on the Rao-Stirling diversity index [34]. The Rao-Stirling diversity index provides a more robust and nuanced measure of interdisciplinarity than the Simpson index [32] and Shannon entropy [33], making it a

Algorithm 5 Discipline Recommendation

```
Input: $t$. textual description of the project,
    $M$. distance matrix,
    $V1$. list of disciplines,
    $distance\_threshold$. distance threshold
Output: $v_r$. list of disciplines recommends to the user
{step 1: calculate topic probability distribution.}
$topic\_probability \leftarrow$ calculate topic probability distribution of $t$ based on the unsupervised topic model
{step 2: predict disciplines.}
$v_p \leftarrow$ predict disciplines of $topic\_probability$ based on the classification model
{step 3: find close disciplines based on the distance matrix.}
$v_c \leftarrow \emptyset$
for each $vv$ in $v_p$ do
    $index \longleftarrow$ find index of $vv$ in the $V1$
    $distances \longleftarrow$ get row $index$ of the $M$
    for each $d$ in $distances$ do
        if $d \leq distance\_threshold$ then
            $v_c \leftarrow$ add corresponding discipline in $V1$
        end if
    end for
end for
{step 4: combine disciplines}
$v_r \leftarrow$ combine disciplines: $v_p, v_c$
return $v_r$
```

Algorithm 6 IDR Calculation

```
Input: $v$. list of disciplines with distance values,
    $M$. distance matrix,
    $V1$. list of disciplines
Output: $idr$. Rao-Stirling diversity score of disciplines
1: $discipline\_size \longleftarrow size(disciplines)$
2: $sum\_distances \longleftarrow$ sum of distances of disciplines
3: $discipline\_vector \longleftarrow$ create list of $discipline\_size$ zeros
4: for each $d$ in $disciplines$ do
5:     $index \longleftarrow$ find index of $d$ in $V1$
6:     $val \longleftarrow$ distance of discipline $d$ divide by $sum\_distances$
7:     assign $val$ to $discipline\_vector$ at the position $index$
8: end for
9: for each $vv$ in $v$ do
10:     $index \longleftarrow$ find index of $vv$ in $V1$
11:     $val \longleftarrow$ probability of discipline $vv$
12:     $dv \leftarrow val$ at the position $index$
13: end for
14: for each pair of $i, j$ in $dv$ do
15:     $idr \longleftarrow idr + dv[i] * dv[j] * M[ij]$
16: end for
17: return $idr$
```

valuable tool for researchers studying interdisciplinary collaboration and innovation [10]. It is well-suited for measuring interdisciplinarity as it considers not only the number of disciplines and their probability distribution but also incorporates the pairwise distances between them. Specifically, the *idr* is calculated as follows:

$$idr = \sum_{i,j(i \neq j)}^{K} f(d_i) * f(d_j) * d(d_i, d_j), \tag{4}$$

where K is number of disciplines; $f(d_i), f(d_j)$ are probability distribution of discipline d_i and d_j, respectively; $d(d_i, d_j)$ is the distance between disciplines d_i and d_j.

Algorithm 6 shows in detail the steps of IDR calculation. Given a list of disciplines, v, we first need to create a vector where each element of the vector is a probability distribution of a discipline over a project. This task is done from lines 1 to 13. The discipline vector, dv, is then used in interdisciplinarity calculation based on (4). This task is done from lines 14 to 16.

IV. EXPERIMENTAL WORK

The proposed framework has been simulated using python. In order to preprocess data, train the topic models, and train the classification model, we implemented various python packages such as nltk,[5] spacy,[6] gensim,[7] and sklearn.[8] The packages are presented in detail in the related step. All experiments were conducted on a laptop with Intel Core i5-10210U, CPU 1.60GHz, 16GB memory, and Windows operating system.

A. DATA AND MODEL SELECTION

1) DATASETS

In order to evaluate the model, we collected data from two databases: FRIS and Dimensions. FRIS is the regional portal for researchers and their research in Flanders, Belgium. It provides information on researchers, scientific projects, and publications since 2008. The data on the FRIS Research portal is open access and offered through web services. Each research object, e.g., a publication or a project, in FRIS is classified into one or more research disciplines from VODS classification schema [3].

FRIS provides web services that allowed us to extract data from its database. The procedures to extract project data from the FRIS database are given in appendix V-C. For analysis purposes, we only selected projects that started in 2010 and later on and that were assigned more than one research discipline code of level 2 of the VODS. As a result, 5702 projects were selected for analysis.

Dimensions is a large research information system. It covers millions of research publications connected by more

[5] https://www.nltk.org/index.html

[6] https://spacy.io/

[7] https://radimrehurek.com/gensim/

[8] https://scikit-learn.org/stable/index.html

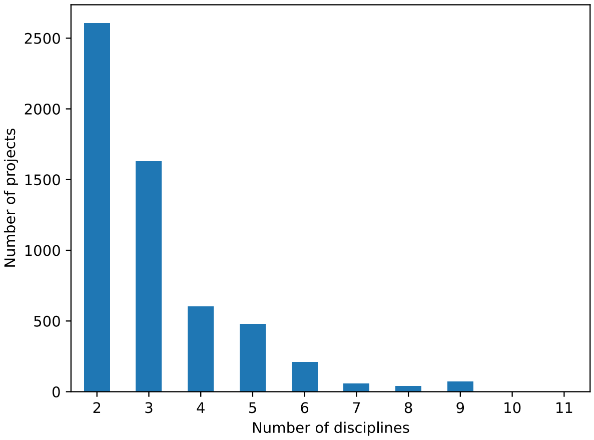


FIGURE 3. Distribution of the number of disciplines in FRIS data.

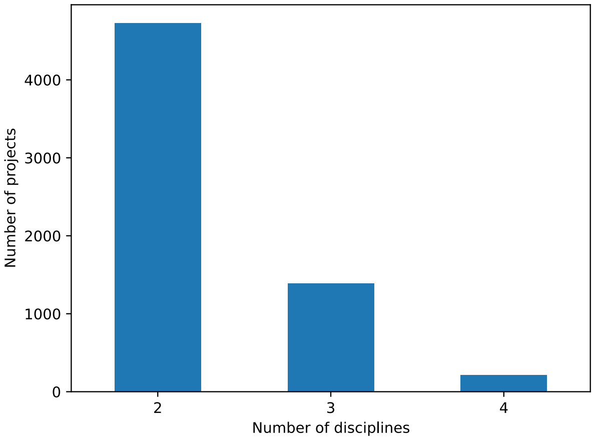


FIGURE 4. Distribution of the number of disciplines in Dimensions data.

than 1.7 billion citations, supporting grants, datasets, clinical trials, patents, and policy documents. Dimensions database applies the ANZSRC classification schema [2] to describe and classify research objects. Each grant is classified into one or more fields of research (FOR). For experimental purposes, in this study, we used disciplines of level 2 of FOR as labels for grants. To limit the size of collected data, in this work, we selected grants with the following conditions:

- funded by Belgian funding organizations,
- started in 2018 or later.

The query[9] to request grants data from Dimensions is given in appendix V-D. From the output of the query, we excluded the grants that contained only 1 FOR. As a result, 6332 records were selected.

2) DATA PREPROCESSING

We first filtered out short abstracts in datasets. The number of words in the abstracts can be determined by the users or the research information system administrators. Based on the project data, we can recommend an appropriate number to be used in such research information systems. In the experimental work and for the sake of simplicity, we filtered out projects with abstracts of less than 200 words. We then excluded projects containing a large number of disciplines. As shown in Fig. 3 and 4, most of the projects in both databases had two disciplines. If we used more than two encoded labels there would have been a high number of projects encoded with a NULL label. As a result, the data could have become imbalanced. To avoid this problem, we filtered out of projects that include more than two disciplines. After this step, there were 2606 and 4727 projects selected in FRIS and Dimensions, respectively.

In multi-label classification systems, the imbalanced number of labels in data significantly affects the performance. To further improve the quality of training data, we excluded projects that contained low frequent disciplines. In particular, for each dataset, we excluded projects that involved disciplines that appeared less than 20 times. After this step, there were 2571 and 4248 projects selected from FRIS and Dimensions, with 37 and 121 disciplines, respectively.

3) DISCIPLINES ENCODING

For each project, we converted the original discipline labels into two encoded integer labels. Particularly, each discipline label was encoded by an integer of a range from 0 to $K-1$. For example, the number of unique disciplines in FRIS data is 37, whereas in Dimension it is 121. Lists of encoded disciplines in FRIS and Dimensions are given in Appendix V-A and V-B, respectively.

4) FEATURE EXTRACTION

In this framework, we applied Top2vec [38] to extract topic probability distribution over the projects. Top2Vec is an algorithm for topic modeling and semantic search. It automatically detects topics present in the text and generates jointly embedded topics, documents, and word vectors. The advantage of Top2vec over other topic models, e.g. LDA, is that it does not require the input K number of topics. The model is able to automatically detect topics present in the text. In addition, Top2vec provides pre-trained models that can be declared by the parameter *embedding_model*. This will determine which model is used to generate the document and word embeddings. The embedding model can be chosen according to the size, language, etc. of the data.

5) DISTANCE MATRIX CREATION

In order to extract discipline probability distributions over projects and term probability distributions over disciplines, we applied the L-LDA model [40]. The L-LDA model produces a disciplines-term matrix that allowed us to calculate the distance between disciplines in data.

[9] In order to access data, we must have an account provided by Dimensions.

6) DISCIPLINE PREDICTION

In this component, we could apply one of the multi-label classification models mentioned in subsection III-B5 to build a classifier. These algorithms were implemented using the sklearn library.[10] For each algorithm, we needed to choose a set of parameters for better performance. The most commonly adjusted parameter with k Nearest Neighbor is *n_neighbors*. It regulates how many neighbors should be checked when an item is being classified. For Random Forest and Gradient Boosting, the most important parameters are *n_estimators* and *max_features*.

- *n_estimator*: number of trees inside the classifier.
- *max_features*: the number of features to consider when looking for the best split.

B. EVALUATION SETUP

In order to evaluate the performance of the machine learning classification models, we set up experiments as follows:

- Data: we used data finally obtained from the preprocessing step for evaluating and comparing the performance of the models.
- Top2vec parameters:
  - $embedding\_model\ =' \ universal - sentence - encoder'$
- L-LDA parameters:
  - $\alpha = 0.001$,
  - $\eta = 0.001$.
- Feature matrix: in addition to features produced by Top2vec, for each dataset, we created a term frequency-inverse document frequency (TF-IDF) matrix. We used TfidfVectorizer[11] from sklearn library to create TF-IDF with parameters: $n\_gram = (1, 2)$, $max\_features = 1000$. Further, we utilized Doc2Vec[12] with $vector\_size = 1000$ to extract another feature matrix. Doc2Vec was chosen because it extends Word2Vec by capturing not only word embeddings but also document-level embeddings, allowing for more comprehensive representations of texts and enabling analysis at both the word and document levels.
- Target labels: apart from encoded integer labels, we created binary labels for each target label set. To transform labels in the dataset to binary representation, we used MultiLabelBinarizer[13] from sklearn library.

  The experimental setup for classification performance evaluation is as follows:
  - Features:
    1) Doc2Vec: features extracted by Doc2Vec
    2) TF-IDF: features extracted by term frequency and inverse document frequency
    3) Topic: features extracted by Top2vec
  - Target labels:
    1) Binary: one-hot encoded binary labels
    2) Int: encoded integer labels
  - Performance measure:
    1) *CPDP*: correctly predicted discipline percentage
    2) *CPDP_D*: correctly predicted discipline percentage using the distance matrix
  - kNN parameters:
    - ∗ *n_neighbors* is ranging from 3 to 9
    - ∗ $weight =' auto'$
    - ∗ $algorithm =' auto'$
  - RF and GB parameters:
    - ∗ $n\_estimators = 100$
    - ∗ $max\_features = \sqrt{feature\_size}$
    - ∗ $random\_state = 0$
  - For other algorithms, e.g., SVM, LR, DT, ET, we used default parameters.

TABLE 4. Result of Top2vec model.

Data	#Project	#Topics	#Terms
FRIS	2571	28	512
Dimensions	4248	48	512

TABLE 5. Result of L-LDA model.

Data	#Project	#Topics	#Terms
FRIS	2571	37	31,585
Dimensions	4248	121	35,210

We first evaluated the performance of kNN with various values of *n_neighbors* in order to assess the effectiveness of features and target labels to the model. Then we compared the performance of the models on the two datasets.

C. RESULTS

1) RESULTS OF TOPIC MODELS

We first summarize the outputs of the topic models. With the given parameters, Top2vec discovered 28 topics in the FRIS dataset, whereas, for the Dimensions dataset, it found 48 topics. These topic probability distributions were used as the input features of the classifiers. The size of vectors (#Terms) was equal for both datasets, i.e., 512. The output of Top2vec is shown in Table 4. Regarding L-LDA, the numbers of topics were exactly the same as the number of disciplines in the input datasets. In particular, there were 37 and 121 topics for FRIS and Dimensions, respectively. The number of terms trained in each dataset was different, e.g., 31,585 terms for FRIS and 35,210 for Dimensions. The output of L-LDA is summarized in Table 5.

[10] https://scikit-learn.org/stable/modules/classes.html

[11] https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

[12] https://www.tutorialspoint.com/gensim/gensim_doc2vec_model.htm

[13] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MultiLabelBinarizer.html

TABLE 6. Performance of the kNN on FRIS data.

	Binary encoded labels	Integer encoded labels
Doc2Vec	29.54	42.64
TF-IDF	36.41	49.31
Topic probability	39.30	51.39

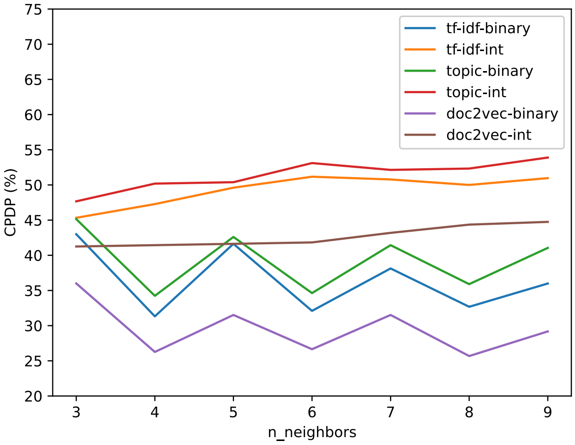


FIGURE 5. Performance of kNN with various *n_neighbors* values on FRIS data.

2) RESULTS OF MODEL EVALUATION

Experiment 1: the performance of the model on types of features and target labels

We used kNN model as a test case. The performance of the model was calculated based on the average *CPDP* of the model with the *n_neighbors* ranging from 3 to 9. For each value of *n_neighbors*, we used 10-Folk validation and repeated three times. The average performance of the kNN on the FRIS dataset is shown in Table 6. As shown, we can see that the performance of the model with topic probability was slightly better than the model with features extracted by TF-IDF and Doc2Vec. The performance of model with TF-IDF was better than that with Doc2Vec. With the same feature matrix, the performance of the model with numerical labels was better than the one with binary-encoded labels. The average *CPDP* of kNN for each value of *n_neighbors* is shown in Fig. 5. As can be seen the performance of kNN could slightly change according to the values of *n_neighbors*.

Table 7 shows the average *CPDP* of the model on the Dimensions dataset. Similar to the results on the FRIS dataset, in all cases, the model training with topic probability and integer encoded labels achieved higher *CPDP*. It is worthwhile to mention that the performance of the kNN model on the Dimensions dataset was better than it was on FRIS data. The highest *CPDP* on the Dimensions dataset was 63.39%. The average values of *CPDP* of the model with various values of *n_neighbors* are shown in Fig. 6.

To summarize, these experimental results show that the kNN model trained with topic probability and encoded integer labels achieved better *CPDP* than the model trained with features created by TF-IDF and Doc2Vec and encoded binary labels.

TABLE 7. Performance of the kNN model on Dimensions data.

	Binary encoded labels	Integer encoded labels
Doc2Vec	33.13	43.11
TF-IDF	50.71	59.75
Topic probability	56.25	63.39

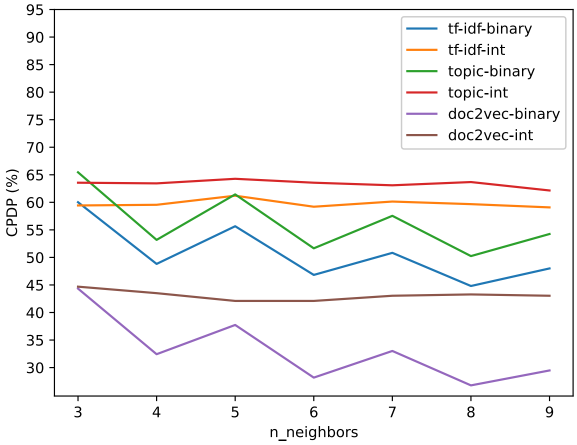


FIGURE 6. Performance of kNN with various values of *n_neighbors* on Dimensions data.

TABLE 8. Performance of classification models.

	FRIS	Dimensions
RF-D	58.12	69.87
ET-D	56.36	67.98
kNN-D	53.61	64.82
GB-D	51.82	54.23
DT-D	48.03	59.13
SVM-D	46.47	33.22
LR-D	25.34	13.43

Experiment 2: the performance of classification models

To evaluate the models mentioned in Section III-B5, we trained them with topic probability and encoded integer labels. The performance was measured by *CPDP_D* from cross-validation with 10-FOLK and repeated three times. To distinguish between a classifier with a distance matrix and a traditional classifier, we append the letter D after the name of the traditional classifier. For example, kNN-D is the kNN associated with the distance matrix.

Table 8 shows the average performance of the models. As can be seen, RF-D achieved the highest performance on both datasets. In particular, the *CPDP_D* of RF-D on the FRIS and Dimensions datasets were 58.12% and 69.87%, respectively. They were slightly higher than the performance of ET-D (56.36% for FRIS, 67.98% for Dimensions) and kNN-D (53.61% for FRIS, 64.82% for Dimensions). The performance of LR-D was the worst on both datasets (25.34% for FRIS and 13.43% for Dimensions).

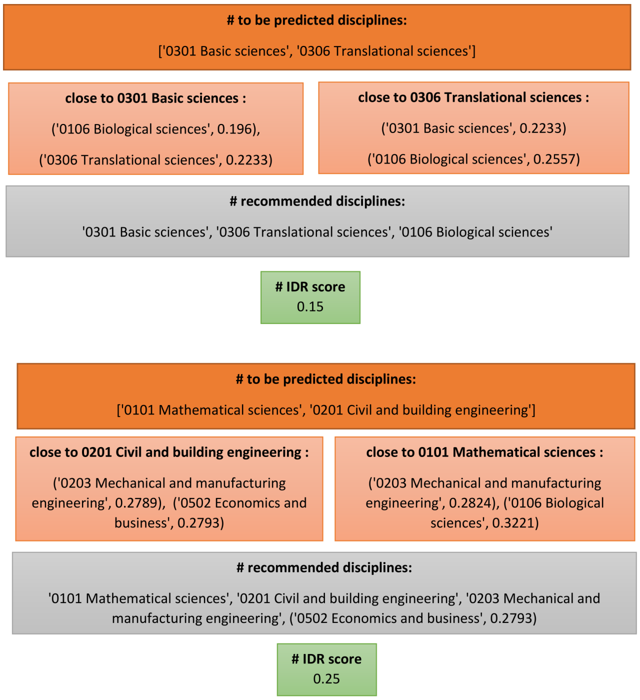

# to be predicted disciplines:
['0301 Basic sciences', '0306 Translational sciences']
close to 0301 Basic sciences :
('0106 Biological sciences', 0.196),
('0306 Translational sciences', 0.2233)
close to 0306 Translational sciences :
('0301 Basic sciences', 0.2233)
('0106 Biological sciences', 0.2557)
# recommended disciplines:
'0301 Basic sciences', '0306 Translational sciences', '0106 Biological sciences'
# IDR score
0.15
# to be predicted disciplines:
['0101 Mathematical sciences', '0201 Civil and building engineering']
close to 0201 Civil and building engineering :
('0203 Mechanical and manufacturing engineering', 0.2789), ('0502 Economics and business', 0.2793)
close to 0101 Mathematical sciences :
('0203 Mechanical and manufacturing engineering', 0.2824), ('0106 Biological sciences', 0.3221)
# recommended disciplines:
'0101 Mathematical sciences', '0201 Civil and building engineering', '0203 Mechanical and manufacturing engineering', ('0502 Economics and business', 0.2793)
# IDR score
0.25

FIGURE 7. System recommendation examples for FRIS Data.

Based on the training models, the recommendation algorithm was able to discover the most relevant disciplines of a project.

Given textual data of an unseen project, the discipline recommendation algorithm outputs results as follows:

- # predicted disciplines: the list of disciplines predicted by the machine learning (ML) classifier.
- # recommended disciplines based on distance matrix: the list of recommended disciplines. For each predicted discipline, the system recommends the two closest disciplines according to distances.
- # IDR score: the discipline diversity calculated by the Rao-Stirling diversity index.

Due to space limitations, we only display a few examples of the discipline recommendation algorithm. Fig. 7 shows two examples of recommending disciplines to two projects in the FRIS dataset. In the first example, the ML classifier predicted two disciplines: '0301 Basic sciences', '0306 Translational sciences'. Based on the distance matrix, we found other disciplines which were close to them. We can notice that the close disciplines are duplicated. The reason is that '0301 Basic sciences' and '0306 Translational sciences' are very similar in practice. As a result, they are close to the same disciplines. After filtering out duplicates, the disciplines recommended to the user were '0301 Basic sciences', '0306 Translational sciences', '0106 Biological sciences'. With these three disciplines, the diversity score (#IDR score) was 0.15.

In the case of the second example, the predicted disciplines were '0101 Mathematical sciences' and '0201 Civil and building engineering'. These two disciplines were far from each other in practice. Based on predicted disciplines and close disciplines the algorithm recommended four disciplines to the users. Since these recommended disciplines were different in terms of the research fields, the IDR score (0.25) was higher than that in the first example.

We can observe from the FRIS dataset and distance matrix that discipline '0301 Basic sciences' occurred in 600 projects and had 336 projects in common with discipline '0306 Translational sciences'. The second most similar discipline to '0301 Basic sciences' was '0106 Biological sciences', which occurred in 86 projects with '0301 Basic sciences'. Furthermore, the third closest discipline, '0302 Clinical sciences', had 74 projects in common with '0301 Basic sciences', according to the distance matrix. Similarly from the obtained results, we can see that when the distance between two disciplines increases, the number of common projects decreases.

3) COMPARISON TO RELATED WORK

A typical approach to compare our research findings to related work is by evaluating the performance of the proposed machine learning models based on the commonly used evaluation metrics such as accuracy, precision, recall, and F1-score on various datasets. This approach is widely adopted in almost every research in the area of machine learning models design [45], [46], [47], [48]. In a similar manner, we compared the proposed approach to related work by applying them to the same dataset and using the same metric to evaluate their performance. However, in the proposed model, the labels are encoded as numbers, some factors such as True negative, False positive, and False negative are not applicable. As a result, these metrics: accuracy, precision, recall, F1-score could not be applied to evaluate the performance of our proposed framework. Therefore, we proposed a new metric to count the true positive values in a way suitable for the context of our framework. We are confident that our approach is appropriate for this paper.

To compare the proposed approach to other studies, we first selected the typical ML classification models employed in those studies. The algorithms were then performed on the same data using two approaches: one traditional way that utilized TF-IDF and binary labels, and one that employed topic probability distribution and the proposed encoded numerical labels. The performance of the algorithms was measured by *CPDP* and *CPDP_D*.

We first compared the performance of the proposed approach to the approaches used in [14]. In order to do that, we ran three traditional classification models: RF, DT, and ET on FRIS data. The performance of the models was measured

TABLE 9. Comparison to [14].

RF-D	RF [14]	DT-D	DT [14]	ET-D	ET [14]
58.12	18.49	48.03	35.87	58.36	16.93

by *CPDP* before and after using the distance matrix. The comparison result is shown in Table 9. As can be seen, the performance of the models based on the proposed approach outperforms the models of [14]. In all cases, the performance of the proposed approach classifiers achieved a better *CPDP* score.

Besides, the performance of the proposed approach was compared to that of the classification models used in [1]. In [1], besides using traditional classification models such as SVM, and LR, the authors applied a modified character-based convolutional deep neural network to classify research documents by fields of study (disciplines). For the sake of comparison only, in this paper, we considered the improvement of [1] when using deep learning over the same traditional classification models as a means of comparison. The improvements of [1] over the SVM and LR were 24.21% and 28.61%, respectively. With the proposed approach, these two classifiers: SVM-D, and LR-D achieved a higher improvement than the improvement of [1]. In particular, the improvements of SVM-D, LR-D over SVM, and LR on FRIS data were 32.61% and 42.7%, respectively.

It is important to acknowledge that the models' performance was not optimal. This is attributed to the inherent difficulties associated with predicting multiple disciplines pertaining to research projects. For instance, in [5], although the problem was treated as a multiclass classification task where each publication was assigned only one discipline, the highest reported F1-score achieved was 80%. In our study, we tackle a more challenging task of multi-label classification. The results we obtained align with those reported in related studies [1], [14]. We have already conducted a comparative analysis of our approach against similar works. There are several possible reasons why the results may not be highly accurate. Firstly, the research disciplines used in the database may be very similar, making it challenging for the classifier to distinguish between them correctly. Secondly, the distribution of research disciplines in the projects may be imbalanced, which can affect the accuracy of the models. Another factor that may have impacted the performance is the quality of the data. For instance, in some databases like FRIS, research disciplines are assigned to projects by research administrators due to time constraints of principal researchers, which can lead to inconsistencies in discipline assignments.

D. THREATS TO VALIDITY

There are several validity concerns regarding the framework proposed in this study. First of all, the datasets used in this study were obtained by querying project metadata containing titles, keywords, and abstract with certain conditions. It is possible that the dataset is not representative of all project data in databases and may contain bias towards certain research disciplines or demographics. In future work, it is important to collect data that not only has high volume but also high quality to void this bias issue.

The performance of the machine learning models was evaluated using CPDP. However, this metric does not capture the trade-off between false positives and false negatives, which may be important in certain applications. In addition, the models were trained using a specific set of hyperparameters. There may be other hyperparameter configurations that yield better performance.

Another threat to validity is that the machine learning models were trained and evaluated on specific datasets and may not generalize well to other datasets which do not have labeled data. Our research predicts research disciplines using machine learning and distance metrics. Combining and applying these methods to data from two real systems is the important contribution. A conceptual framework with several components is implemented in a simulated environment on real data. Thus, our experimental approach is not to extend our technique to fit all datasets or systems, but to design and evaluate a solution to an actual challenge for research project metadata systems like FRIS. In that sense, we compared our technique to similar work within this context of the two systems to demonstrate our contribution and utility in this simulated environment, not to conclude that our approach works on other research project systems.

V. CONCLUSION

In this paper, we proposed a generic framework not only for multiple disciplines prediction but also for interdisciplinarity calculation. The proposed framework consisted of a number of components which combined different machine learning techniques and distance metrics to find the most relevant research disciplines to a research project based on its textual description metadata. To the best of our knowledge, this work is the first to apply distance metrics to improve research discipline prediction. We evaluated the proposed framework on two different scientific databases; FRIS and Dimensions. Empirical results show that in the proposed framework, the ML classifier performs better than conventional approaches like TF-IDF as features and binary encoded labels as output variables. Further, the proposed approach was found to outperform related work.

This study has some limitations, which can be considered in the future. First of all, the size of the experimental data is limited. i.e. only 2571 records from FRIS and 4248 records from Dimensions were considered. The limited size of the dataset may have affected the performance of the ML classification models to a certain extent. The second limitation is that the distance threshold used to determine close disciplines is selected experimentally. For each predicted discipline we select two disciplines that are closest to it. Finally, the outputs of the topic models change every time. This can affect

TABLE 10. Discipline code table for FRIS dataset.

Discipline label	Encoded number
0101 Mathematical sciences	0
0102 Information and computing sciences	1
0103 Physical sciences	2
0104 Chemical sciences	3
0105 Earth sciences	4
0106 Biological sciences	5
0107 Environmental sciences	6
0199 Other natural sciences	7
0201 Civil and building engineering	8
0202 Electrical and electronic engineering	9
0203 Mechanical and manufacturing engineering	10
0204 (Bio)chemical engineering	11
0205 Materials engineering	12
0206 (Bio)medical engineering	13
0207 Biotechnology, bio-engineering and biosystems engineering	14
0208 Computer engineering, information technology and mathematical engineering	15
0299 Other engineering and technology	16
0301 Basic sciences	17
0302 Clinical sciences	18
0303 Health sciences	19
0304 Paramedical sciences	20
0305 Pharmaceutical sciences	21
0306 Translational sciences	22
0401 Agriculture, forestry, fisheries and allied sciences	23
0402 Veterinary sciences	24
0501 Psychology and cognitive sciences	25
0502 Economics and business	26
0503 Pedagogical and educational sciences	27
0504 Sociology and anthropology	28
0505 Law and legal studies	29
0506 Political sciences	30
0508 Media and communications	31
0599 Other social sciences	32
0601 History and archaeology	33
0602 Languages and literary studies	34
0603 Philosophy, ethics and religious studies	35
0604 Arts	36

the ML classification models since they use the output of the topic model as the input. In this paper, we excluded projects with low frequent disciplines which means that the system does not predict those disciplines at the moment. We expect with using the system more data will be added and accordingly we will update the system to include associated disciplines.

Besides, different research directions can be considered for future work. First of all, investigating methods to improve data quality as well for training ML classification models is the most essential task. With high-quality data, e.g, abstracts containing rich information, labels correctly assigned, labels balanced, etc., the ML classification model performance can improve. Improving the quality of the distance matrix through applying more sophisticated distance metric techniques is another essential research. This matrix plays an important role in finding close disciplines and in calculating interdisciplinarity. Optimizing the number of topics or controlling the change of topics produced by topic models is also important. This can help improve the performance of the ML classification model as well as make the recommendation more reliable.

ACKNOWLEDGMENT

The authors would like to thank Dimensions for granting access to their database.

APPENDIX

A. DISCIPLINE CODE TABLE FOR FRIS DATASET

See Table 10.

B. DISCIPLINE CODE TABLE FOR DIMENSIONS DATASET

See Table 11.

C. EXTRACT PROJECT DATA FROM FRIS

```
def fetch_from_service(url, headers,
    body, max_pages, destination):
with requests.Session() as session:
page = 0
while page < max_pages:
body_for_this_page = body % page
response = session.post(url, data=
    body_for_this_page, headers=headers)
f = open(destination % page, 'wt')
f.write(response.text)
```

TABLE 11. Discipline code table for Dimensions dataset.

Discipline label	Encoded number	Discipline label	Encoded number
3004 Crop and Pasture Production	0	4101 Climate Change Impacts and Adaptation	61
3006 Food Sciences	1	4102 Ecological Applications	62
3007 Forestry Sciences	2	4104 Environmental Management	63
3101 Biochemistry and Cell Biology	3	4105 Pollution and Contamination	64
3102 Bioinformatics and Computational Biology	4	4201 Allied Health and Rehabilitation Science	65
3103 Ecology	5	4203 Health Services and Systems	66
3104 Evolutionary Biology	6	4205 Nursing	67
3105 Genetics	7	4206 Public Health	68
3106 Industrial Biotechnology	8	4207 Sports Science and Exercise	69
3107 Microbiology	9	4301 Archaeology	70
3108 Plant Biology	10	4302 Heritage, Archive and Museum Studies	71
3109 Zoology	11	4303 Historical Studies	72
3201 Cardiovascular Medicine and Haematology	12	4402 Criminology	73
3202 Clinical Sciences	13	4403 Demography	74
3204 Immunology	14	4404 Development Studies	75
3205 Medical Biochemistry and Metabolomics	15	4405 Gender Studies	76
3206 Medical Biotechnology	16	4406 Human Geography	77
3207 Medical Microbiology	17	4407 Policy and Administration	78
3208 Medical Physiology	18	4408 Political Science	79
3209 Neurosciences	19	4410 Sociology	80
3211 Oncology and Carcinogenesis	20	4602 Artificial Intelligence	81
3214 Pharmacology and Pharmaceutical Sciences	21	4603 Computer Vision and Multimedia Computation	82
3215 Reproductive Medicine	22	4604 Cybersecurity and Privacy	83
3301 Architecture	23	4605 Data Management and Data Science	84
3302 Building	24	4606 Distributed Computing and Systems Software	85
3303 Design	25	4608 Human-Centred Computing	86
3304 Urban and Regional Planning	26	4611 Machine Learning	87
3401 Analytical Chemistry	27	4612 Software Engineering	88
3402 Inorganic Chemistry	28	4613 Theory Of Computation	89
3403 Macromolecular and Materials Chemistry	29	4701 Communication and Media Studies	90
3404 Medicinal and Biomolecular Chemistry	30	4702 Cultural Studies	91
3405 Organic Chemistry	31	4703 Language Studies	92
3406 Physical Chemistry	32	4704 Linguistics	93
3502 Banking, Finance and Investment	33	4705 Literary Studies	94
3507 Strategy, Management and Organisational Behaviour	34	4801 Commercial Law	95
3509 Transportation, Logistics and Supply Chains	35	4803 International and Comparative Law	96
3601 Art History, Theory and Criticism	36	4804 Law In Context	97
3605 Screen and Digital Media	37	4805 Legal Systems	98
3701 Atmospheric Sciences	38	4806 Private Law and Civil Obligations	99
3705 Geology	39	4807 Public Law	100
3707 Hydrology	40	4901 Applied Mathematics	101
3708 Oceanography	41	4902 Mathematical Physics	102
3709 Physical Geography and Environmental Geoscience	42	4904 Pure Mathematics	103
3801 Applied Economics	43	5001 Applied Ethics	104
3802 Econometrics	44	5002 History and Philosophy Of Specific Fields	105
3803 Economic Theory	45	5003 Philosophy	106
3901 Curriculum and Pedagogy	46	5004 Religious Studies	107
3903 Education Systems	47	5005 Theology	108
4001 Aerospace Engineering	48	5101 Astronomical Sciences	109
4003 Biomedical Engineering	49	5102 Atomic, Molecular and Optical Physics	110
4004 Chemical Engineering	50	5104 Condensed Matter Physics	111
4005 Civil Engineering	51	5106 Nuclear and Plasma Physics	112
4006 Communications Engineering	52	5107 Particle and High Energy Physics	113
4007 Control Engineering, Mechatronics and Robotics	53	5108 Quantum Physics	114
4008 Electrical Engineering	54	5109 Space Sciences	115
4009 Electronics, Sensors and Digital Hardware	55	5110 Synchrotrons and Accelerators	116
4010 Engineering Practice and Education	56	5201 Applied and Developmental Psychology	117
4011 Environmental Engineering	57	5202 Biological Psychology	118
4012 Fluid Mechanics and Thermal Engineering	58	5204 Cognitive and Computational Psychology	119
4016 Materials Engineering	59	5205 Social and Personality Psychology	120
4018 Nanotechnology	60		

```
f.close()

def fetch_projects():
url = "https://frisr4.researchportal.be/
    ws/ProjectService"
```

```
headers = {"content-type": "application/
    xml"}
body = """
<soap:Envelope xmlns:soap="http://
    schemas.xmlsoap.org/soap/envelope/">
```

```
<soap:Body>
<ns1:getProjects xmlns:ns1="http://fris.ewi.be/">
<projectCriteria xmlns="http://fris.ewi.be/criteria">
<window>
<pageSize>1000</pageSize>
<pageNumber>%s</pageNumber>
<orderings>
<order>
<id>entity.created</id>
<direction>DESCENDING</direction>
</order>
</orderings>
</window>
</projectCriteria>
</ns1:getProjects>
</soap:Body>
</soap:Envelope>
"""
fetch_from_service(url, headers, body,
    max_pages=50, destination='project_data/page%s.xml')
```

D. EXTRACT PROJECT DATA FROM DIMENSIONS

```
query = f"""
search grants
where researchers is not empty and
    start_year >= {start_year}
and funder_org_name in {funders}
and research_org_countries.name={country
    }
return grants[id+title+abstract+
    category_for_2020+concepts_scores]

res = dsl.query_iterative(query)
"""
```

REFERENCES

[1] 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, May 2021, Art. no. e0251493.

[2] *Australian and New Zealand Standard Research Classification (ANZSRC)*, Australian Bureau Statist., Australia, 2020.

[3] S. Vancauwenbergh and H. Poelmans, "The flemish research discipline classification standard: A practical approach," *Knowl. Org.*, vol. 46, no. 5, pp. 354–363, 2019.

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

[5] J. Eykens, R. Guns, and T. Engels, "Article level classification of publications in sociology: An experimental assessment of supervised machine learning approaches," in *Proc. 17th Int. Conf. Scientometrics Informetrics*, 2019, pp. 1–12.

[6] D.-T. Vo and C.-Y. Ock, "Learning to classify short text from scientific documents using topic models with various types of knowledge," *Expert Syst. Appl.*, vol. 42, no. 3, pp. 1684–1698, Feb. 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417414005764

[7] M. M. Mirończuk and J. Protasiewicz, "A recent overview of the state-of-the-art elements of text classification," *Expert Syst. Appl.*, vol. 106, pp. 36–54, Sep. 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S095741741830215X

[8] P. Raento, "Interdisciplinarity," in *International Encyclopedia of Human Geography*. Amsterdam, The Netherlands: Elsevier, 2020, pp. 357–363.

[9] W. Glänzel and K. Debackere, "Various aspects of interdisciplinarity in research and how to quantify and measure those," *Scientometrics*, vol. 127, no. 9, pp. 5551–5569, Sep. 2021.

[10] I. Rafols and M. Meyer, "Diversity and network coherence as indicators of interdisciplinarity: Case studies in bionanoscience," *Scientometrics*, vol. 82, no. 2, pp. 263–287, Jun. 2009.

[11] J. Adams, T. Loach, and M. Szomszor, "Digital research report: Interdisciplinary research - methodologies for identification and assessment," Digital Sci., London, U.K., Tech. Rep., 2016. Accessed: Jan. 12, 2022. [Online]. Available: https://www.digital-science.com/resource/methodologies-for-identification-and-assessment/

[12] L. Zhang, B. Sun, Z. Chinchilla-Rodríguez, L. Chen, and Y. Huang, "Interdisciplinarity and collaboration: On the relationship between disciplinary diversity in departmental affiliations and reference lists," *Scientometrics*, vol. 117, no. 1, pp. 271–291, Jul. 2018.

[13] Q. Wang and J. Schneider, "Consistency of interdisciplinarity measures," 2018, *arXiv:1810.00577*.

[14] T. Weber, D. Kranzlmüller, M. Fromm, and N. T. de Sousa, "Using supervised learning to classify metadata of research data by field of study," *Quant. Sci. Stud.*, vol. 2020, pp. 1–26, May 2020.

[15] FRIS. *Flanders Research Information Space*. Accessed: Jan. 10, 2022. [Online]. Available: https://researchportal.be/en

[16] *Dimensions*. Accessed: Jan. 10, 2022. [Online]. Available: https://www.dimensions.ai/

[17] K. W. Boyack and R. Klavans, "Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately?" *J. Amer. Soc. Inf. Sci. Technol.*, vol. 61, no. 12, pp. 2389–2404, Dec. 2010.

[18] S. Xu, Y. Li, and Z. Wang, "Bayesian multinomial Naïve Bayes classifier to text classification," in *Advanced Multimedia and Ubiquitous Engineering* (Lecture Notes in Electrical Engineering). Singapore: Springer, 2017, pp. 347–352.

[19] K. P. Bennett and C. Campbell, "Support vector machines: Hype or hallelujah?" *ACM SIGKDD Explor. Newslett.*, vol. 2, no. 2, pp. 1–13, 2000, doi: 10.1145/380995.380999.

[20] T. K. Ho, "Random decision forests," in *Proc. 3rd Int. Conf. Document Anal. Recognit.*, 1995, p. 278.

[21] L. Mason, J. Baxter, P. Bartlett, and M. Frean, "Boosting algorithms as gradient descent," in *Proc. 12th Int. Conf. Neural Inf. Process. Syst.* Cambridge, MA, USA: MIT Press, 1999, pp. 512–518.

[22] A. L. Porter, A. S. Cohen, J. David Roessner, and M. Perreault, "Measuring researcher interdisciplinarity," *Scientometrics*, vol. 72, no. 1, pp. 117–147, Jun. 2007.

[23] G. Abramo, C. A. D'Angelo, and F. Di Costa, "Identifying interdisciplinarity through the disciplinary classification of coauthors of scientific publications," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 63, no. 11, pp. 2206–2222, Nov. 2012, doi: 10.1002/asi.22647.

[24] Z. Ba, Y. Cao, J. Mao, and G. Li, "A hierarchical approach to analyzing knowledge integration between two fields—A case study on medical informatics and computer science," *Scientometrics*, vol. 119, no. 3, pp. 1455–1486, May 2019.

[25] L. G. Nichols, "A topic model approach to measuring interdisciplinarity at the national science foundation," *Scientometrics*, vol. 100, no. 3, pp. 741–754, May 2014.

[26] H. Xu, T. Guo, Z. Yue, L. Ru, and S. Fang, "Interdisciplinary topics of information science: A study based on the terms interdisciplinarity index series," *Scientometrics*, vol. 106, no. 2, pp. 583–601, Jan. 2016.

[27] A. Bonaccorsi, N. Melluso, and F. A. Massucci, "Detecting interdisciplinarity in top-class research using topic modeling," in *Proc. ISSI*, vol. 42, W. Glänzel, S. Heeffer, P.-S. Chi, and R. Rousseau, Eds. Heidelberg, Germany: Springer, 2021, pp. 160–169.

[28] M. Suhaidi, R. Abdul Kadir, and S. Tiun, "A review of feature extraction methods on machine learning," *J. Inf. Syst. Technol. Manage.*, vol. 6, no. 22, pp. 51–59, Sep. 2021.

[29] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proc. 31st Int. Conf. Int. Conf. Mach. Learn.*, vol. 32, 2014, pp. II-1188–II-1196.

[30] C.-C.-J. Kuo, "Understanding convolutional neural networks with a mathematical model," *J. Vis. Commun. Image Represent.*, vol. 41, pp. 406–413, Nov. 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1047320316302267

[31] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Inf. Process. Manage.*, vol. 24, no. 5, pp. 513–523, Jan. 1988. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0306457388900210

[32] E. Simpson, "Measurement of diversity," *Nature*, vol. 163, no. 4148, p. 688, 1949.

[33] S. Ortiz-Burgos, *Shannon-Weaver Diversity Index*. Dordrecht, The Netherlands: Springer, 2016, pp. 572–573, doi: 10.1007/978-94-017-8801-4_233.

[34] A. Stirling, "A general framework for analysing diversity in science, technology and society," *J. Roy. Soc. Interface*, vol. 4, no. 15, pp. 707–719, Feb. 2007.

[35] S. K. Ashenden, A. Bartosik, P.-M. Agapow, and E. Semenova, "Introduction to artificial intelligence and machine learning," in *The Era of Artificial Intelligence, Machine Learning, and Data Science in the Pharmaceutical Industry*. Amsterdam, The Netherlands: Elsevier, 2021, pp. 15–26.

[36] J. T. Hancock and T. M. Khoshgoftaar, "Survey on categorical data for neural networks," *J. Big Data*, vol. 7, no. 1, pp. 1–10, Apr. 2020.

[37] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

[38] D. Angelov, "Top2 Vec: Distributed representations of topics," 2020, *arXiv:2008.09470*.

[39] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," 2022, *arXiv:2203.05794*.

[40] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning, "Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2009, pp. 248–256. [Online]. Available: https://aclanthology.org/D09-1026

[41] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory*, vol. IT-13, no. 1, pp. 21–27, Jan. 1967.

[42] L. Rokach and O. Maimon, "Decision trees," in *Data Mining and Knowledge Discovery Handbook*. Cham, Switzerland: Springer, 2005, pp. 165–192.

[43] D. R. Cox, "The regression analysis of binary sequences," *J. Roy. Stat. Soc., B, Methodol.*, vol. 20, no. 2, pp. 215–232, Jul. 1958, doi: 10.1111/j.2517-6161.1958.tb00292.x.

[44] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Mach. Learn.*, vol. 63, no. 1, pp. 3–42, Mar. 2006.

[45] L. Romeo, R. Marani, T. D'Orazio, and G. Cicirelli, "Video based mobility monitoring of elderly people using deep learning models," *IEEE Access*, vol. 11, pp. 2804–2819, 2023.

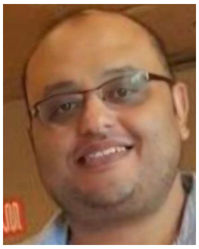
[46] P. R. Mendes, L. Bondi, P. Bestagini, S. Tubaro, and A. Rocha, "An in-depth study on open-set camera model identification," *IEEE Access*, vol. 7, pp. 180713–180726, 2019.

[47] S. Rajora, D. Li, C. Jha, N. Bharill, O. P. Patel, S. Joshi, D. Puthal, and M. Prasad, "A comparative study of machine learning techniques for credit card fraud detection based on time variance," in *Proc. IEEE Symp. Ser. Comput. Intell. (SSCI)*, Nov. 2018, pp. 1958–1963.

[48] I. K. Nti, A. F. Adekoya, and B. A. Weyori, "A comprehensive evaluation of ensemble learning for stock-market prediction," *J. Big Data*, vol. 7, no. 1, pp. 1–12, Mar. 2020.

HOANG-SON PHAM received the Engineering degree in information technology and the M.S. degree in information systems from Can Tho University, Can Tho, Vietnam, in 2004 and 2013, respectively, and the Ph.D. degree in computer sciences from Rennes 1 University, Rennes, France, in 2017. From 2018 to 2020, he was a Postdoctoral Researcher with the ICTEAM, UCLouvain, Belgium where he was developing pattern mining algorithms for analyzing legacy software systems. Since 2021, he has been a Postdoctoral Researcher with ECOOM, Hasselt University. His research interests include AI, mainly focusing on developing unsupervised learning algorithms, natural language processing, and applying AI to solve complex real-life issues.

HANNE POELMANS was born in Tongeren, Belgium, in 1982. She received the B.S. and M.S. degrees in biological psychology from Maastricht University, The Netherlands, in 2004, the M.S. degree in clinical psychology from Ghent University, Belgium, in 2005, and the Ph.D. degree in medicine from KU Leuven, Belgium, in 2012. From 2005 to 2007, she was a Research Assistant with Maastricht University. From 2012 to 2014, she was a Postdoctoral Researcher with KU Leuven. In 2015, she joined the Directorate Research, Library and International Office, Hasselt University, Belgium, where she is currently a Coordinator of the Information Management and Strategic Data-Analysis Team. Her research interests include research information systems, research classifications, open science, research assessment, indicator development, and university rankings.



AMR ALI-ELDIN (Senior Member, IEEE) received the B.Sc. degree in electronics engineering and the M.Sc. degree in automatic control engineering from Mansoura University, Mansoura, Egypt, in 1997 and 2001, respectively, and the Ph.D. degree in computer and information systems from the Delft University of Technology, The Netherlands, in 2006. He has been with ECOOM, Hasselt University, as a Senior Researcher, since August 2022. Further, he is currently affiliated with Mansoura University, as an Associate Professor. Over the last 20 years, he gained wide international experience working in information and communication technology consultancy and academia for a number of international companies and universities. His research interests include research information systems, data science, semantic interoperability, and software engineering.