DESIDOC Journal of Library & Information Technology, Vol. 42, No. 3, May 2022, pp. 154-167, DOI : 10.14429/djlit.42.3.17554

Analysing and Examining Taxonomy and Folksonomy Terms in the Hybrid Subject Device using Machine Learning Techniques

Swarnali Chatterjee[#] and Rajesh Das[$,*]

[#]*Bengal Music College, Kolkata - 700 029, India*

[$]*Department of Library and Information Science, The University of Burdwan, Burdwan - 713 104, India*

[*]*E-mail: rajeshdas99@gmail.com*

ABSTRACT

The information retrieval system contains either a list of subject terms (taxonomy) or a list of collaborative tags (folksonomy) or both. The taxonomy and folksonomy come together as called hybrid subject devices. The main purpose of this paper is to apply machine learning techniques in the dataset from the library domain like others and analyse a large quantity of data for critical problems with accuracy. This research reveals to perform EDA (Exploratory data analysis), prediction analysis, and similarity measurement between folksonomy and taxonomy terms with new emerging technologies. Data science deals with big data that means unstructured data, messy data, a large volume of data. The size is of a large amount of data in terms of GB, TB. Machine learning tools manage this type of data. Usually, the Excel, or other spreadsheets package could not manage the file size in GB or TB, and that's why ML tools, and techniques are applied. At present, the library science domain also contains a large amount of data like 20/30 years of circulation data or subject descriptors, collaborative tags etc. Library professionals can apply machine learning tools for analysing this kind of data in the library domain. In this paper, the authors have introduced the applications of tools and techniques in the library domain and they have tested with 2642 taxonomy and folksonomy terms. This research work includes – EDA, prediction analysis, and similarity measurement of a folksonomy and taxonomy dataset. In the EDA part, the research work has performed a lot of analysis that includes frequency of LCSH (Library of Congress Subject Heading - taxonomy) terms, pair plots, joint plots, and heat map of LCSH and folksonomy terms. The logistic regression (LR) model for prediction analysis has been used in the folksonomy and taxonomy dataset. These 2642 terms of folksonomy and taxonomy both terms are taken as data for this research work. The EDA has been performed with the attributes in the dataset. The accuracy value of logistic regression (f1- score) is 0.37 at the training percentage of 69. The percentage of similarity between LCSH terms and folksonomy terms is 30 per cent (0.30151134), and the angle between these two vectors is 27 degrees. The novelty of this research work is that library data has been analysed using machine learning techniques the ever used before.

Keywords: Folksonomy; Taxonomy; Information retrieval; Machine learning techniques; Similarity measurement

1. INTRODUCTION

Generally, a subject device is constructed by a controlled vocabulary device or a standard schedule of terms in a digital library environment[1]. Dr S.R. Ranganathan defined the subject device in his book entitled "Prolegomena to Library Classification" for both idea plane and notational plane. In the idea plane the subject device belongs to focal ideas in an array. The subject device is implemented on the basis of subject characteristics in the idea plane. In the notational plane, the subject device of the idea plane is implemented by using a class number as the focal number in an array[2]. In both planes, the subject device refers to a controlled vocabulary. There are many controlled vocabulary devices like LCSH, Sear's List of Subject Headings, UNESCO thesaurus, classification schedules like Dewey Decimal Classification (DDC), Universal Decimal Classification (UDC), Library of Congress Classification (LCC) in the field of library and Information science. The taxonomy refers to a controlled vocabulary device that must be generated by a single or group of experts from a particular subject domain. Other hand, folksonomy refers to an uncontrolled vocabulary device containing a list of terms contributed by non-experts in that domain[3]. The folksonomy is collaborative tags/keywords/terms that are enriched by the general users[4].

The subject device is the key pillar of an Information Retrieval System (IRS)[5] in a traditional library and digital environment. The subject device contains the list of terms from a controlled vocabulary, whereas the hybrid subject device[6] contains the list of terms from the controlled vocabulary as well as uncontrolled vocabulary. When a subject device is populated in a library management system or digital library system or any other digital applications for libraries, a taxonomy (controlled vocabulary device) must follow[7]. But in a hybrid subject device, both taxonomy and folksonomy (uncontrolled vocabulary device- users' collaborative tags) are followed. The hybrid subject device is nothing, but it allows users to

Received : 02 October 2021, Revised : 03 January 2022
Accepted : 04 March 2022, Online published : 25 April 2022

contribute their tags to an existing subject device. This research work has been done with two different domains, i.e., the first is the subject device that belongs to an information retrieval system which is from library and information science[8]. The second is data analysis which belongs to Machine Learning (ML) which is from the data science domain[9]. This research work was done on an existing hybrid subject device that contains 2642 terms from both LCSH and folksonomy terms, including RT (related terms), NT (narrower terms), RS (related subjects), BT (broader terms) in a digital library environment. Three major tasks have been performed in the research work - (i) exploratory data analysis on above said data, (ii) prediction analysis of folksonomy and taxonomy dataset, and (iii) cosine similarity measurement between LCSH terms and folksonomy terms by using ML techniques.

2. BACKGROUND OF THE STUDY

The information retrieval system plays an important role to access information in a digital library environment as well as a traditional library. It consists of a controlled vocabulary device that refers to taxonomy. With the advent of Web 2.0, the concept of collaborating tagging has been introduced in a web-based system. The collaborating tagging is nothing but folksonomy that is a completely uncontrolled vocabulary device[10]. The combination of taxonomy and folksonomy may increase the efficiency of the information retrieval system. Before the incorporation of folksonomy terms in an information retrieval system, we should analyse, examine and measure similarities between taxonomy and folksonomy terms in a digital library environment. Some related research works have been done in this area that is mentioned in the next section.

3. LITERATURE REVIEW

Some recent related literature were studied in ML techniques in library and other fields, and they are discussed below:

Virkus and Garoufallou[11] established a relationship between data science and library and information science, and they performed a content analysis of research publications on data science was made of papers published on the "Web of Science" database to identify the main things discussed in the publications from the LIS (Library and Information Science) perspective. The authors have taken 80 publications for the work and divided into six categories- data science; education and training; knowledge and skills of the data professional; the role of libraries and librarians; data science measurement tools, techniques; application of data science; data science from the knowledge management perspective; and the data science from the perspectives of health science. The 80 publications were kept in those categories. The category of tools, techniques, and application of data science was most addressed by the authors. Daimari *et al.*,[12] have developed a system, which can predict the possible availability of the issued books. The users are from the library of the Central Institute of Technology Kokrajhar. The authors have predicted the date of books availability by using machine learning techniques. Random forest, support vector machine (SVM), and neural network are used, and the resulting trend is compared using 'Keres' and 'S.K. Learn'. Vaidy and Harinarayana[13] wrote an article on the role of social tags in web resources discovery. They have discussed social tags and library subject heading terms and made a comparison study of how they are matched and how they are non-matched. They identified the frequencies of social tags and LCSH terms. They applied the co-sign similarity measurement techniques in the social tags and LCSH terms datasets and found the unit of similarity and also the distance the edges of data. Their work may be included with ML models. But they applied the formula of cosine similarity measurement usually, and they limited the data range of social tags and LCSH terms.

Choi and Choi[14] have discussed in their research work on prediction analysis using ML techniques. Their work aimed to develop a reliable prediction model for job involvement in the H.R. management of employees using ML techniques. This is very much useful in the H.R. Management of an organisation. The top-level authority of an organisation could measure and predict the job involvement of the employees of their organisation. ML methods and models are based on mathematics and statistics. Malhotra and others[15] have done work on customer loans in the banking field. They have evaluated the customer loans using ML techniques. They have applied the decision trees and SVMs to identify potential bad loans. They have shown the results of various techniques and compared the good credit clients and bad credit clients.

In the Medical Science field, there is so much literature available as closed access and open access regarding machine learning techniques. One of the most important subdomains is "Breast Cancer" from Medical Science. Jaison and others[16] wrote an article on the detection of breast cancer using ML techniques, and they applied various models like- Naive Bayes, Random Forest, KNN, and SVMs to find the malignant and non-malignant tumours and predict the trend of breast cancer. They also tested their models as to how much it was accurate by classification methods. Another work-related on breast cancer machine learning technique was published as a journal article by the author Seid Hassan Yesuf, and he also applied various Machine Learning models in his breast cancer data Sets. As a result, he also found 97.6 per cent - 98.8 per cent of the accuracy of the models by the classification method[17].

4. STATEMENT OF THE PROBLEM

The information retrieval is built to retrieve organised resources from an information system. The indexing process and subject device play a vital role in accessing the resources. An existing hybrid subject device is used for research this research work. The hybrid subject device contains standard subject terms (LCSH terms – taxonomy terms) and users' collaborative tags (folksonomy terms). The standard subject terms are populated by library professionals, and users' collaborative terms are populated by the users. The result of such a population of terms is reflected as a hybrid subject device as well as the bulky subject device. The hybrid subject device will be an efficient information retrieval if the users' approach terms (folksonomy) and subject terms are different from each other's. Otherwise, if they are very closed to each other, then

the subject device is to be a less efficient information retrieval system. Given this truth, some questions are raised.

4.1 Primary Questions

- How do we determine that a hybrid subject device could increase or decrease the efficiency of the information retrieval system?
- How do we analyse a large volume of the subject device in terms of GB (Gigabyte) or TB (Terabyte)?

4.2 Secondary Questions

- How do we perform to perform an exploratory data analysis with the hybrid subject device?
- How do we measure the similarity and dissimilarity between taxonomy terms and folksonomy terms using ML techniques?
- How do we make a prediction analysis of taxonomy and folksonomy terms using ML models?

5. HYPOTHESIS

The hypothesis is stated to determine the efficiency of the information retrieval system of a hybrid subject device that contains a large amount of taxonomy, as well as folksonomy terms by using the machine learning technique. For the purpose of this study, the machine learning technique includes performing exploratory data analysis, similarity and dissimilarity between subject terms and folksonomy terms and prediction analysis of taxonomy and folksonomy terms.

6. OBJECTIVES

The general objective of the research work, as a whole, is to determine the efficiency of the information retrieval system of a hybrid subject device that contains a large amount of taxonomy, as well as folksonomy terms by using machine learning techniques.

The above objective can be derived from some specific objectives:

- To perform an exploratory data analysis of the hybrid subject device (taxonomy terms and folksonomy terms).
- To measure the similarity and dissimilarity between subject terms and folksonomy terms by using machine learning techniques.
- To make a prediction analysis of taxonomy and folksonomy terms using ML models.

7. SIGNIFICANCE OF THE STUDY

In this research work, machine learning-based methodology has been introduced for analysing and examining the dataset. Data science deals with a large volume of data, and machine learning deals with the analytics part of the datasets. The analytics part includes statistical and mathematical algorithms to apply in the dataset. In the present day, we can observe that every domain concentrates on their data, and they apply machine learning-based tools and techniques to analyse their data and get more accuracy on results. We have explored the EDA by using the Python machine learning tool and used the logistic regression model for prediction analysis. We have discussed results and findings in Section 10.

8. METHODOLOGY

The methodology includes three parts. The first part includes an exploratory data analysis between taxonomy terms and folksonomy terms by using advanced data analytical tools – machine learning tools. The second part involves the prediction analysis of taxonomy and the folksonomy dataset. The final part includes measuring the similarity and dissimilarity between taxonomy terms and folksonomy terms by using machine learning techniques. In this research work, Python (https://www.python.org) is used as a machine learning tool for exploratory data analysis, prediction analysis, and similarity measurement. Figure 1 shows a proposed framework for methodology exploratory data analysis, prediction analysis, and similarity measurement with its various components for this research work.

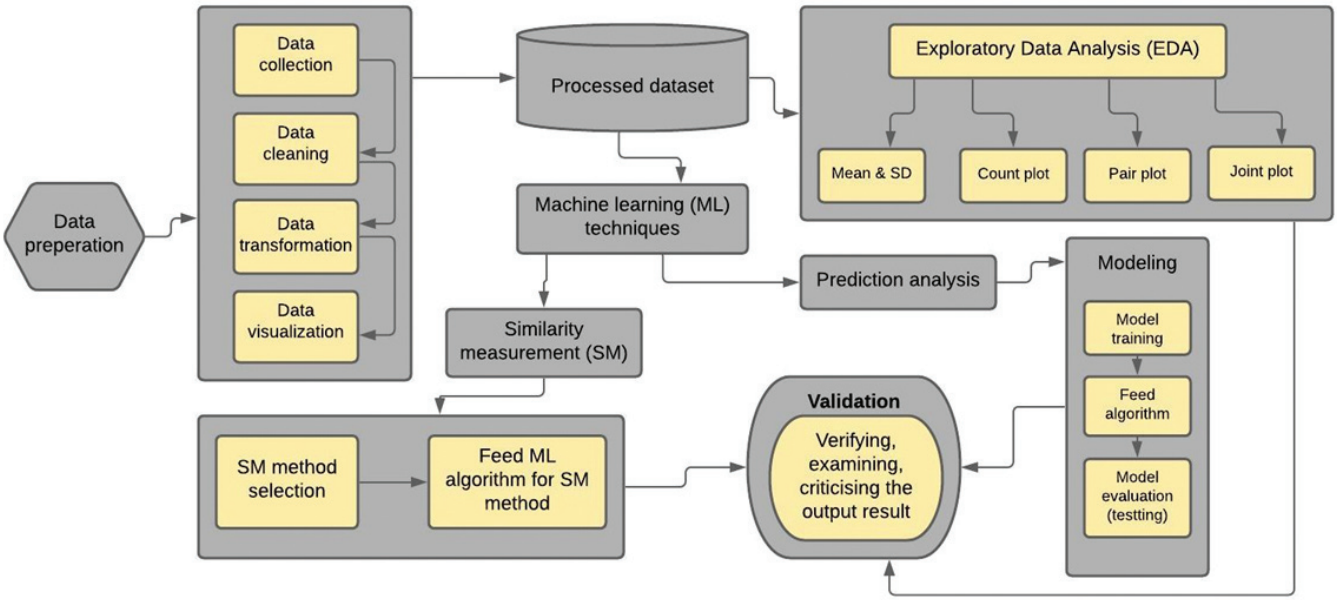

Figure 1. The proposed framework for a methodology for EDA, prediction analysis, and similarity measurement of the hybrid subject device (folksonomy and taxonomy).

The authors have answered all the research questions in this methodology section. The first question is how to perform EDA, and the answer had revealed in Sections 8.3 and Section 8. The second question is how to measure the similarity and dissimilarity between taxonomy and folksonomy terms, and the answer is received in Section 8.5 and Section 8.5 and Section 10.2. The third question is how to make a prediction analysis of taxonomy and folksonomy terms and the answer is received in Section 10 and section 10.1.

8.1 Experimental Data Collection

This research includes two primary concepts – taxonomy and folksonomy. Here, data contain taxonomy terms and folksonomy terms. The taxonomy terms also include BT, NT, RT, and RS of subject terms that are collected from LCSH by the researcher. The folksonomy terms are collected by conducting a survey. The survey has been done in the Jadavpur University (http:///www.jaduniv.edu.in). The folksonomy terms are collected from students and faculty members by providing a form that mentions twenty unique titles of books from six different departments, i.e., Civil engineering, Architecture, Mathematics, Chemistry, English, and History. In case of collection of subject terms, searched and retrieved the subject keyword (LCSH terms) of those twenty book titles from each department from OPAC of Library of Congress (loc.gov). The total number of terms, both subject terms and folksonomy terms, is 2642 that including B.T., NT, R.T., and R.S. in subject terms. The collected data have been transformed as cleaned and filtered and divided into two datasets. The cleaning task has included with correction of spell mistakes, and filtering has decided to take attributes for datasets. The first dataset is named with 'dataset_terms_1.csv' for EDA and prediction analysis, and the second is 'dataset_terms_2.csv' for similarity measurement between folksonomy and taxonomy. The file 'dataset_terms_1.csv' has contained the frequencies of subject terms along with BT, NT, RT, and RS Another file, 'dataset_terms_2.csv' has contained 2642 terms with its frequencies in two vectors -LCSH terms and folksonomy terms. Figure 2 shows the data collection process of the research work graphically.

8.2 Selection of Machine Learning Tool

In terms of 'open', the authors have selected open-source software, i.e., Python programming language from Google Colaboratory (in short 'colab'). The Google Colaboratory (https://colab.research.google.com) is open to access for every Google account holder. It provides a web-based IDE (Integrated Development Environment) to write Python scripts without installation and configuration in our computer/laptop. It also provides GPU (Graphics Processing Unit) in place of CPU (Central Processing Unit) and a minimum of 12 GB RAM to write Python scripts for machine learning models.

8.3 Exploratory Data Analysis

The exploratory data analysis (EDA) is a task performed by the data analyst before starting the machine learning models /techniques to get familiar with the dataset. The EDA includes the practice of analysing quantitative data and visualisation of the dataset without making any assumptions about its content[18]. It is a very important step before going to machine learning or statistical modelling because it helps to build an appropriate dataset to develop an appropriate model/technique for prediction as well as interpret the outcomes, findings, and results. In this research work, EDA has performed with the dataset ('dataset1_terms.csv') and found a lot of results. The EDA has been discussed in detail in the next section.

8.4 Prediction Analysis Through ML Model

Machine learning includes mathematical techniques, statistical techniques, and prediction models. The prediction analysis performs the trends and patterns in data[19]. To analyse prediction, authors have used ML prediction models. The most common and widely used prediction models are: i) Decision tree, ii) Regression (linear and logistic), iii) Neural networks, iv) Random forest, v) SVM (support vector machine), vi) Clustering and more[20]. In this research work, logistic regression has been used both to perform the trends and patterns in data. The details are discussed in Section 7.

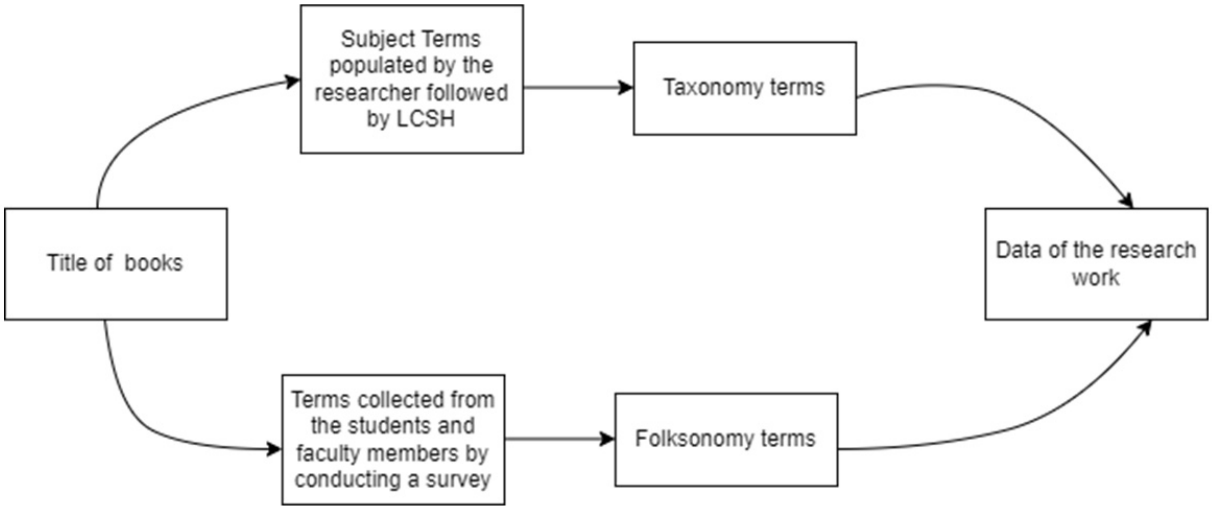

Figure 2. Data collection process of the research work.

8.5 Similarity Measurement

The similarity measure refers to a wide variety of meanings among mathematics, probability, data science, and machine learning techniques. In machine learning, generally, the similarity is applied in unsupervised models, which provides unlabelled data that the algorithms try to make sense of by extracting features and patterns of their own[21]. The training dataset is a collection of examples without a specific desired outcome or correct answer. There are many popular similarity measurements for machine learning techniques like – Cosine similarity, Manhattan distance, Euclidean distance, Minkowski distance, Jaccard similarity, etc.[22]. The authors have applied cosine similarity measurement because there are two vectors–LCSH terms and folksonomy terms. In the next sections, the authors have discussed the data analysis and interpretation for EDA, prediction analysis, and similarity measurement of the two data files.

9. DATA ANALYSIS AND INTERPRETATION

As before said that the dataset has been divided into two datasets the first dataset ('dataset_terms_1.csv' (Table 1)) contains the frequencies of subject terms along with B.T., NT, R.T., and R.S.; and the second dataset 'dataset_terms_2.csv' (Table 2)) contains 2642 terms with its frequencies in two vectors -LCSH terms and folksonomy terms. The EDA has been made on the first dataset – dataset_terms_1.csv file.

Table 1. dataset_terms_1.csv

	Acc_no	Departments	LCSH_terms	Folksonomy_terms	RT	NT	RS	BT
0	D1	Civil Engineering	37	93	7	26	0	2
1	D2	Civil Engineering	15	107	0	12	0	2
2	D3	Civil Engineering	23	71	2	18	0	2
3	D4	Civil Engineering	16	64	1	13	0	2
4	D5	Civil Engineering	21	88	3	5	0	3
...	...	...	...	...	...	...	...	...
115	D276	Mathematics	30	94	2	26	0	1
116	D277	Mathematics	21	48	15	2	0	3
117	D278	Mathematics	46	54	3	50	0	1
118	D279	Mathematics	30	75	44	2	0	4
119	D280	Mathematics	42	51	25	7	0	3
120 rows × 8 columns								

Table 2. dataset_terms_2.csv

ID	Departments	Terms(LCSH+Folksonomy)	LCSH_terms	Folksonomies_terms
0	Civil Engg	Soil mechanics	13.0	128.0
1	Civil Engg	Geotechnical engineering	11.0	93.0
2	Civil Engg	Mechanics	11.0	1.0
3	Civil Engg	Anchorage (Structural engineering)	9.0	0.0
4	Chemistry	Chemistry, Physical and theoretical	9.0	0.0
...	...	...	...	...
2636	Civil Engg	Weldments-Residural Stress	0.0	1.0
2637	Civil Engg	Western World History	0.0	1.0
2638	History	WFF 'N PROOF (Game)	0.0	1.0
2639	History	World Peace Of German	0.0	1.0
2640	English	Zoning	0.0	1.0
2641 rows × 4 columns				

Table 3. Display the statistical values of the dataset

df.describe()						
	LCSH_terms	**Folksonomy_terms**	**RT**	**NT**	**RS**	**BT**
count	120.000000	120.000000	120.000000	120.000000	120.000000	120.000000
mean	17.050000	67.625000	1.358333	13.958333	0.050000	2.141667
std	21.835077	18.173331	4.905486	23.178662	0.313559	1.764714
min	1.000000	11.000000	0.000000	0.000000	0.000000	0.000000
25%	2.000000	56.000000	0.000000	0.000000	0.000000	1.000000
50%	7.000000	64.500000	0.000000	2.000000	0.000000	2.000000
75%	28.500000	79.000000	1.000000	23.000000	0.000000	3.000000
max	143.000000	125.000000	44.000000	139.000000	2.000000	10.000000

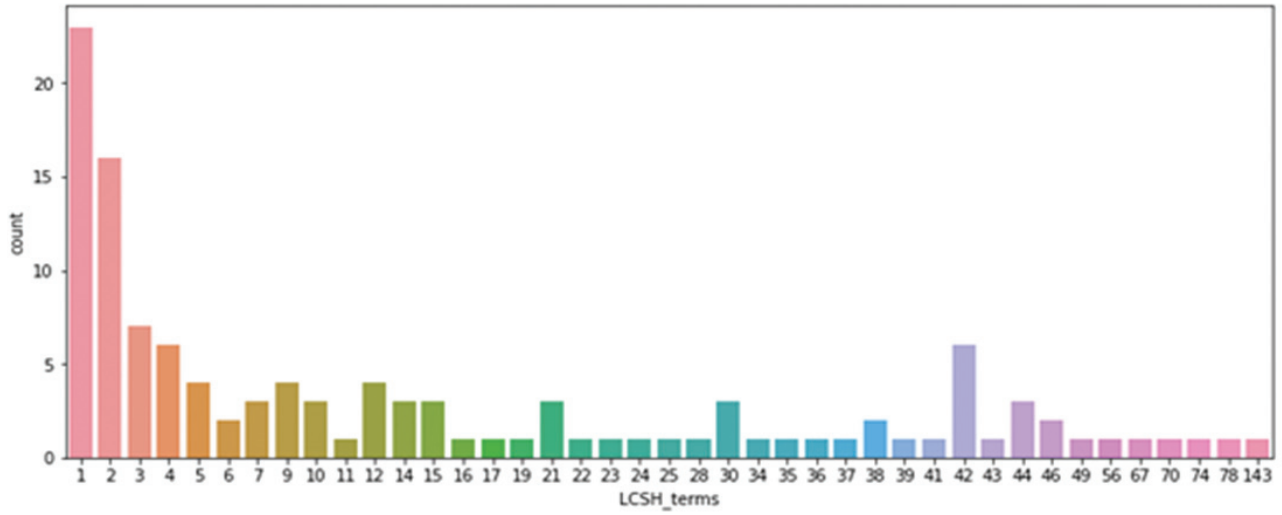

Figure 3. Display the bar diagram of frequencies of LCSH terms.

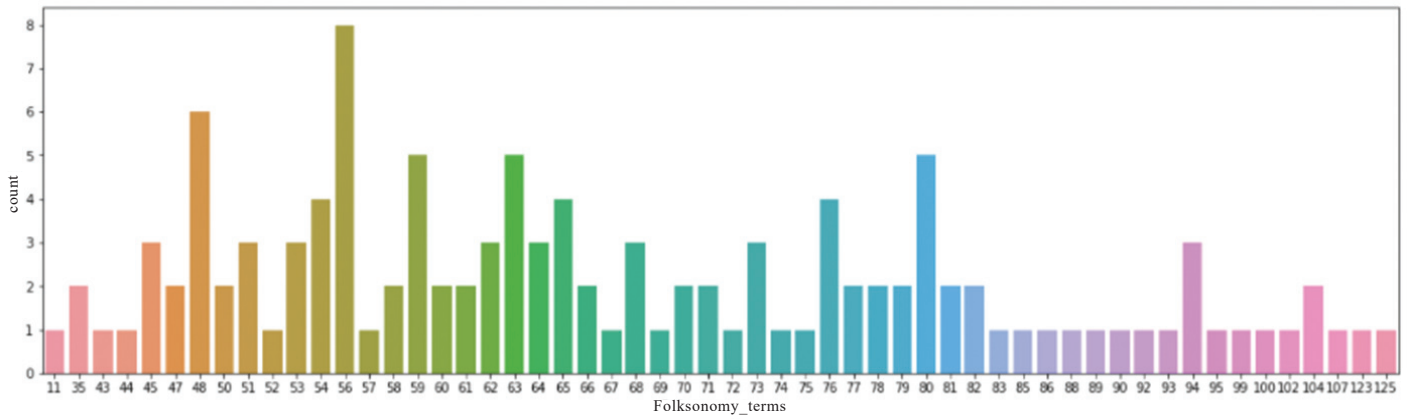

Figure 4. Display the bar diagram of frequencies of folksonomy terms.

The Google Colab Research Python notebook is used as a scripting language to analyse the dataset file.

Before starting the EDA, machine learning models, and similarity measurement, imported four python libraries - NumPy, pandas, seaborn, and matplotlib[23]. The libraries are imported as different variable names like np, pd, sns, and plt that expressed as follows:

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

In the next step, the dataset file – 'dataset_terms_1.csv' has been uploaded and stored in the 'df' variable as follows:

```
df = pd.read_csv('dataset_terms_1.csv')
```

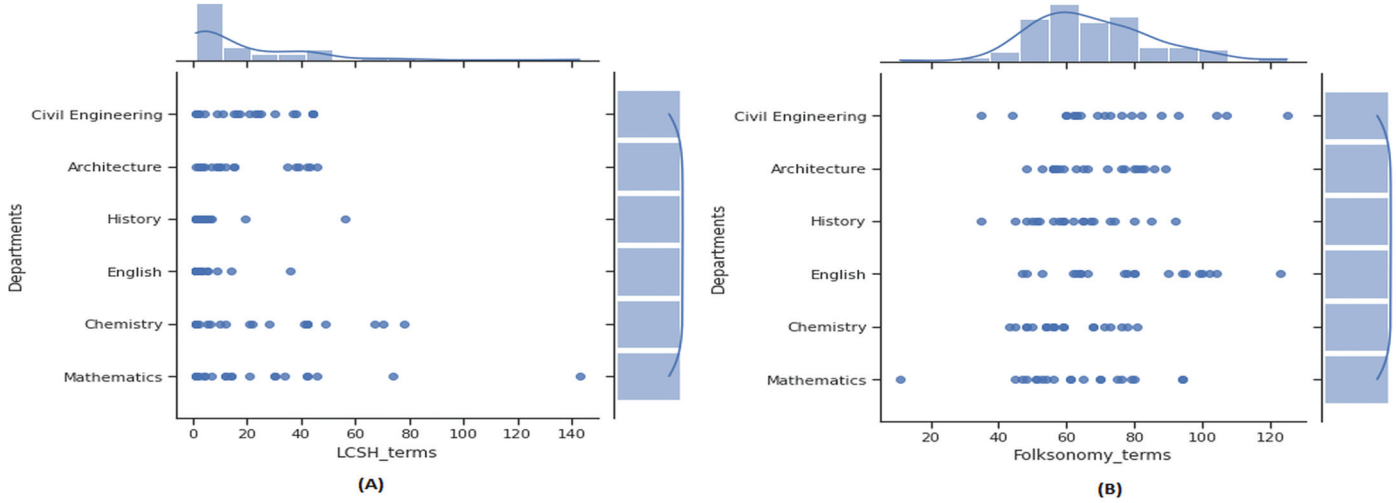

Figure 5. Display the join plot between attributes 'Departments' & 'LCSH_terms' (A) and 'Departments' & 'Folksonomy_terms' (B).

Statistical information is an important step to study a dataset. The statistical information of the dataset has been retrieved using the "Pandas describe()" library to retrieve the count, mean, standard deviation, minimum and maximum values, and percentile. The highest mean (67.625000) has been found of the attribute 'Folksonomy_terms', and the highest standard deviation, 23.178662, has been found of the attribute 'NT' (Table 3).

In the next step, a bar diagram has been shown by using the "Seaborn Countplot()" library of the attribute 'LCSH_terms' column of the dataset. Figure 3 depicts the bars of each frequency of the LCSH terms where the minimum and maximum frequencies are 1 and 143, respectively. The command to display the bar diagram (Fig. 3) as follows:

```
sns.countplot(x='LCSH_terms',data=df)
set_size(10,4)
plt.show()
```

Another bar diagram in Fig. 4 has been populated with respect to folksonomy terms and using the same library, i.e. 'Seaborn Countplot()'. The minimum and maximum frequencies of the folksonomy terms are 11 and 125, respectively. The command to display the bar diagram (Fig. 4) as follows:

```
sns.countplot(x='Folksonomy_terms',data=df)
set_size(15,4)
plt.show()
```

In the next step, a pair plot (Annexure I) has been displayed by using the 'Seaborn pairplot()' library. The pair plot function creates a grid of axes such that for each variable, the data will be shared in the y-axis across a single row and in the x-axis across a single column. The command of pairplot (Annexure I) is expressed as follows:

```
sns.set(style="ticks", color_codes=True)
sns.pairplot(df, hue='LCSH_terms')
```

In the following section (Fig. 5), the authors have drawn two joint plots using the 'Seaborn joinplot()' library. The joint plot refers to a relationship between two bivariate variables with several canned plot kinds. Here, the authors put the attribute 'Departments' column on the y-axis and the attribute 'LCSH_terms' on the x-axis (Fig. 5A), and another bivariate variable 'Folksonomy_terms' put in X-axis (Fig. 5B). The following commands are used for these plots:

```
sns.jointplot(x='LCSH_terms',y='Departments',data=df,kind='reg')
sns.jointplot(x='Folksonomy_terms',y='Departments',data=df,kind='reg')
```

In the Fig. 6, authors have drawn four major join plots between the attributes 'LCSH_terms' and 'NT' (Fig. 6A); 'LCSH_terms' and 'BT' (Fig. 6B); 'LCSH_terms' and 'RT' (Fig. 6C) and 'LCSH_terms' and 'RT' (Fig. 6D).

In the next step, the authors have computed the pairwise correlation of the columns. Table 4 is displaying the values of correlation of the attributes – 'LCSH_terms', 'Folksonomy_terms', 'RT', 'NT', 'RS' and 'BT'.

After computing the values of correlation of columns, authors have put the values in a matrix called 'heatmap' using the 'Seaborn heatmap()' library. The heatmaps are typically used to visualise correlation matrices. Figure 7 has the heatmap correlation matrices of the dataset. The highest value of correlation is 0.891038 between the attributes 'LCSH_terms' and 'NT'.

The above analyses are made as EDA of the dataset 'dataset_terms_1.csv'. The authors have found a lot of analyses and plots the dataset through countplot(), pairplot, jointplot(), heatmap() functions. After EDA, authors have moved to prediction analysis by machine learning techniques. For this dataset, the authors have used a logistic regression model the dataset to find out a clear idea.

10. PREDICTION ANALYSIS USING LOGISTIC REGRESSION

In this research work, the authors have used the "Logistic regression" (LR) model to predict the trend of the dataset[24].

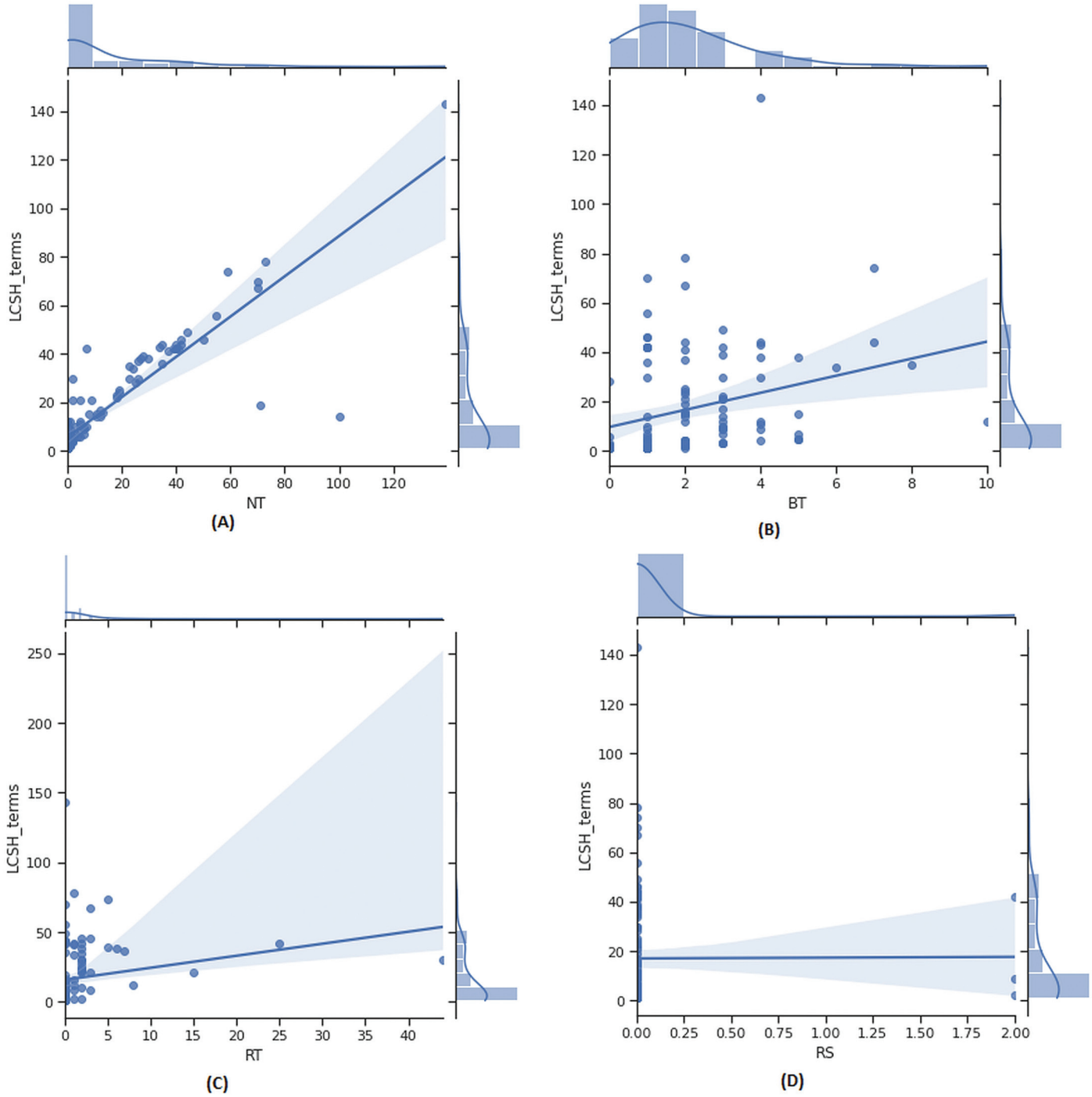

Figure 6. Display the join plots between attributes 'LCSH_terms' & 'NT' (A); 'LCSH_terms' & 'BT' (B); 'LCSH_terms' & 'RT' (C) and 'LCSH_terms' & 'RS'.

The LR is the fundamental classification technique. The classification is an area of supervised machine learning that tries to predict which class or category some entity belongs to, based on its features[25]. The LR is a supervised learning algorithm and is used to predict the probability of dependable variables. The supervised data comprises labelled data. There are six actual classes in 'Departments' attribute– (i) Architecture, (ii) Chemistry, (iii) Civil Engineering, (iv) English, (v) History, (vi) Mathematics. The attribute 'LCSH_terms' is a dependable variable of NT, BT, RT, and R.S. attributes. The undependable variables are 'Folksonomy_terms', NT, BT, RT, RS. The number of inputs of 'Folksonomy_terms' depends on the user's choices, and NT, BT, RT, RS are dependable on the subject device. Between these features, there is no certain value increasing/decreasing pattern. So, logistic regression is the best-fitted model for this prediction analysis.

The logistic regression model is expressed as:

$$P = \frac{e^{y}}{1+e^{y}} \quad (1)$$

where 'P' is the probability of an event occurring (e.g., the

Table 4. Display the correlation values of the all columns from the dataset

df.corr()						
	LCSH_terms	**Folksonomy_terms**	**RT**	**NT**	**RS**	**BT**
LCSH_terms	1.000000	-0.076359	0.193692	0.882850	0.004541	0.279617
Folksonomy_terms	-0.076359	1.000000	-0.014505	-0.015299	0.015116	0.092331
RT	0.193692	-0.014505	1.000000	-0.002528	-0.011746	0.175612
NT	0.882850	-0.015299	-0.002528	1.000000	0.004914	0.139436
RS	0.004541	0.015116	-0.011746	0.004914	1.000000	-0.073655
BT	0.279617	0.092331	0.175612	0.139436	-0.073655	1.000000

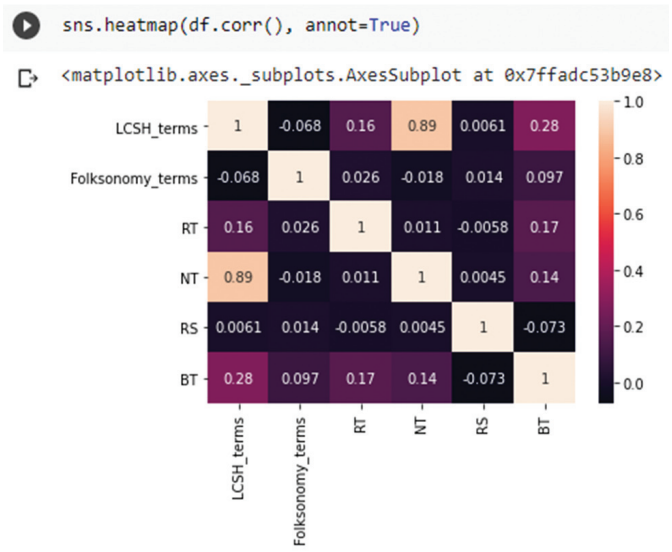

Figure 7. Display the heatmap matrices of correlation values of the dataset.

probability of the presence of a species), 'e' is the basis of the natural logarithm, and 'y' is a regression equation of the form

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots\ldots + \beta_n x_n \quad (2)$$

where α is a constant and β1, β2, ..., βn are the coefficients of the n predictor variables x_1, x_2, ..., x_n. This linear regression also be expressed by removing e and adding natural algorithm ln[26]

$$y = \ln\left(\frac{p}{1-p}\right) \quad (3)$$

Before the implementation of logistic regression in machine learning, the authors have split the dataset into two sets – training and test dataset and assumed the size of the test is 31 per cent. The authors have declared the training and test dataset through 'sklearn' model selection as follows:

```
from sklearn.model_selection import train_test_split
X_train,X_test,y_train,y_test = train_test_split(X,y,test_
size=0.31,random_state=101)
```

The authors assigned 'LogisticRegression()' function in a variable 'LogisR' as follows:

```
from sklearn.linear_model import LogisticRegression
logisR=LogisticRegression()
```

The authors have computed intercept values of the six classes that cross the y-axis. The intercept function with values is expressed as follows:

```
print(logisR.intercept_)
[ 0.02450552 0.36850955 -0.46037401 -0.33671379 0.33624438
0.06782836]
```

The highest intercept value is 0.36850955 that belongs to the 'Chemistry' class, and the least value '-0.46037401' of intercept belongs to the 'Civil Engineering' class.

The regression coefficient is a measurement that uses to measure the average functional relationship among variables. There are different values coefficient of classes in L.R. is found as shown in Table 5.

Table 5. Display the different values of coefficient of classes in LR

logisR.coef_
array([[0.51696756, -0.00172758, 0.159257, -0.51831441, 0.3246581, -0.59073472],
[0.20954505, 0.00390634, -0.12777273, -0.17582313, -0.19103773, -0.60882288],
[0.43234569, 0.00266878, 0.45591397, -0.43634703, -0.14441032, -0.42157989],
[-0.65149861, 0.02111755, -0.55564148, 0.61250979, -0.18297263, 0.72487102],
[-0.73855475, -0.0043732 , -0.6459207 , 0.71679149, -0.09638117, 1.04800362],
[0.23119506, -0.02159189, 0.71416394, -0.19881671, 0.29014374, -0.15173715]])

10.1 Classification Report

The Classification report is an important part of a classification algorithm, and it is used to measure the accuracy of predictions from a classification algorithm[27]. It provides a report that indicates how many predictions are True and how many are False. The classification report consistive of True Positives (TP), False Positives (FP), True negatives (TN), and False Negatives (FN) are used to predict the metrics of

a classification report[28]. The classification report for the L. is expressed as follows:

from sklearn.metrics import classification_report

The output of the classification report is mentioned in Table 6.

Table 6. Display the classification report of the different classes

print(classification_report(y_test,prediction))				
	precision	**recall**	**f1-score**	**support**
Architecture	0.40	0.40	0.40	5
Chemistry	0.33	0.75	0.46	4
Civil Engineering	0.20	0.17	0.18	6
English	0.43	0.86	0.57	7
History	0.33	0.11	0.17	9
Mathematics	0.50	0.14	0.22	7
accuracy			0.37	38
macro avg	0.37	0.40	0.33	38
weighted avg	0.37	0.37	0.32	38

Table 6 shows the different classification metrics - precision, recall, and f1-score on a class basis. The metrics are based on true and false positives, true and false negatives. The combination of positive and negative is used for predicted classes. There are four combinations of positive and negative if the predictions are right or wrong[29]:

- TN (True negative): when a case was negative and predicted negative
- TP (True positive): when a case was positive and predicted positive
- FN (False negative): when a case was positive but predicted negative
- FP (False positive): when a case was negative but predicted positive

The precision indicates the correct prediction of each class, and it refers to the accuracy of positive predictions. Precision is defined as the ratio of true positives to the sum of true and false positives. It is expressed below:

$$\mathrm{Pr}\,ecision = \frac{TP}{(TP + FP)} \tag{4}$$

The recall indicates the percentage of cases that have been retrieved. It is a fraction of positives that were correctly identified. The recall is defined as the ratio of true positives to the sum of true positives and false negatives.

$$\mathrm{Re}\,call = \frac{TP}{(TP + FN)} \tag{5}$$

The f1 score indicates the percentage of the correctness of positive predictions. It is a weighted harmonic mean of precision and recall. The best score f1 is 1.0, and the worst is 0.0. The f1 score should be used to compare classifier models, not global accuracy. The f1 score is defined as follows:

$$f1 - score = 2 \times \frac{(\mathrm{Re}\,call \times \mathrm{Pr}\,ecision)}{(\mathrm{Re}\,call + \mathrm{Pr}\,ecision)} \tag{6}$$

10.2 Similarity Measurement between LCSH and Folksonomy Terms

In this research work, the cosine similarity technique is used to measure the similarity between two types of vectors- 'LCSH terms' (A) and 'Folksonomy terms'(B) and this is the best fit for the cosine similarity technique. The formula of Cosine-Similarity is:

$$Similarity = \cos\theta = \frac{\mathrm{A} \times \mathrm{B}}{\|\mathrm{A}\|\|\mathrm{B}\|} = \frac{\sum_{i=1}^{n} \mathrm{A}_i \times \mathrm{B}_i}{\sqrt{\sum_{i=1}^{n} (\mathrm{A}_i)^2} \times \sqrt{\sum_{i=1}^{n} (\mathrm{B}_i)^2}} \tag{7}$$

The entire code is written in Python script to find out the cosine similarity distance and angle(degree) between two vectors (LCSH terms (A) and Folksonomy terms (B)). The machine learning model of cosine similarity distance is shown as follows:

```
from sklearn.metrics.pairwise import cosine_similarity,cosine_
distances
cos_sim=cosine_similarity(A.reshape(1,-1),B.reshape(1,-1))
print (f"Cosine Similarity between A and B:{cos_sim}")
print(math.degrees(0.30151134*(math.pi/2)))

Cosine Similarity between A and B:[[0.30151134]]
27.136020600000002
```

As a result, the authors found a 30 per cent (Score: 0.30151134) similarity between LCSH terms (A) and Folksonomy terms (B) and 27 degrees (approx.) between the two vectors (Fig. 8).

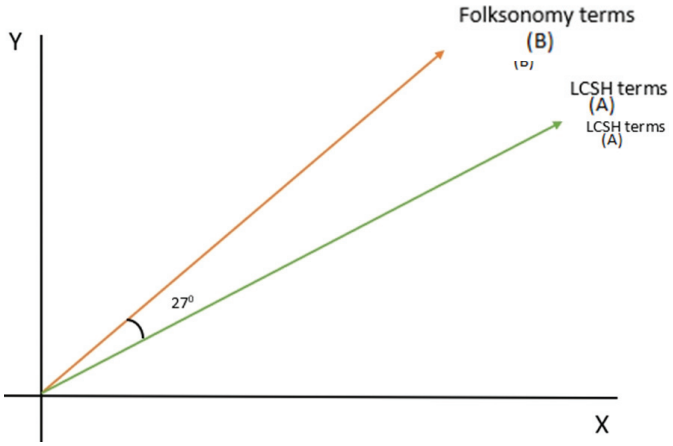

Figure 8. Cosine angle between LCSH terms and Folksonomy terms.

11. RESULTS AND DISCUSSION

As said before, the research work has mainly three parts - exploratory data analysis, prediction analysis, and similarity measurement of folksonomy and taxonomy dataset. In the EDA

part, there are a lot of works have been shown. The dataset includes various attributes like- 'Acc_no',' Departments', 'LCSH_terms', 'Folksonomy0_terms', 'RT', 'NT', 'RS', 'BT'. The unique values have found 120, 6, 41, 56, 11, 38, 2, and 10 of 'Acc_no', 'Departments', 'LCSH_terms', 'Folksonomy_ terms' 'RT', 'NT', 'RS', 'BT' respectively. We have found that folksonomy terms (56) are higher than taxonomy terms (41) for 120 titles of books. The narrower terms (38) have taken a leadership position among other subject terms – RT, RS, BT. Table 3 shows the statistical information. The highest value of mean is '67.625000' of the attribute 'Folksonomy_terms', and the highest standard deviation (S.D.) is 23.178662 of the 'NT' attributes. Figure 3 depicts the bar diagram of the frequency of LCSH terms where the minimum and maximum frequencies are 1 and 143 respectively. The highest value of count is '24' of the frequency 1 and lowest value is 1 of many frequencies like 11, 16, 17, 19, 22, 25, 28, 34, 35, 36, 37, 39, 41, 43, 49, 56, 67, 70, 74, 78, 143. The taxonomy terms have been populated by the researcher followed by the LCSH of 120 titles of books from different subjects and naturally, we have found heterogeneous types of frequencies and appearances of terms.

Similarly, Fig. 4 shows the minimum and the maximum frequencies of the folksonomy terms 1 and 125 respectively. The frequency 56 is the highest count value 8 and lowest count value is 1 of many frequencies - 11, 43, 44, 52, 57, 67, 69, 72, 74, 75, 83, 85, 86, 88, 89, 90, 92, 93, 95, 99, 100, 102, 107, 123, 125. Being folksonomy terms are collaborative contributions and it contains the approaches of the people. We can observe that the terms that have 56 frequencies are mostly 8 times appeared among folksonomy terms.

Figure 5 depicts a pair plot that grid of axis such that each variable the data will be shared in the y-axis across a single class. The pair plot has been generated among the six attributes of the dataset and it has produced 36 graphs. Six well-produced sub-graphs have appeared diagonally (left top to right bottom). These sub-graphs are the same attribute in both the x-axis and y-axis. One sub-graph (4th row and 1st column) between NT (y-axis) and LCSH_terms (x-axis) has produced a positive slope that when LCSH_terms are increased, the NT terms also be increased.

Figure 6 shows two joint plots for the relationship between 'Departments' and 'LCSH_terms' attributes and 'Departments' and 'Folksonomy_terms' attributes. Figure 6A depicts the range of frequencies is 0-60 which is impacted on six departments. Figure 6B shows that the relationship between 'Departments' and 'Folksonomy_terms' and the impacted range is 40- 100 which most frequencies on the six departments. The 'NT', 'BT', 'RT', 'RS' is strongly related to LCSH terms. The LCSH terms have mostly appeared in the frequency range 0-70 from all six departments (Fig. 6A). The folksonomy terms have mostly appeared in the frequency range 4-80 from all six departments (Fig. 6B).

Figure 7 depicts the four joint parts between the attributes 'LCSH_terms' and 'NT'; 'LCSH_terms' and 'BT'; 'LCSH_ terms' and 'RT' and 'LCSH_terms' and 'RS'. Figure - 7A shows a positive slope between the attributes 'LCSH_terms' and 'NT' and the range of 0-90 is more impacted on the slope. Figure 7B and Fig. 7C shows also a positive slope between the attributes 'LCSH_terms' and 'BT'; 'LCSH_terms' and 'RT' respectively. But the Fig. 7D shows a parallel slope between the attributes 'LCSH_terms' and 'RS'. There is no impact of the attribute 'RS' and 'LCSH_terms'. A positive slope has been seen in Fig. 7A between LCSH_terms and NT with the frequency range 0-60. The RT of LCSH_terms mostly appears in the frequency range 0-5 (Fig. 7C). Table 4 depicts correlation values between all attributes in the dataset. The highest value of correlation is 0.891038 between the attributes 'LCSH_terms' and 'NT'.

The second part of this research work is about predicting the trend of the dataset. There are six actual classes in 'Departments' attribute and they are - i) Architecture ii) Chemistry iii) Civil Engineering iv) English v) History vi) Mathematics. The logistic regression has been applied to predict the actual classes data. The authors have split the dataset into two sets - training and test dataset and assumed the size of the test is 31 per cent. The classification report consists of TP, FP, TN, and FN and it produces classification metrics - precision (4), recall (5), and f1-score (6). The precision values of the classes 'Architecture', 'Chemistry', 'Civil Engineering', 'English', 'History' and 'Mathematics', are 0.40, 0.33, 0.20, 0.43, 0.33 and 0.50 respectively. These are very low precision values. The recall and f1-score values of the classes - 'Architecture', ' Chemistry',' Civil Engineering',' English', 'history', 'Mathematics' are 0.40 & 0.40; 0.75 & 0.46; 0.17 & 0.18, 0.86 & 0.57; 0.11 & 0.17; and 0.14 & 0.22 respectively. The accuracy value of f1- score is 0.37 at the training percentage as 69. Being the value of f1- score is below 0.7, a negative prediction has been found of the dataset.

In the final part, a similarity measurement has been computed by using the cosine similarity technique (7). There are two vectors- A (for LCSH terms) and B (Folksonomy terms). We found the similarity score and degrees of the angle between two vectors. The similarity score is found as 30 per cent (0.03151134) and the degree of angle between two vectors is 27 degrees (approx). Being very less percentage of similarity, it would be recommended for hybrid subject device and efficiency of information retrieval must be increased.

12. CONCLUSION

Data science, machine learning, artificial intelligence in the modern age are rapidly being applied in various domains, like-business domain, science and technology domain, educational domain, sports domain, medical science domain, entertainment domain, and also a library and information science domain. This research work has mainly three parts, one is exploratory data analysis, prediction analysis, and similarity measurement. The Library and Information Science (LIS) field handles big volume data, where it may be bibliographic data, authority data, circulation data, or maybe linked data or any other data that are related to the LIS field.

Folksonomy data is a group of collaborative data by the people and expected it is built with a large volume of data and it may be controlled easily by applications of machine learning techniques. The machine learning technique is used to analyse large amounts of data in terms of GB (GigaByte), TB (TeraByte) in smart ways. The library contains such kinds of data like 20 or 30 years of circulation data, subject terms

of two lakhs collection, large amount users' tags of electronic documents, etc. These kinds of data are called big data and could not perform analysis for exploratory data analysis, classification task, recommending system, etc in simple spreadsheet software of alike.

This research work is a primary attempt to apply machine learning techniques in folksonomy and taxonomy datasets which are managed by LIS professionals. The authors have shown a lot of data analysis through EDA in Python script easily in the above sections. The EDA includes frequency of LCSH terms, folksonomy terms pair plot and joint plot of LCSH and folksonomy terms, heat map. The prediction analysis in machine learning techniques used a logistic regression model to predict the trend of the data set. In prediction analysis, it found a very poor accuracy score – 0.37, which means there is no predictable pace between 'Folksonomy' and 'LCSH' terms. Finally, the authors have found the similarity distance and angle between two vectors (LCSH terms and Folksonomy terms). It is found that the cosine similarity distance between LCSH terms and Folksonomy terms is 0.30151134 and the angles between LCSH terms and Folksonomy terms is 27 degrees (approx). It means that we need to keep the folksonomy terms as approach terms in the system for efficient information retrieval service. Presently, LIS professionals handle big volumes of data and it is very difficult to analyse or another advanced task that may not available in the system. Suppose, a library has built an Institutional Digital Repository which does not visualize the tag cloud in the interface, and tags become a big volume of terms. How do they develop a tag cloud visualisation of those tags? This kind of problem could be solved through machine learning techniques. This research work is a trial to use machine learning techniques for library data. How library professionals could use these techniques on their specific problems for better library services. So, it may be concluded that the LIS professional could apply data science and Machine Learning techniques to compute the accurate result and values, EDA, prediction analysis from their large amount of data.

REFERENCES

1. Buchanan, S. Planning strategically, designing architecturally: A framework for digital library services. Advances in Librarianship (Advances in Librarianship, Vol. 32), edited by A. Woodsworth. Emerald Group Publishing Limited, Bingley, 2010, 159-180. doi: 10.1108/s0065-2830(2010)0000032010.
2. Ranganathan, S.R. & Gopinath, M.A. Prolegomena to library classification. Asia Publishing House, Bombay, 1967, 346-347.
3. Chatterjee, Swarnali. Design and development of a folksonomy driven subject access system in digital environment. University of Kalyani. 2017. PhD Thesis. 209p. URI: http://shodhganga.inflibnet.ac.in:8080/jspui/handle/10603/248534. (Accessed on 4 May 2021).
4. Bates, J. & Rowley, J. Social reproduction and exclusion in subject indexing. *J. Doc.*, 2011, **67**(3), 431–448. doi: 10.1108/00220411111124532.
5. Markey Drabenstott, K. & Vizine-Goetz, D. The future of subject headings for online information retrieval. Using subject headings for online retrieval: Theory, practice and potential (Library and Information Science, Vol. 94B), edited by K. Markey Drabenstott & D. Vizine-Goetz. Emerald Group Publishing Limited, Bingley, 1994, 330-342. doi: 10.1108/S1876-0562(1994)000094B016.
6. Batch, Y.; Yusof, M.M. & Noah, S.A. ICDTag: A prototype for a web-based system for organizing physician-written blog posts using a hybrid taxonomy-folksonomy approach. *J. Med. Internet Res.*, 2013, **15**(2), e41. doi: 10.2196/jmir.2353
7. Mukherjee, S. & Das, R. Integration of domain-specific metadata schema for cultural heritage resources to DSpace: A prototype design. *J. Libr. Metadata*, 2020, **20**(2–3), 155–178. doi: 10.1080/19386389.2020.1834093
8. Roman, D.; Reeves, N.; Gonzalez, E.; Celino, I.; Abd El Kader, S.; Turk, P.; Soylu, A.; Corcho, O.; Cedazo, R.; Re Calegari, G.; Scandolari, D. & Simperl, E. An analysis of pollution citizen science projects from the perspective of data science and open science. *Data Technol. Appl.*, 2021, **55**(5), 622-642. doi: 10.1108/dta-10-2020-0253
9. Lantz, B. Overview of machine learning tools. The machine age of customer insight, Edited by M. Einhorn, M. Loffler, E de Bellis, A. Herrmann, & P. Burghartz. Emerald Publishing Limited, Bingley, 2021, 79-90. doi: 10.1108/9781839096945
10. Bates, J. & Rowley, J. Social reproduction and exclusion in subject indexing: A comparison of public library OPACs and LibraryThing folksonomy, *J. Doc.*, 2011, **67**(3), 431-448. doi: 10.1108/00220411111124532
11. Virkus, S. & Garoufallou, E. Data science and its relationship to library and information science: A content analysis. *Data Technol. Appl.*, 2020, **54**(5), 643–663. doi: 10.1108/dta-07-2020-0167
12. Daimari, D.; Narzary, M.; Mazumdar, N.; & Nag, A. A machine learning based book availability prediction model for library management system. *Libr. Philos. Pract.*, 2021, 4982. https://digitalcommons.unl.edu/libphilprac/4982 (Accessed on 20 August 2021)
13. Vaidya, P. & Harinarayana, N.S. The role of social tags in web resource discovery: An evaluation of user-generated keywords. *Ann. Libr. Inf. Stud.*, 2016, **63**, 289–297. http://nopr.niscair.res.in/handle/123456789/39765 (Accessed on 25 July 2021)
14. Choi, Y. & Choi, J.W. A study of job involvement prediction using machine learning technique. *Int. J. Organ. Anal.*, 2020, **29**(3), 788–800. doi: 10.1108/ijoa-05-2020-2222
15. Malhotra, D. K.; Malhotra, K. & Malhotra, R. Evaluating consumer loans using machine learning techniques. Applications of Management Science (Applications of Management Science, Vol. 20), edited by K.D. Lawrence, & D.R. Pai. Emerald Publishing Limited, Bingley, 2020, 59-69. doi: 10.1108/S0276-8976202020

16. Jaison, S.S.; R, N. & Mounusha, S. Machine learning techniques to detect breast cancer. *Int. J. Adv. Res. Comp. Sci.*, 2020, **11**(1), 58–62. doi: 10.26483/ijarcs.v11i0.6546
17. Yesuf, S.H. Breast cancer detection using machine learning techniques. *Int. J. Adv. Res. Comp. Sci.*, 2019, **10**(5), 27–33. doi: 10.26483/ijarcs.v10i5.6464
18. Liu, Q. An application of exploratory data analysis in auditing – credit card retention case. Rutgers studies in accounting analytics: Audit analytics in the financial industry (Rutgers Studies in Accounting Analytics), edited by J. Dai, M.A. Vasarhelyi, & A.F. Medinets. Emerald Publishing Limited, Bingley, 2019, 3-15. doi: 10.1108/978-1-78743-085-320191001
19. Eye, A.; Brandtstädter, J. & Rovine, M.J. Models for prediction analysis. *J. Math. Sociol.*, 1993 **18**(1), 65–80. doi: 10.1080/0022250x.1993.9990116
20. Park, E. IRIS: A goal-oriented big data business analytics framework. 2017. https://utd-ir.tdl.org/handle/10735.1/5407 (Accessed on 22 June 2021).
21. Gupta, S. Top 5 distance similarity measures implementation in machine learning, 2019. https://medium.com/@gshriya195/top-5-distance-similarity-measures-implementation-in-machine-learning-1f68b9ecb0a3 (Accessed on 20 May 2021)
22. Volpi, G.F. Similarity and distance metrics for data science and machine learning. 2019. https://medium.com/dataseries/similarity-and-distance-metrics-for-data-science-and-machine-learning-e5121b3956f8 (Accessed on 5 June 2021).
23. Nelli, F. Python data analytics: with Pandas, NumPy, and Matplotlib. Apress, Italy, 2018. doi: 10.1007/978-1-4842-3913-1
24. Forsyth, David. Probability and statistics for computer science. Springer, USA, 2018. doi: 10.1007/978-3-319-64410-3
25. Alpaydin, Ethem. Introduction to machine learning. The MIT Press, Cambridge, 2010. https://www.pdfdrive.com/introduction-to-machine-learning-e166961950.html (Accessed on 15 June 2021).
26. Real, R.; Barbosa, A.M. & Vargas, J.M. Obtaining environmental favourability functions from logistic regression. *Environ. Ecol. Stat.*, 2006, **13**(2), 237–245. doi: 10.1007/s10651-005-0003-3
27. Jaison, S.S.R.; Nayana., S.; Mounusha & Kodabagi, Mallikarjuna. Machine learning techniques to detect breast cancer. *Int. J. Adv. Res. Comp. Sci.*, 2020, **11**(1), 58–62. http://www.ijarcs.info/index.php/Ijarcs/article/view/6546/5279 (Accessed on 19 June 2021)
28. Gedikli, F. Recommender system and the social web. Springer Vieweg, Germany, 2012. doi: 10.1007/978-3-658-01948-8
29. Raschka, Sebastian. Python machine learning. Packet publishing, Birmingham, UK, 2016. https://www.pdfdrive.com/python-machine-learningpdf-e34337331.html (Accessed on 12 July 2021).

CONTRIBUTORS

Dr Swarnali Chatterjee has obtained her Ph.D. from University of Kalyani. She is working as Librarian at Bengal Music College, Kolkata. Her area of interest includes: knowledge classification, knowledge management, information retrieval, taxonomy, folksonomy, digital library management.
Contributions in this current study: collected data from various departments in Jadavpur University, analysed the folksonomy and taxonomy terms in term of quantitative. She has performed EDA (Exploratory Data Analysis) by using Python scripts.

Dr Rajesh Das has obtained his Ph.D. from Jadavpur University, Kolkata. Presently he is working as Assistant Professor at the Department of Library and Information Science, The University of Burdwan, Burdwan. His area of interest includes: data Science, machine learning, open source software, semantic web, linked open data, digital library, web database applications, computer networking, Linux server administration system, etc.
He has designed the framework of the study, selected the machine learning algorithms and performed the similarity measurement between taxonomy and folksonomy terms and concluded the entire study.

Annexure I

Display the pair plot of the dataset