## CHAPTER-9

# SEMI-AUTOMATED CLASS NUMBER PREDICTION OF BIBLIOGRAPHICAL RESOURCES: A FRAMEWORK DEPLOYING ANNIF

**Meghna Biswas**
Global Library, O.P. Jindal Global University
Sonipat Narela Road, Near Jagdishpur Village, Sonipat, Haryana- 131001
*Email: mbiswas.library@gmail.com / meghna.biswas@jgu.edu.in*

## ABSTRACT

This study investigates an AI/ML-based semi-automated indexing system for libraries to efficiently process large document collections. Using supervised learning within Python's Annif framework, we trained models on manually classified MARC bibliographic records organized by Dewey Decimal Classification (DDC) standards. The implementation involved collecting and processing records containing titles, summaries, DDC numbers and subject descriptors, then dividing them into training and test datasets. We evaluated four algorithms (TF-IDF, Omikuji, FastText and NN Ensemble) using standard retrieval metrics (F1@5 and NDCG), finding that Omikuji and NN Ensemble significantly outperformed the others in indexing accuracy. The complete open-source framework demonstrates the viability of machine learning for library classification tasks, offering an efficient alternative to manual indexing while maintaining accuracy. These results suggest promising applications for AI in knowledge organization systems, with potential for expansion to other classification schemes and larger datasets to further enhance performance.

**Keywords:** Supervised Machine Learning, Semi-Automated Classification, Automated Subject Indexing, DDC, Annif, Ensemble approach.


E Book ISBN: 978-81-981814-6-6| DOI: https://doi.org/10.34256/vadlibs.25.9.83

## INTRODUCTION

Artificial Intelligence and Machine learning (AI/ML) are arguably the most beneficial technologies to have gained momentum in recent times. An AI/ML-based automated indexing system may help professionals manage skill-oriented, labour-intensive and time-consuming activities related to the processing of documents. The magnitude of collections and the corresponding bibliographic records are accelerating in value, volume and variety in libraries of all types and sizes all over the world. As a result, workloads related to technical processing activities involving classification and subject indexing are increasing manifold. Digitization has changed the way we process and analyze information. There is a gradual increase in online availability of information. From web pages to emails, science journals, e-books, learning content, news and social media, are all full of textual data. The idea is to create, analyze and report information fast. A number of projects on semi-automated systems of classification have already been applied in the field of LIS. This generally means that the system will predict and offer a set of suggestive subject descriptors on the basis of a given vocabulary system (like LCSH, MeSH, Agrovoc, UDC and so on), but the final decision of selecting the appropriate descriptor(s) will be the privilege of a LIS professional (Ahmed et al., 2022). Nowadays modern businesses are using Machine Learning (ML) based solutions to help automate operations and make the whole process of document management faster and more effective. These latest systems are incorporating Artificial Intelligence (AI) to "read" documents like a human identify and classify the type of document and extract key data. Such systems can efficiently and accurately convert the varied content those sources.

The rapid advancements in AI have revolutionised various sectors and libraries are no exception. The integration of AI in libraries has the potential to significantly transform library functions and enhance user experiences (Subaveerap, n.d.). The present society is a post-Industrial society, which represents a new era of innovation in technology, particularly, AI-driven technology. The term "Industry 4.0" typically refers to the present trend of adopting modern technology to automate processes and exchange information (Sarker, 2022). This includes deployment of AI chatbots, intelligent library systems, robots and various other AI applications in library services. With the increasing availability of open-source analysers like Natural Language Processing (NLP) toolkits, and backend algorithms, the potential to transform

libraries into dynamic, user-centered spaces where information is easily accessible and services are tailored to individual needs is increasing at a steady pace. One example of an open-source AI/ML application is the Annif framework for automated indexing, which was developed by the National Library of Finland. Annif combines different open-source tools to predict subject headings or class numbers for documents, based on widely used knowledge organization systems (KOSs) such as LCSH, UDC, MeSH and Agrovoc. This research is an attempt to apply Annif (http://annif.org/) as an open-source AI/ML framework to auto-generate class numbers for documentary resources based on the Dewey Decimal Classification (DDC) scheme. The training dataset was made by accessing Marc bibliographic records from different sources. There are considerations to keep in mind when applying machine learning to automatic subject indexing, such as choosing an appropriate machine learning backend, preparing and pre-processing the dataset, and evaluating the model's performance. However, with the right approach, machine learning can be a powerful tool for automating the process of subject indexing (Ahmed et al., 2022). When using machine learning (ML) for automatic subject indexing, there are a few things to keep in mind, such as: 1) Choosing the right ML algorithm for the task.; 2) Preparing and cleaning the dataset that will be used to train the ML model and 3) Evaluating the model's performance to ensure that it is accurate and reliable. There are considerations to keep in mind when applying machine learning to automatic subject indexing, such as choosing an appropriate machine learning backend, preparing and pre-processing the dataset and evaluating the model's performance.

## REVIEW OF RELATED LITERATURE

Automated document classification, i.e. the construction of a call number by computer, has been one of the dreams of library professionals. After the emergence of computers and more particularly, after the development of artificial intelligence, the hopes of success in automatic classification have reached the apex. This chapter delineates a selective review of literature published in scholarly publications on research carried out in automated document classification. Ranganathan (1965) suggested that the "future work of FID/CR should be to encourage the design to improved schemes of classification, whether they are going to be faceted, analytic synthetic, or whether new brand it will be. In doing this, there should be *frequent consultations with the machine specialists*." Compared to printed media, digital collections present numerous challenges regarding their

preservation, curation, organization and resource discovery and access. According to (Pong et al., 2008), "With the explosive growth in the number of electronic documents available on the Internet and digital libraries, it is increasingly difficult for library practitioners to categorize both electronic documents and traditional library materials using just a manual approach....To improve the effectiveness and efficiency of document categorization at the library setting, more in-depth studies of using automatic document classification methods to categorize library items are required."

Early efforts in automatic subject classification (including subject indexing) date back to the 1970s and various approaches have been explored, including rule-based methods, statistics based methods and information retrieval based methods (Desale & Kumbhar, 2013). Almost all of these early efforts were mainly directed towards term classification – grouping and arraying terms for machine-readable databases to improve computer search efficiency or to retrieve information from the internet. However, more concentrated efforts are seen in literature published since the turn of the 20$^{th}$ Century. Automatic classification research has made prominent progress in the text-categorization (TC) field in recent years (Sebastiani, 2002), especially the supervised machine learning approach, which has shed new light on the resolution of this problem (Golub, 2021). "The most frequent approach to automated classification is machine learning. It, however, requires training documents and performs well on new documents only if these are similar enough to the former" (Golub, 2007). (Cheng & Wu, 1995) introduced ACS: an automatic classification system for school libraries. First, they critically reviewed the various approaches towards automatic classification, namely (i) rule-based, (ii) browse and search and (iii) partial match. The central issues of scheme selection, text analysis and similarity measures were discussed. A novel approach towards detecting book-class similarity with Modified Overlap Coefficient (MOC) was also proposed. Finally, the design and implementation of ACS is presented. The test result of over 80% correctness in automatic classification and a cost reduction of 75% compared to manual classification suggested that ACS was highly adoptable. The major problem with these techniques is that they require data in machine readable form, training documents and still they often fail to produce unique class numbers needed to place a document on library shelves. Most of current research in automated classification is in text categorization. Major techniques used in text categorization are: support vector machine models, k-nearest

neighbour, machine learning, frequency measure and weighing technique. The best approach for document classification without training documents, which could be useful for document classification in a library environment, was suggested by Golub (2007). According to her, "in document classification, matching is conducted between a controlled vocabulary and text of documents to be classified. A major advantage of this approach is that it does not require training documents. If using a well-developed classification scheme, it will also be suitable for subject browsing in information retrieval systems. Golub et.al. (2024) conducted a recent study on the effectiveness of semi-automated subject indexing, where they applied the open-source tool **Annif** to over 230,000 Swedish union catalogue records using Dewey Decimal Classification (DDC) as the target system. Five approaches were tested, including lexical algorithms, support vector machines, fast Text, Omikuji Bonsai and an ensemble method. The ensemble approach achieved the highest accuracy (66.82%) for three-digit DDC classification. A qualitative evaluation involving librarians and LIS students revealed low inter-rater agreement but supported the use of automated classes as useful supplementary access points in information retrieval. The study underscores the practical value of integrating semi-automated indexing into library catalogues.

It is accepted fact in the literature of library and information science that traditional classification schemes are also useful for automated classification. (Gedam & Paradkar, 2013) compared the three web-based document classification schemes, viz., WebDewey 2.0 (Web-based product of DDC (Dewey Decimal Classification), 23rd ed.), UDC (Universal Decimal Classification) Online and Classification Web of LCC (Library of Congress Classification). These schemes are constantly updated, revised and up-to date web-based library classification schemes. It concluded that DDC is the most updated library classification scheme in the world. (Walt, 1997) highlighted the advantages of library classification schemes for organization of information resources in the web environment. According to Van der Walt, the knowledge organization tools developed and used by web search engines often feature shallow hierarchies and uneven coverage of topics. On the other hand, web search engines often respond to popular topics more quickly than traditional library knowledge organization tools do. In the context of hierarchical browsing based on a classification scheme, having too many classes assigned to a document would place one document in many different places, which

would create the opposite effect of the original purpose of a classification scheme (grouping similar documents together).

Hunter (2009) [Quoted in Desale & Kumbhar (2010)] wrote that the principles upon which Colon Classification is based are important. It is clear from literature that hierarchical library classification schemes are useful for hierarchical browsing but are of little use in automatically producing class numbers. Faceted classification based on sound principles might be useful in automation. (Kim & Lee, 2002) designed a knowledge base for an automatic classification in the library and information science field, by using the facet classification principles of Colon Classification. (Nasir Uddin & Janecek, 2007) found that "faceted classification allows the users of a website to access information more efficiently than the simple taxonomic hierarchy of information object." (Panigrahi & Prasad, 2007) demonstrated the techniques of fixing the facet sequence in developing an automatic classification system to construct classification numbers for document titles, which appear in natural language. These studies suggest that little success is achieved in automatic class number generation by using faceted classification schemes. Rather, a faceted structure is more suitable for automated classification and it also needs a relatively small vocabulary for knowledge representation. (Bianchini, 2023) presented the Wikidata gadget, CCLitBox, for the automated classification of literary authors and works by faceted classification and using Linked Open Data (LOD). The tool reproduced the classification algorithm of Class "O" Literature of the Colon Classification of S. R. Ranganathan and used data freely available in Wikidata to generate Colon Classification class numbers. CCLitBox is completely free and allows any user to classify literary authors and their works. (Halder & Biswas, 2023) have also reported their research study exploring the applications of CCLitBox in generating CC 6th edition-based class numbers for Indian literary works. Panigrahi (2000) argued that natural language processing could be used in the automatic identification of noun phrases from the expressive title. (Kim & Lee, 2002) argued that book titles usually have an immediate connection to their contents in that they often encapsulate the entire work. (Wang, 2003) also had similar views that the title of a document usually summarizes its contents and reveals its central topics. (Gupta et al., 2019) proposed a method that gains knowledge from a large number of words from the books and transforms them into a feature matrix. During transformation, the size of the initial matrix is reduced using Wordnet and Principle Component Analysis. Then, AdaBoost classifier is applied to predict the genres of

the books. (Pokorný, 2018) described a method for machine-based creation of high quality subject indexing and classification for both electronic and print documents using tables of contents (ToCs). (Tsuji, 2019) reported the development of a system for recommending books based on the articles Wikipedia users read in libraries that aims to encourage students to read library books as a more reliable source of information rather than relying on Wikipedia articles. This review indicates that basically three types of research are ongoing on automatic classification: 1) hierarchical classification by using different library classification schemes, 2) text categorization and document categorization by using different type of classifiers with or without using training documents and 3) automatic document classification. Predominantly, the research is directed towards solving problems of organization of digital documents in an online environment.

## OBJECTIVES

This research can be recognised as a pilot study to test whether automatic class number generation (at least up to third summary divisions) is possible using the DDC classification scheme by applying machine learning technologies.

The objectives of the research can be summarised as follows: -

- To load Dewey Decimal Classification as Linked Open Dataset inside the AI/ML framework (here Annif).
- To prepare a large bibliographic dataset covering Marc bibliographic records from different sources across the world, preferably with subject descriptors, abstract/ summary notes and DDC notations in a format suitable for Annif.
- To examine and assess the accuracy of subject descriptors and class numbers suggested by Annif, and to design a mechanism for large-scale use of the framework.

The tasks to accomplish these specified goals can be divided into the following categories: a)Marc data collection and pre-processing ; b) preparing the text-corpus from which the training dataset and vocabulary would be formed; c)designing the Annif framework by selecting appropriate analyzer and backend algorithm; d)developing a sizeable training dataset; e) preparing a backend KOS, here DDC in SKOS format, to feed into the framework and f) testing and measuring the model framework's indexing efficacy using a test dataset.

## METHODOLOGY

The foregoing section of this study discusses about the broad tasks that were required to be performed keeping in view of the fulfilment of the research objectives. The activities related to accomplishing the objectives could be broadly classified into the following sub-sections: 1) First was to collect as many MARC formatted bibliographic records as possible, preferably with DDC class numbers (tag 082), title of the document (tag 245), subject descriptors (tag 650 $a) and summary notes (tag 520 $a); 2)merge MARC files to generate a single consolidated file and then export the file in a format suitable(tsv or csv) for OpenRefine data wrangling software by using the MarcEdit tool; 3)Curating the Marc data and making a Text-corpus (a combined field of title and summary note); 4) Creating a training dataset and a test dataset in the format suitable for Annif framework as given in ;5)As DDC vocabulary is not available as LOD, a vocabulary was created using the subject descriptors and their corresponding notations (with fictitious created URI's of those subject descriptors) available in the Marc data in the format suitable for Annif framework; 6)load the SKOS-compliant vocabulary (here DDC) generated earlier into the Annif framework; 7)train the framework with the curated MARC file training dataset and finally 8)measure the system's indexing efficiency using a set of appropriate retrieval metrics. The veracity of the afore-mentioned procedures can be justified by the fact that a similar approach was also followed by Golub et.al. (2024) in their study.

### *Preparation of Virtual Environment*

Continual work is dedicated to discovering the most effective algorithms and setups for optimal quality (Suominen et al., 2022). Manually indexing documents for subject-based access is a labour-intensive process that can be automated using AI technology. Annif (present version 1.0) is an open-source tool designed for automated subject indexing and classification. Annif can be installed through three methods- 1) Virtual Box install [This process requires to install Virtual box software for Windows, Linux or Mac. It is best recommended for most of the users as it is one of the easiest way to install Annif and work with it.] ;2) Docker install [This process requires the Docker software to be installed on the local machine and it is a good way of getting Annif set up with all the dependencies, included in a pre-built container.] and finally the 3) Local linux install [It requires a recent version of Python installation with support for virtual environments and suitable for experienced Linux-users as it allows maximum

flexibility]. In this research study, I have installed Annif using the third method, i.e., the local Linux installation. As already mentioned earlier, this process is suitable for those people who have some prior knowledge and experience in using Linux command line and also familiar with installing Python packages.

The Annif official website (https://annif.org/) provides a user-friendly and collaborative environment, by making users aware of the installation process, usage of the toolkit, the different approaches it adopts. It leverages a blend of existing natural language processing and machine learning tools. The fundamental Annif package encompasses various learning backends such as TF-IDF, analyzers like Simplemma and components like TensorFlow and Gensim. Additionally, the framework can utilize NLTK punctuation rules (punkt) and advanced backend algorithms like FastText, Omikuji, neural network ensemble, among others. It is capable of handling multiple languages and is adaptable to any subject vocabulary, whether in SKOS or a straightforward TSV format. Annif offers a command-line interface, a user-friendly Web UI and a microservice-style REST API. In this study I have made use of four machine learning backend algorithms which are popularly known as associative approaches to machine learning– TF-IDF (statistical method, term frequencies), FastText (neural machine learning algorithm for text classification), Omikuji (tree-based machine learning algorithm for classification) and NN Ensemble (combination of the three above algorithms).

***Creation of the Vocabulary***

The Annif framework needs a structured standard vocabulary to start with. In Annif, a standard vocabulary may be added in two ways: 1) feeding a SKOS-compliant vocabulary in any common RDF serialization format (like RDF/XML (.xml), N-Triple (.nt), Turtle (.ttl), etc.); or 2) using a vocabulary file in a UTF-8 encoded TSV file, where the first column contains a subject URI and the second column includes the corresponding label (subject descriptor) (Ahmed et al., 2022). Most of the standard vocabularies that are being used in the LIS domain, like LCSH, MeSH, Agrovoc and the UNESCO thesaurus are available as SKOS compliant KOS. But in this aspect, where I have focussed on the automatic classification using DDC, there is an additional workload of preparing the TSV file in an Annif compliant format. Constructing a Vocabulary is critical to invest in both the data structure and the data used to populate the data elements in that

structure; the data should survive through a succession of computer systems over time.

**Table 1:** Components of a virtual framework

| **Target** | **Dataset and Tools** | **Process and Purpose** |
|---|---|---|
| Framework for automated subject indexing | Python Virtual Environment (Python 3.8.16 version and PIP) | Requires to install and configure Python virtual environment with Python (3.8+) and PIP (22.0+)for Annif and its associated components. |
| | Annif (version 1.0) (with NLP and ML tools) https://github.com/NatLibFi/Annif/ | The main component of the framework is available as an open-source tool including components like TensorFlow and Gensim |
| | Language Models and Tools (Annif virtual environment will select appropriate versions) | NLTK model for punctuation rules (punkt) and machine learning backends like fastText; Omikuji and Neural network ensemble. |

**Table 2:** Structure of the vocabulary dataset required for the framework

| | | |
|---|---|---|
| <http://dewey.inf/class/342/e23/> | Constitutional law. | 342 |
| <http://dewey.inf/class/028.9/e23/> | Books and reading -- Technological innovations. | 028.9 |

The Annif framework gnerally supports three forms of vocabulary formats - 1) Subject vocabulary as TSV; 2) Subject vocabulary as CSV and 3) Subject vocabulary as SKOS. For this research, I have selected the TSV format. Annif doesn't care much about the internal structure of a subject vocabulary, it just needs to know the URIs and preferred labels (terms or descriptors) of each subject/class/concept. If the vocabulary includes also notion codes, e.g., as in any classification schemes, also they can be given. The format for subject vocabulary as prescribed by Annif includes three columns of which the first column contains a subject URI, the second column its label and the third column is the notation code as shown in 'Figure2'. Finally, this vocabulary would now be required to load inside the annif framework. The command to feed the ready vocabulary inside the Annif framework is -

*annif load-vocab <path/to/TTL file>*.

***Preparation of Training Dataset***

The accuracy and efficiency of any model is directly proportional to the quality of the training dataset. The quality of this data has profound implications for the model's subsequent development, setting a powerful example for all future applications that use the same training data. Just as humans learn better with practical examples, machine also needs a set of data to identify the patterns and learn from it. Any application of Artificial Intelligence or Machine Learning requires certain basic requirements, one of them being the training dataset. Training dataset refers to the initial set of data fed to any machine learning model from which the the final model is created (Rizzoli, 2022). The framework requires a training dataset as a TSV file with the first column containing a text corpus and the second column containing the URIs of the subject descriptors.

**Table 3:** Datasets and Tools to develop the KOS

| Target | Dataset and Tools | Process and Purpose |
|---|---|---|
| Vocabulary dataset preparation | MARC Bibliographic Database for DDC and Text corpus. | The SKOS-compliant DDC in TSV format is deployed to develop the backend KOS for the framework. |
| | MarcEdit Tool | To convert MARC records from XML format to MRC format; By using MarcEditor application, the records were then converted into TSV format so that it could be worked upon by Openrefine. |

(to be enclosed with angular brackets <>). Unlike other Knowledge organisation systems like LCSH or UDC, whose vocabulary are readily available over the internet under OdbL licensing, DDC neither have its own vocabulary nor it has its own URI for subject descriptors. Therefore, in this research work there was the need to create fictitious URI against the subject descriptors (Tag 650). This was a very difficult task because of the lack of standardisation of Marc data.

After the URI was created against the subject descriptors using the DDC class numbers, the training dataset was formed in the format prescribed by Annif. Annif refers this training dataset as document corpus. This corpus is needed for training statistical or machine learning based models as well as for evaluating how well those models work. Annif supports two document corpus formats: one that is more suitable for longer documents (full text or long abstracts) and another that is better suited for short texts such as when the document titles are only available. I have used here the short text document corpus (TSV file) which is especially useful for metadata about documents, when only titles are known, or for very short documents. The final data set has been prepared using Data Wrangling software named as

OpenRefine with the help of a Python/jython script. OpenRefine, an open-source data wrangling software, allows us to select only the rows having certain tags. The first column contains the text of the document (e.g., title or title + abstract) while the second column contains a whitespace-separated list of subject URIs (again within angle brackets) for that document. Another important thing to remember here is that there should not be any column name in the final training dataset. The final dataset contains 373454 (around 0.3 million) bibliographic records (out of 381237 of gathered data) with titles (tag 245), notes (block 5xx, especially tag 520) and assigned subject descriptors from DDC (tag 650).

**Table 4:** Final Training Dataset inside Openrefine

| | | |
|---|---|---|
| 1. | Anti-racist Shakespeare ## Anti-Racist Shakespeare argues that Shakespeare is a productive site to cultivate an anti-racist pedagogy. Our study outlines the necessary theoretical foundations for educators to develop a critical understanding of the longue durée of racial formation so that they can implement anti-racist pedagogical strategies and interventions in their classrooms. This Element advances teaching Shakespeare through race and anti-racism in order to expose students to the unequal structures of power and domination that are systemically reproduced within society, culture, academic disciplines, and classrooms. We contend that this approach to teaching Shakespeare and race empowers students not only to see these paradigms but also to take action by challenging and overturning them. This title is also available as Open Access on Cambridge Core. | <http://dewey.info/class/822.3/3/e23/> |
| 2. | Industrial agriculture and ape conservation ## Social and economic systems worldwide are changing rapidly. These changes are accompanied by an increasing demand for natural resources, including land, water, minerals, energy sources, food and timber. Today's foremost challenge lies in finding the tools to address the complexity of these interrelated trends, and in implementing strategies to balance environmental and socioeconomic needs. This volume contributes to this search by presenting original research, topical case studies and emerging best practice from a range of key stakeholders to examine the interface between ape conservation and industrial agriculture. In assessing the drivers behind agricultural expansion and land investments, it sheds light on governance challenges and legal frameworks that shape land use. Intended for policy makers, industry experts, decision makers, academics, researchers and NGOs, it is designed to inform debate, practice and policy to help reconcile the goals of industrial agriculture with those of ape conservation and welfare, and social and economic development. | <http://dewey.info/class/599.88/e23/> |
| 3. | Malarial subjects : empire, medicine and nonhumans in British India, 1820-1909 ## Malaria was considered one of the most widespread disease-causing entities in the nineteenth century. It was associated with a variety of frailties far beyond fevers, ranging from idiocy to impotence. And yet, it was not a self-contained category. The reconsolidation of malaria as a diagnostic category during this period happened within a wider context in which cinchona plants and their most valuable extract, quinine, were reinforced as objects of natural knowledge and social control. In India, the exigencies and apparatuses of British imperial rule occasioned the close interactions between these histories. In the process, British imperial rule became entangled with a network of nonhumans that included, apart from cinchona plants and the drug quinine, a range of objects described as malarial, as well as mosquitoes. Malarial Subjects explores this history of the co-constitution of a cure and disease, of British colonial rule and nonhumans, and of science, medicine and empire. This title is also available as Open Access. | <http://dewey.info/class/616.9/362/009541409034/e23/> |
| 4. | Academic brands : distinction in global higher education ## The first comprehensive analysis of the emergence of academic brands, this book explores how the modern university is being transformed in an increasingly global economy of higher education where luxury is replacing access. More than just a sign of corporatization and privatization, academic brands provide a unique window on the university's concerns and struggles with conveying 'excellence' and reputation in a competitive landscape organized by rankings, while also capitalizing on its brand to generate revenue when state support dwindles. This multidisciplinary volume addresses topics including the uniqueness of academic brands, their role in the global brand economy of distinction, and their vulnerability to problematic social and political associations. By focusing on brands, the volume analyzes the tensions between the university's traditional commitment to public interest values - education, research, and the production of knowledge - and its increasingly managerial culture framed by corporate, private values. Available as Open Access on Cambridge Core. | <http://dewey.info/class/378.1/01/e23/> |
| 5. | Law and policy for the quantum age ## It is often said that quantum technologies are poised to change the world as we know it, but cutting through the hype, what will quantum technologies actually mean for countries and their citizens? In Law and Policy for the Quantum Age, Chris Jay Hoofnagle and Simson L. Garfinkel explain the genesis of quantum information science (QIS) and the resulting quantum technologies that are most exciting: quantum sensing, computing, and communication. This groundbreaking, timely text explains how quantum technologies work, how countries will likely employ QIS for future national defense and what the legal landscapes will be for these nations, and how companies might (or might not) profit from the technology. Hoofnagle and Garfinkel argue that the consequences of QIS are so profound that we must begin planning for them today. | <http://dewey.info/class/530.12/e23/> |
| 6. | Trademark and unfair competition conflicts : historical-comparative, doctrinal, and economic perspectives ## With the rise of internet marketing and e-commerce around the world, international and cross-border conflicts in trademark and unfair competition law have become increasingly important. In this groundbreaking work, Tim Dornis - who, in addition to his scholarly pursuits, has worked as an attorney, a public prosecutor, and a judge, giving him experience in both civil and common-law jurisdictions - presents the historical-comparative, doctrinal, and economic aspects of trademark and unfair competition conflicts law. The book should be read by any scholar or practitioner interested in the international aspects of intellectual property generally, and trademark and unfair competition law specifically. This title is available as Open Access. | <http://dewey.info/class/346.04/88/e23/> |
| 7. | Open access and the humanities : contexts, controversies and the future ## If you work in a university, you are almost certain to have heard the term 'open access' in the past couple of years. You may also have heard either that it is the utopian answer to all the problems of research dissemination or perhaps that it marks the beginning of an apocalyptic new era of 'pay-to-say' publishing. In this book, Martin Paul Eve sets out the histories, contexts and controversies for open access, specifically in the humanities. Broaching practical elements alongside economic histories, open licensing, monographs | <http://dewey.info/class/001.30285/e23/> |

javascript:{}

Once training dataset was created, it's time to formulate the test dataset. This dataset evaluates the performance of the model and ensures that the model can generalize well with the new or unseen dataset. In my research study I have taken around 0.02% (around 1800 records) of the training dataset to form the test dataset. The test dataset provides the gold standard used to evaluate the model.

***Measuring the Prediction Efficiencies***

The complex process of creating these datasets will only be fruitful only when the predictive automated subject headings and class numbers would be accurate enough. The next step is to train the

machine with a dataset. Before starting the training procedure, a project needs to be created and configured separately for each algorithm defined in the 'projects.cfg' file located under in the current directory where Annif is executed. A project here is used to tell Annif which kind of algorithm vocabulary language and other settings are required for training. The projects are usually identified by project Id's that are usually a short string, for example, 'ddc-tfidf'. All the settings that are configured with each backend algorithm are compulsory for Annif framework. As mentioned earlier Annif supports a number of backend algorithms and retrieval metrics, each having its own set of advantages and disadvantages. After successful training of the projects inside the virtual environment, Annif can predict not only subject descriptor(s) as well as DDC class number(s) against a given text corpus and can rank those according to accuracy scores based on those metrics, where the scores ranges from 0 to 1 (Figure 4).

**Table 5:** Complete Training of projects

```
(annif-venv) dlisku@dlisku-HP-280-Pro-G6-Microtower-PC:~/annif$ annif list-projects
Project ID    Project Name               Vocabulary ID  Language  Trained  Modification time
-----------------------------------------------------------------------------------------------
ddc-tfidf     DDC TFIDF project          ddc            en        True     2023-09-18 16:44:43
ddc-fastText  DDC fastText English       ddc            en        True     2023-09-18 19:37:28
ddc-omikuji   DDC Omikuji Bonsai Project ddc            en        True     2023-09-18 20:33:12
ddc-nn        DDC NN ensemble English    ddc            en        True     2023-09-19 19:39:53
(annif-venv) dlisku@dlisku-HP-280-Pro-G6-Microtower-PC:~/annif$
```

**Table 6:** Automatic indexing by TF-IDF model along with their accuracy scores.

```
meghna@meghna-inspiron-15-3567: ~/annif
(annif-venv) meghna@meghna-Inspiron-15-3567:~/annif$ echo "Classical and medieval literature criticism.$nVolume 191  ## Presents literary criticism on
the works of classical and medieval philosophers, poets, playwrights, political leaders, scientists, mathematicians, and writers from other genres. Cri
tical essays are selected from leading sources, including published journals, magazines, books, reviews, and scholarly papers. Criticism includes early
 views from the author's lifetime as well as later views, including extensive collections of contemporary analysis." | annif suggest ddc-tfidf
<http://dewey.inf/class/880.09/e23/>    English literature -- Medieval influences.      880.09  0.9617
<http://dewey.inf/class/809.005/e23/>   English literature -- Medieval influences.      809.005 0.9501
<http://dewey.inf/class/820.9003/e23/>  English literature -- Medieval influences.      820.9003        0.9501
<http://dewey.inf/class/820.9001/e23/>  Epic literature, English.       820.9001        0.9501
<http://dewey.inf/class/820.9/e23/>     English literature -- Medieval influences.      820.9   0.8973
<http://dewey.inf/class/813.309/e23/>   American fiction -- 19th century -- History and criticism.      813.309 0.7375
<http://dewey.inf/class/813.034/e23/>   American fiction -- 19th century -- History and criticism.      813.034 0.7063
<http://dewey.inf/class/820.9008/e23/>  Literature -- History and criticism$vEarly works to 1800.       820.9008        0.7063
<http://dewey.inf/class/809/e23/>       Comparative literature. 809     0.6986
<http://dewey.inf/class/809.034/e23/>   Literature, Modern--19th century.        809.034 0.6670
(annif-venv) meghna@meghna-Inspiron-15-3567:~/annif$
```

## FINDINGS

Subject indexing, a fundamental task in library and information science (LIS), refers to the systematic process of assigning specific terms or descriptors to documents or resources in order to represent their

content accurately. It plays a crucial role in facilitating efficient information retrieval for library users. The process of subject indexing is a complex one, even for the same subject content, two LIS professionals may not assign the same descriptors. This indicates that subject indexing is not a purely objective or standardized process. It involves a degree of interpretation, judgment and expertise on the part of the professionals involved. Several factors contribute to this variability. Different professionals may have varying levels of expertise, experience, or knowledge in a particular subject area. They may also approach the indexing process with different perspectives, preferences, or interpretations of the content. Same principle is applied in case of machine leaning models, where the report provided by each backend differs from the other one.

**Table 7:** Comparison of performances for different backends

| **Retrieval metrics** | **NN Ensemble** | **Omikuji** | **FastText** | **TF-IDF** |
|---|---|---|---|---|
| **Precision (doc avg):** | **0.14** | **0.07** | **0.06** | **0.05** |
| **Recall (doc avg):** | **0.61** | **0.74** | **0.56** | **0.54** |
| **F1 score (doc avg):** | 0.22 | 0.13 | 0.1 | 0.1 |
| **Precision (subj avg):** | 0 | 0 | 0 | 0 |
| **Recall (subj avg):** | 0 | 0 | 0 | 0 |
| **F1 score (subj avg):** | 0 | 0 | 0 | 0 |
| **Precision (weighted subj avg):** | 0.24 | 0.18 | 0.13 | 0.19 |
| **Recall (weighted subj avg):** | 0.61 | 0.74 | 0.56 | 0.54 |
| **F1 score** | 0.32 | 0.26 | 0.19 | 0.25 |

| | | | | |
|---|---|---|---|---|
| **(weighted subj avg):** | | | | |
| **Precision (microavg):** | 0.11 | 0.07 | 0.06 | 0.05 |
| **Recall (microavg):** | 0.61 | 0.74 | 0.56 | 0.54 |
| **F1 score (microavg):** | 0.19 | 0.13 | 0.1 | 0.1 |
| **F1@5:** | **0.24** | **0.22** | **0.16** | **0.14** |
| **NDCG:** | **0.48** | **0.55** | **0.4** | **0.35** |
| **NDCG@5:** | 0.47 | 0.53 | 0.37 | 0.31 |
| **NDCG@10:** | 0.48 | 0.55 | 0.4 | 0.35 |
| **Precision@1:** | 0.34 | 0.37 | 0.25 | 0.18 |
| **Precision@3:** | 0.19 | 0.19 | 0.14 | 0.12 |
| **Precision@5:** | 0.16 | 0.13 | 0.09 | 0.09 |
| **True positives:** | **1122** | **1367** | **1036** | **1001** |
| **False positives:** | **8781** | **17053** | **17350** | **17419** |
| **False negatives:** | **720** | **475** | **806** | **841** |
| **Documents evaluated:** | 1842 | 1842 | 1842 | 1842 |

The output report comes with several metrics along with their score, as because automated indexing is a multi-label classification problem. There are various alternatives on how to compute metrics in detail. This is why the final report includes Precision, Recall and F1 scores obtained with different averaging ways. The test dataset with human-assigned index terms (as the Gold standard) is utilized in OpenRefine to generate suggested descriptors in Annif by using using different backend algorithms. A comparative study of the scores, generated on the basis of an array of retrieval metrics by the *'eval'* command for the major backend algorithms, is given in the Table 9 to understand the relative performances.

## CONCLUSION

The wiki of Annif says that the two most important values from the array of retrieval results are F1 @5 and NDCG. Generally, NDCG and F1 @5 are two of the best retrieval metrics because they are graded, normalized and easy to understand and interpret. NDCG, for instance, stands out for its ability to consider both the relevance and position of retrieved items in a ranked list. This characteristic is particularly crucial in applications like recommendation systems search engines and even in automated indexing systems, where the order of results significantly impacts user satisfaction. On the other hand, F1@5 strikes a balance between precision and recall, providing a comprehensive evaluation of system performance, especially in the context of the initial search results. The scores of these models ranges from 0.0-1.0, the closer to 1.0, the better and more accurate the results are. Therefore, while evaluating the performance of the backend models, more emphasis is being given on NDCG and F1@5. The comparative scores for these retrieval metrics show that the Omikuji and NN-Ensemble backends of the AI/ML-based automated indexing framework have performed better than the TF-IDF and fastText backend, considering the evaluation parameters (given in bold text) against human-assigned indexing terms that were considered as the 'Gold standard'. On a similar note, Golub et.al. (2024), in their study, ranked the ensemble approach highest in terms of retrieval performance.

The AI/ML based indexing system, although in its early stages of development, is already making strides by demonstrating remarkable potential. It's akin to a young seedling, just beginning to sprout in the soil. In this dynamic environment, the AL/ML-based indexing system emerges as a promising technological innovation. It serves as a digital assistant, equipped with advanced algorithms and machine learning capabilities. This system promises to revolutionize how libraries manage and organize their extensive collections, moving towards a new era of efficiency and accessibility for library users. Till now, the AI/ML based applications have been either commercial projects or large-scale organizational initiatives, but with the advancement of open-source software and open datasets, the horizons have expanded. By embracing these emerging technologies, LIS professionals and schools have the chance to not only enhance their proficiency but also to significantly elevate their capabilities in managing and leveraging

data. Essentially, the research aims to shed light on the potential and capabilities of this AI and ML tool, Annif.

## FUTURE SCOPE

The study reflects the promising fusion of two critical fields: data carpentry and AI/ML-based knowledge processing, within the domain of Library and Information Science (LIS). This convergence is anticipated to bring about significant advancements in the field of LIS. In the upcoming times, it lays the foundation for creating systems that can autonomously generate class numbers and appropriate subject descriptors for extensive collections of documents. It is expected to revolutionize how information is organized and categorized in the realm of LIS. This signifies a major leap forward in automating and streamlining information management processes in the field. Some of the future possibilities includes: auto-generation of UDC-based class numbers (as UDC summary is available as a LOD dataset); using MeSH for developing automated indexing systems for bio-medical literature (MeSH is available as a LOD dataset), and so on. There are thousands of open datasets of bibliographic records available in the web (https://lod-cloud.net/). Thus, they can be used to integrate with AI/ML technologies and provide insightful facilites in different subject domains. The Annif framework may also be utilized by integrating it with Koha as the backend indexing system to generate suggestions for the subject access field (Tag 650) on the basis of title (Tag 245$a) and summary note (Tag 520$a). It can also be incorporated within an institutional repository system like Dspace or Eprints for populating the DC.Subject metadata element automatically on the basis of text input in the DC.Title and DC.Description. Additionally, it can be also used as recommender system in Koha OPAC, that is the very feature of a Library Discovery service, that would eventually provide a new perspective to LIS professionals.

***Acknowledgement***

I would like to express my sincere gratitude to Prof. Parthasarathi Mukhopadhyay, from the Department of Library and Information Science at the University of Kalyani, for his invaluable guidance and support throughout my research work.

## REFERENCES

[1] Ahmed, M., Mukhopadhyay, M., & Mukhopadhyay, P. (2022). Automated knowledge organization: AI/ML-based subject indexing system for libraries. *DESIDOC Journal of Library & Information Technology, 42*(1), 75–82.

[2] Bianchini, C. (2023). CCLitBox: A Wikidata gadget to classify world literature. *SRELS Journal of Information Management, 60*(3), 133–141. https://doi.org/10.17821/srels/2023/v60i3/171024

[3] Cheng, P. T. K., & Wu, A. K. W. (1995). ACS: An automatic classification system. *Journal of Information Science, 21*(4), 289–299. https://doi.org/10.1177/016555159502100405

[4] Desale, S. K., & Kumbhar, R. M. (2013). Research on automatic classification of documents in library environment: A literature review. *Knowledge Organization, 40*(5), 295–304. https://doi.org/10.5771/0943-7444-2013-5-295

[5] Gedam, P. B., & Paradkar, A. (2013). A study of web-based library classification schemes. *International Journal of Library and Information Science, 5*(10), 386–393. https://doi.org/10.5897/IJLIS2013.0336

[6] Golub, K. (2007). Automated subject classification of textual documents in the context of web-based hierarchical browsing. *Knowledge Organization, 34*(3), 230–244. https://doi.org/10.5771/0943-7444-2011-3-230

[7] Golub, K. (2021). Automated subject indexing: An overview. *Cataloging & Classification Quarterly, 59*(8), 702–719. https://doi.org/10.1080/01639374.2021.2012311

[8] Golub, K., Suominen, O., Mohammed, A. T., Aagaard, H., & Osterman, O. (2024). Automated Dewey Decimal Classification of Swedish library metadata using Annif software. *Journal of Documentation, 80*(5), 1057–1079. https://doi.org/10.1108/JD-01-2022-0026

[9] Gupta, S., Agarwal, M., & Jain, S. (2019). Automated genre classification of books using machine learning and natural language processing. In *2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence)* (pp. 269–272). IEEE. https://doi.org/10.1109/CONFLUENCE.2019.8776935

[10] Halder, D., & Biswas, M. (2023). Machine-generated colon class numbers: Automatic classification of Indian literary works in the Wikidata environment. *Journal of Information and Knowledge, 63*(3), 143–149. https://doi.org/10.17821/srels/2023/v60i3/171025

[11] Kim, J., & Lee, K. (2002). Designing a knowledge base for automatic book classification. *The Electronic Library, 20*(6), 488–495. https://doi.org/10.1108/02640470210454010

[12] Nasir Uddin, M., & Janecek, P. (2007). Faceted classification in web information architecture: A framework for using semantic web tools. *The Electronic Library, 25*(2), 219–233. https://doi.org/10.1108/02640470710741340

[13] Panigrahi, P., & Prasad, A. R. D. (2007). Facet sequence in analytico-synthetic scheme: A study for developing an AI-based automatic classification system. *Proceedings of the International Conference on Semantic Web & Digital Libraries.*

[14] Pokorný, J. (2018, June 15). Automatic subject indexing and classification using text recognition and computer-based analysis of tables of contents. In *22nd International Conference on Electronic Publishing*. https://doi.org/10.4000/proceedings.elpub.2018.19

[15] Pong, J. Y.-H., Kwok, R. C.-W., Lau, R. Y.-K., Hao, J.-X., & Wong, P. C.-C. (2008). A comparative study of two automatic document classification methods in a library setting. *Journal of Information Science, 34*(2), 213–230. https://doi.org/10.1177/0165551507082592

[16] Ranganathan, S. R. (1965). Discussion on Neelameghan and Rigby. In P. Atherton (Ed.), *Classification research: Proceedings of the Second International Study Conference* (pp. 540–542). Copenhagen: Munksgaard.

[17] Rizzoli, A. (2022). Training data quality: Why it matters in machine learning. *V7 Labs*. https://www.v7labs.com/blog/quality-training-data-for-machine-learning-guide

[18] Sarker, I. H. (2022). AI-based modeling: Techniques, applications and research issues towards automation, intelligent and smart systems. *SN Computer Science, 3*(2), 158. https://doi.org/10.1007/s42979-022-01043-x

[19] Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys, 34*(1), 1–47. https://doi.org/10.1145/505282.505283

[20] Subaveerap, I. A. (n.d.). Application of artificial intelligence (AI) in libraries and its impact on library operations: A review. *International Journal of Library and Information Science*. (Please update with publication year and volume/issue if available.)

[21] Suominen, O., Lehtinen, M., & Inkinen, J. (2022). Annif and Finto AI: Developing and implementing automated subject indexing. *JLIS.it, 13*(1), 1–11. https://doi.org/10.4403/jlis.it-12740

[22] Tsuji, K. (2019). Automatic classification of Wikipedia articles by using convolutional neural network. *Qualitative and Quantitative Methods in Libraries, 6*(3). http://www.qqml-journal.net/index.php/qqml/article/view/566

[23] Van der Walt, M. S. (1997). The role of classification in information retrieval on the Internet. *ShaneyCrawford.com*. https://www.shaneycrawford.com/2003/02/the-role-of-classification-in-information-retrieval-on-the-internet-by-marthinus-s-van-der-walt/

[24] Wang, J. (2003). A knowledge network constructed by integrating classification, thesaurus, and metadata in digital library. *International Information & Library Review, 35*(2–4), 383–397. https://doi.org/10.1080/10572317.2003.10762613

***Published by*** Vyom Hans Publications,
***ISBN*** (Digital Download and Online): 978-81-981814-6-6
***ISBN*** (Book): 978-81-981814-5-9
***Year:*** April, 2025
***DOI:*** https://doi.org/10.34256/vadlibs.25.9.83

## HOW TO CITE

Meghna Biswas. (2025). Semi-Automated Class Number Prediction of Bibliographical Resources: A Framework Deploying Annif. Ed. by Kumar, J. & Sedam, M. V. in "Advancing Library and Information Science: Innovations, Practices, and Future Directions" (pp. 83-104).Vyom Hans Publications. https://doi.org/10.34256/vadlibs.25.9.83;

E-ISBN: 978-81-981814-6-6