ISSN (Print) : 0972-2467
ISSN (Online) : 0976-2477
Journal of Information and Knowledge, Vol 60(2), April 2023, p.69-76
DOI: 10.17821/srels/2023/v60i2/170963

Automated Subject Identification using the Universal Decimal Classification: The ANN Approach

Aditi Roy[1]* and Saptarshi Ghosh[2]

[1]Research Scholar, Department of Library and Information Science, University of North Bengal, Raja Rammohunpur – 734014, West Bengal, India; rs_aditi@nbu.ac.in
[2]Professor, Department of Library and Information Science, University of North Bengal, Raja Rammohunpur – 734014, West Bengal, India; sghosh@nbu.ac.in

Abstract

Universal Decimal Classification (UDC) is a popular controlled vocabulary that is used to represent subjects of documents. Text categorization determines a text's category, as evident from the notation-text label format of the Universal Decimal Classification. With the help of machine learning techniques and the Universal Decimal Classification (UDC), the present work aims to develop an end-user (library professional) based recommender system for automatically classifying documents using the UDC scheme. The proposed work is conceived for determining and constructing a complex class number using the syntax of Universal Decimal Classification (UDC). A corpus of documents classified with the UDC scheme is used as a training dataset. The classification of the documents is done with human mediation having proficiency in classificatory approaches. The BERT model and the KNIME software are used for the study. This study uses the classified dataset to fine-tune the pre-trained BERT model to construct the semi-automatic classification model. The results show that the model is constructed with high accuracy and Area Under Curve (AUC) value, although the prediction represented a low accuracy rate. This study reflected that if the model is explicitly trained by annotating each concept and if the full licensed version of UDC class numbers becomes available, there is a greater potency of developing an automated, freely faceted classification scheme for practical use.

Keywords: Automatic Classification, BERT Model, KNIME, Multi-Label Classification, UDC (Universal Decimal Classification)

1. Introduction

Automated classification and cataloguing have always been thrust areas of research in the LIS domain. However, classification is a mental process. Therefore, automated Subject Indexing using artificial intelligence that mimics human cognition is an area of contemporary research. This study explored using the Universal Decimal Classification (UDC) scheme at the input level of an AI system and observing the accuracy of prediction by the trained AI model for the feasibility study of AI-driven classification in a real scenario. UDC is a widely used library classification system, based on the Dewey Decimal Classification (DDC) system, developed by the Belgian bibliographer's Paul Outlet and Henri La Fontaine in the late 19th century (British Standards I, 2005) and uses a combination of numbers and letters to categorize knowledge in a systematic and organized manner. However, UDC extends and improves the DDC system by allowing more flexibility in classifying knowledge (Slavic A, 2008). In addition, UDC has a much more extensive vocabulary, with over 70,000 possible subject headings (UDC Consortium, 2022). UDC is designed to be used by librarians and other information professionals to organize collections of books and other materials. UDC classifies all knowledge into ten main categories, further subdivided into more specific subject areas (British Standards I, 2005; Slavic A, 2008). Each subject area is given a unique

*Author for correspondence

notation comprising a combination of numbers and letters ((British Standards I, 2005). One of the critical advantages of UDC is its ability to facilitate cross-disciplinary searching. The hierarchical structure of UDC allows for easy navigation between related subject areas, making it easier to find materials on a particular topic (Slavic A, 2008). In addition, UDC is multilingual, with translations available in over 30 languages, making it a truly global classification system. UDC is also highly adaptable and can be customized to suit the needs of different libraries and information centres. New subject areas can be added to the system, and existing subject areas can be modified or merged. This flexibility has contributed to the continued popularity of UDC in libraries and other information centres worldwide. Overall, UDC is a valuable tool for organizing and accessing knowledge in a systematic and organized manner. Its hierarchical structure, extensive vocabulary, and flexibility make it an ideal classification system for libraries and other information centres (UDC Consortium, 2022).

With the help of machine learning techniques and the Universal Decimal Classification (UDC), this work aims to construct a model for the semi-automated classification of texts using the fine-tuned BERT classifier. UDC's structure is very complex, using several punctuation marks as standard and special auxiliaries (British Standards I, 2005); the classification is limited to multi-label classification rather than multi-class classification (Borovic *et al*., 2022). Multi-label classification is a machine-learning problem in which an instance can be assigned multiple labels from a predefined set of categories. This phenomenon contrasts with traditional binary or multi-class classification, where a model is assigned to a single brand or class (Guo Z, 2021). The proposed work is conceived for determining and constructing a complex class number using the syntax of Universal Decimal Classification (UDC). The digital version of the UDC standard edition in RDF/JSON/Turtle format was unavailable. The UDC MRF (Master Reference File) of 2011 as LOD (Linked Open Data) was available (Slavic *et al*., 2021) at Finto AI, i.e., as a Finnish Triple File where the Object (Obj.) contains text in Finnish, English, and Slovak languages and translation was very troublesome. Conversion of UDC standard edition from pdf to text led to invalid labelling of labels and constituent vocabularies. We could have developed a UDC TTL file, but that would violate the copyright principle of UDC, and thus we used the CSV file converted from the pdf version of the UDC standard edition.

2. Literature Review

The automatic classification was initiated using the DDC scheme, which seems to be the primitive stage of classifying documents using machine learning algorithms. In 2011, an unsupervised approach was used to automatically classify scientific literature using the DDC classification scheme by designing a prototype software system (Joorabchi A and Mahdi, 2011). An attempt was made in 2021 to use the BERT model in the automatic classification of higher educational resources belonging to the DDC 905 class (Schrumpf *et al*., 2021). The structure of UDC is based on that of DDC. UDC Consortium states, "UDC is used in bibliographic services, documentation centres, and libraries in around 130 countries worldwide" (UDC Consortium, 2022). A similar study was carried out in 2020 to assign UDC class numbers to older digitized texts by using machine learning algorithms for which they extracted 70,000 scholarly articles and classified them with the help of human experts, and 80% of that corpus was treated as training data and rest 20% as test data. The study only used the primary classes to assign a class number to the selected sample (Kragelj M and Kljajić Borštnar M, 2021). In 2011, the Multilingual UDC Summary was published as LOD. That UDC MRF file was considered to express UDC using SKOS and RDFS for the edge labels to post Knowledge organization as linked data (Slavic *et al*., 2021). By selecting a text corpus of 1,14,485 texts from Slovenian Open Access Infrastructure with UDC class numbers and considering the UDC MRF file 2011, a recommender system was created by performing BM25 ranking and different hybrid recommendation system configurations for decision support in a semi-automatic cataloguing process (Borovic *et al*., 2022). A domain-specific study on Mathematics was carried out based on OntoMathPro ontology to assign UDC class numbers with three core features of method, equation, and problem (Nevzorova O, 2021).

3. Methodology

3.1 Data Preparation

This step includes the following activities

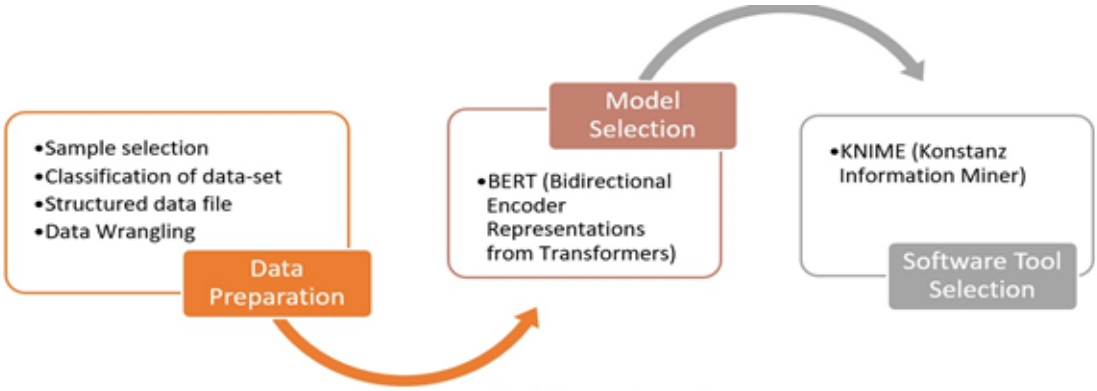


Figure 1. Methodology flow chart.

3.1.1 Sample Selection

For this study, we collected articles published in "Annals of Library and Information Studies" for the last four years, i.e., 2018-2022. A total of 151 articles formed the corpus of our work.

3.1.2 Classification of Dataset

The 151 articles were manually classified using Universal Decimal Classification (UDC) standard edition (2006). Furthermore, the sought terms from the titles of each article were parsed, and notations were assigned using the UDC scheme. Standard vocabularies used in the UDC scheme often must catch up regarding approach terms. For example, no notation is provided for a term like 'Bibliometrics' and thus is required to be made by synthesizing more than one contextual notation for making it an 'approach-to-sought term'. The notation of Bibliometrics was conceived as [(02):519.22] where (02) was taken from "form(Table(d))" common auxiliaries (Table 1d) and 519.22 was taken from mathematics class denoting 'Statistical Measure'. Such occurrences were quite common throughout our study's number construction process.

A few examples are illustrated as follows-

3.1.3 Structured Data File

The next step we followed was creating a CSV file with sought terms and UDC notation. For addressing the approach term, each row of the CSV file having term and UDC notation was further exemplified with a new cell connoting the context of the term with the column header 'Subject.'

3.1.4 Data Wrangling

To remove disambiguated terms, Parts of Speech (POS) removal, and other unsought terms/signs/punctuations, we wrangled the data in Excel using an add-in 'Ablebits Ultimate Suite' and an add-in open-source messy data wrangling tool OpenRefine.

3.2 Model Selection

Our study used the BERT (Bidirectional Encoder

Table 1.

Terminology	Class Number
Scientometrics	[5/6:519.22]
Sentiment Analysis	316.65:303.72
MOOCs	37.018.43:004
Information Seeking Trends	[001.102:001.2]
Law course entrants	[34:37.04]
Information Mashup	[001.102:001.814]
H-index	331.108.52:655.55:519.22

Representations from Transformers) model (Devlin *et al*., 2018; González-Carvajal S and Garrido-Merchán EC, 2020; Gweon H, and Schonlau M, 2022; Koroteev MV, 2021). This model is a pre-trained Natural Language Processing (NLP) model developed by Google in 2018. BERT is based on the Transformer architecture, a deep learning model that uses self-attention mechanisms to process sequential data (Devlin *et al*., 2018; González-Carvajal S and Garrido-Merchán EC, 2020).

What sets BERT apart from other NLP models is its ability to understand the context of words in a sentence (Gweon H, and Schonlau M, 2022; Arora *et al*., 2020). BERT is a bidirectional model, meaning it can process a sentence in both directions (left to right and proper to left), allowing it to capture the full context of the sentence. This pro makes it particularly useful for understanding the nuances of language, such as sarcasm, idioms, and homonyms (Devlin *et al*., 2018; González-Carvajal S and Garrido-Merchán EC, 2020; Koroteev MV, 2021). Using masked language modelling, BERT was trained through a large corpus of text data, including Wikipedia articles and books (Devlin *et al*., 2018). In this approach, some words in a sentence are masked, and the model is trained to predict the missing words based on the surrounding context (Devlin *et al*., 2018; González-Carvajal S and Garrido-Merchán EC, 2020). The pre-trained BERT model can be fine-tuned for specific NLP tasks, such as sentiment analysis, named entity recognition and question answering (Gweon H, and Schonlau M, 2022; Koroteev MV, 2021). By fine-tuning the pre-trained model on a smaller dataset specific to the task at hand, the model can achieve state-of-the-art performance on a wide range of NLP benchmarks (Arora *et al*., 2020). Since its release, BERT has become one of the most widely used NLP models, with applications in various industries, including healthcare, finance, and marketing. Google continues to improve BERT's performance and has released several variants, including RoBERTa, ALBERT, and ELECTRA (Sultan Alrowili E and Vijay-Shanker K, 2021).

3.3 Software Tool Selection

KNIME (Konstanz Information Miner) is a free and open-source data analytics and machine learning software that allows users to create data processing workflows using a drag-and-drop interface. It was first released in 2006 and was developed by the KNIME company based in Konstanz, Germany (Berthold *et al*., 2009). KNIME provides a wide range of tools and techniques for data processing, including data blending, transformation, visualization, and analysis. It also includes advanced machine-learning algorithms for classification, regression, clustering, and text-mining tasks. In addition, the software supports various data formats, including spreadsheets, databases, and big data platforms like Hadoop and Spark. One of the key features of KNIME is its flexibility and extensibility. It allows users to create their nodes and workflows and integrate external tools and libraries using its open API. KNIME also has a large and active community of users who share workflows and extensions through its public hub (KNIME Community Hub). In addition to its open-source version, KNIME also offers a commercial platform called KNIME Analytics Platform, which provides additional features and support for enterprise use. Overall, KNIME is a powerful and user-friendly tool for data analytics and machine learning, suitable for beginners and field experts (Berthold *et al*., 2009).

3.3.1 Designing of Workflow for Data Analytics

The process's first step includes preparing the data table with column rename, i.e., adding the column name subject, then aggregating the columns, and lastly, text processing by string manipulation using regex expression.

The next step of the process includes sampling and partitioning the data table into test and training data for preparing the prediction model.

Lastly, the BERT Model Selector node downloaded a BERT base model with encoded layer 12 (Attention Head 12), Hidden Layer 768M, and 110M Parameters. The pre-trained BERT base model was fed with more specific vocabularies of the UDC classification scheme using BERT Embedder. Contextual embeddings give rise to the relative impact of the model's performance (Arora *et al*., 2020). Therefore, the first BERT Embedder was trained with partitioned trained data, and the second one was trained with classified facets of sample article titles and UDC terms; then, the cross-joiner node joined the trained embedded results. The pre-trained BERT base model and the cross-joiner result are connected to the input node of the BERT Classification Learner node to create a specific model for UDC classification. Next, the

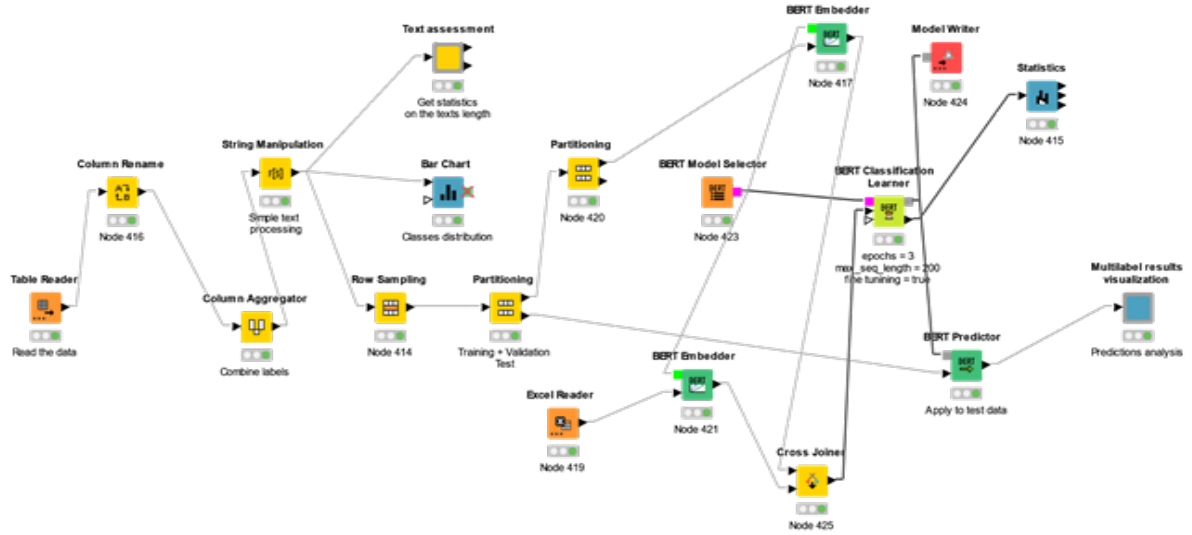


Figure 2. Visual representation of KNIME nodes.

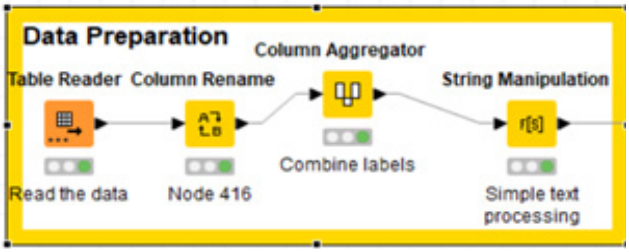


Figure 3. Step 1 of model preparation.

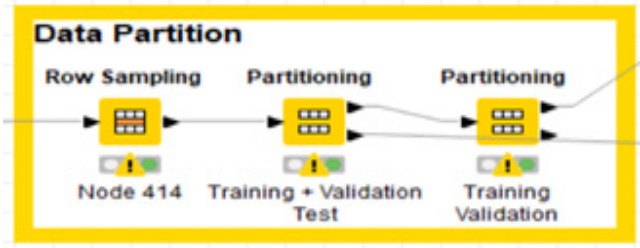


Figure 4. Step 2 of model preparation.

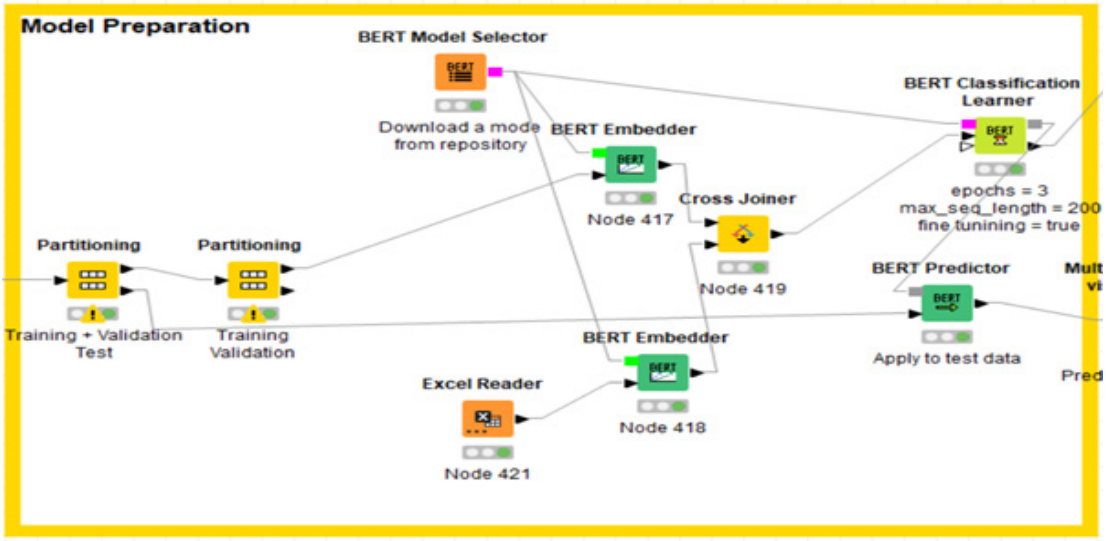


Figure 5. Step 3 of model preparation.

final learned model is connected to the input model node of the BERT Predictor to predict the result of the test partition data table.

4. Results and Discussion

After training the BERT learner node, Figure 6 i.e., the model content represents newly embedded classes

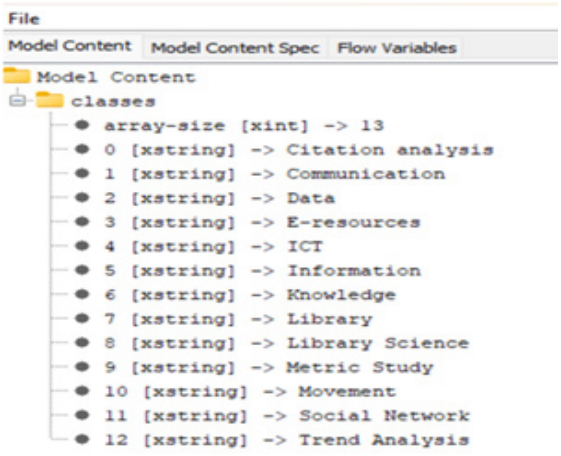
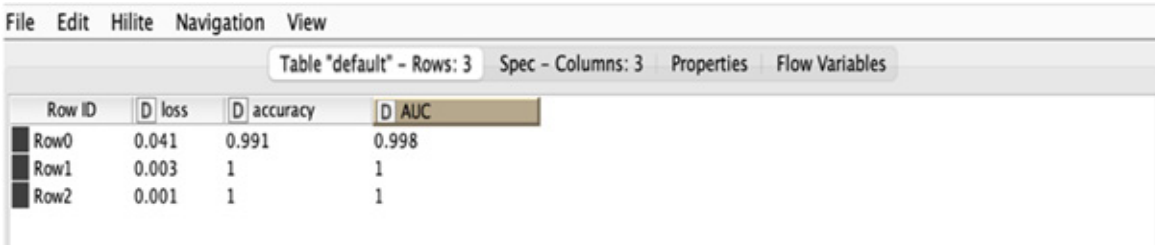


Figure 6. Model content.

required for sample prediction. The model represented the subjects of each content.

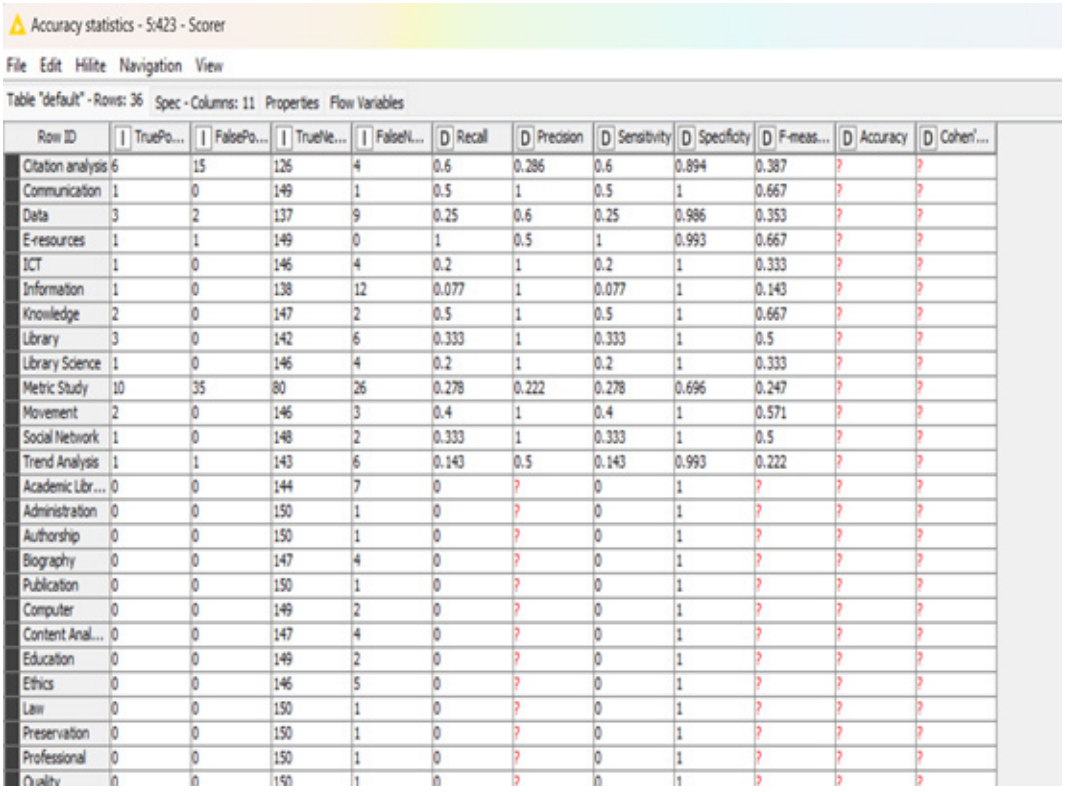
In the input label, our total subjects were 49, and the learner node learned 13 (array size), maintaining homogeneity.

Figure 7 shows the accuracy of the BERT learner model, which depicts that the loss is minimal, accuracy is 1, and AUC is also 1. AUC, i.e., Area Under the Curve, has a statistical significance: AUC represents the degree or measure of separability. It represents the capability of the model in distinguishing between classes. AUC near one means it has a good separability measure (Narkhede S, 2018). Higher the AUC, the better the prediction. In our case, AUC and Accuracy stand one, meaning the model is trained very well.



Row ID	loss	accuracy	AUC
Row0	0.041	0.991	0.998
Row1	0.003	1	1
Row2	0.001	1	1

Figure 7. Statistics of the BERT learner model.



Row ID	TruePo...	FalsePo...	TrueNe...	FalseN...	Recall	Precision	Sensitivity	Specificity	F-meas...	Accuracy	Cohen'...
Citation analysis	6	15	126	4	0.6	0.286	0.6	0.894	0.387	?	?
Communication	1	0	149	1	0.5	1	0.5	1	0.667	?	?
Data	3	2	137	9	0.25	0.6	0.25	0.986	0.353	?	?
E-resources	1	1	149	0	1	0.5	1	0.993	0.667	?	?
ICT	1	0	146	4	0.2	1	0.2	1	0.333	?	?
Information	1	0	138	12	0.077	1	0.077	1	0.143	?	?
Knowledge	2	0	147	2	0.5	1	0.5	1	0.667	?	?
Library	3	0	142	6	0.333	1	0.333	1	0.5	?	?
Library Science	1	0	146	4	0.2	1	0.2	1	0.333	?	?
Metric Study	10	35	80	26	0.278	0.222	0.278	0.696	0.247	?	?
Movement	2	0	146	3	0.4	1	0.4	1	0.571	?	?
Social Network	1	0	148	2	0.333	1	0.333	1	0.5	?	?
Trend Analysis	1	1	143	6	0.143	0.5	0.143	0.993	0.222	?	?
Academic Libr...	0	0	144	7	0	?	0	1	?	?	?
Administration	0	0	150	1	0	?	0	1	?	?	?
Authorship	0	0	150	1	0	?	0	1	?	?	?
Biography	0	0	147	4	0	?	0	1	?	?	?
Publication	0	0	150	1	0	?	0	1	?	?	?
Computer	0	0	149	2	0	?	0	1	?	?	?
Content Anal...	0	0	147	4	0	?	0	1	?	?	?
Education	0	0	149	2	0	?	0	1	?	?	?
Ethics	0	0	146	5	0	?	0	1	?	?	?
Law	0	0	150	1	0	?	0	1	?	?	?
Preservation	0	0	150	1	0	?	0	1	?	?	?
Professional	0	0	150	1	0	?	0	1	?	?	?
Quality	0	0	150	1	0	?	0	1	?	?	?

Figure 8. BERT prediction.

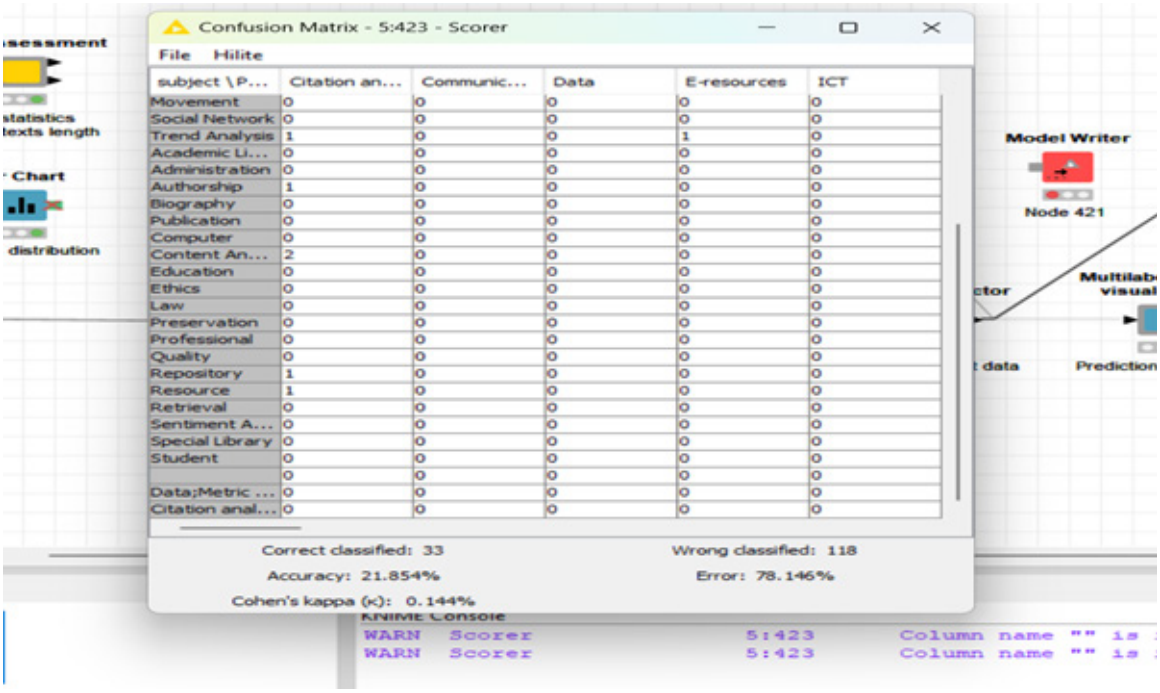


Figure 9. BERT predictor confusion matrix.

Figure 8 portrays the statistical measure of the model representing the recall value, precision, F-measure, sensitivity, and specificity. As only 13 were represented in the model content thus, the rest showed 0 recall value and could not draw the precision value.

Our small corpus sample resulted in one-fourth accuracy in the prediction node. The reason for such a result is the issue with BERT, as it is trained with the common vocabulary of the English language and Wikipedia articles. UDC is a controlled vocabulary, the terminologies fed into BERT were normalized using the 'Subject' Qualifier, which was too small for a pre-trained domain-specific model. However, we were surprised to observe that even a small set of additional terminologies led to a semi-perfect predictive model with high hope of better prediction with a bit extra effort in data feeding.

5. Conclusion

As stated earlier, the hierarchical framework of UDC has always been considered the prime advantage of a polylith-controlled vocabulary. Despite its advantages, UDC still seems complex for notation building because of symbols, signs, punctuation, and other provisions. The classification model we developed is only partially successful because of the non-availability of an authorized UDC scheme. Many a time, we found the results are distracting as the contextual meaning of the content is not provided; however, if trained specifically by annotating each concept and if the full licensed version of UDC class numbers become available (Borovic *et al.*, 2022) there is a greater possibility of developing an automated freely faceted classification scheme for practical use.

6. References

Alrowili, S. E. and Vijay-Shanker, K. (2021). BioM-Transformers: Building large biomedical language models with, 221-227. https://doi.org/10.18653/v1/2021.bionlp-1.24

Arora, S., May, A., Zhang, J. and Ré, C. (2020). Contextual embeddings: When are they worth it? 2650-2663. https://doi.org/10.18653/v1/2020.acl-main.236

Berthold, M. R., Cebron, N., Dill, F., Fatta, G. D., Gabriel, T. R., Georg, F., Meinl, T., Ohl, P., Sieb, C. and Wiswedel, B. (2009). Knime: The Konstanz Information Miner. ACM SIGKDD Explorations Newsletter, *11*(1). https://doi.org/10.1145/1656274.1656280 PMCid:PMC2670301

Borovic, M., Ojstersek, M. and Strnad, D. (2022). A hybrid approach to recommending universal decimal classification codes for cataloguing in Slovenian

digital libraries. IEEE Access. https://doi.org/10.1109/ACCESS.2022.3198706

British Standards, I. (2005). UDC, Universal Decimal Classification. British Standards Institution.

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding.

González-Carvajal, S. and Garrido-Merchán, E. C. (2020). Comparing BERT against traditional machine learning text classification.

Guo, Z. (2021) When can BERT beat SVM? Replication of four text classification papers and fine-tuning RobBERT on Dutch language datasets.;11p. retrieved from: https://theses.liacs.nl/pdf/2020-2021-GuoZhenyu.pdf

Gweon, H. and Schonlau, M. (2022). Automated classification for open-ended questions with BERT.

Joorabchi, A. and Mahdi, A. E. (2011). An Unsupervised Approach to Automatic Classification of Scientific Literature Utilising Bibliographic Metadata. Journal of Information Science, *37*(5), 499-514 https://doi.org/10.1177/0165551511417785

Koroteev, M. V. (2021). BERT: A review of applications in natural language processing and understanding. 188.

KNIME Community Hub. Available from https://hub.knime.com/

Kragelj, M. and Kljajić Borštnar, M. (2021). Automatic classification of older electronic texts into the Universal Decimal Classification-UDC. Journal of Documentation. *77*, 755-776 https://doi.org/10.1108/JD-06-2020-0092

Narkhede, S. (2018). Understanding AUC - ROC Curve. Available from https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

Nevzorova, O. (2021). Towards a recommender system for the choice of UDC code for mathematical articles. 190.

Schrumpf, J., Weber, F. and Thelen, T. (2021). A neural natural language processing system for educational resource knowledge domain classification. 194, 283-283.

Slavic, A. (2008). Use of the universal decimal classification: A world-wide survey. Journal of Documentation. *64*, 211-228 https://doi.org/10.1108/00220410810858029

Slavic, A., Siebes, R. and Scharnhorst, A. (2021). Chapter 5. Publishing a knowledge organization system as linked data. The case of the universal decimal classification. In: Linking Knowledge, Ergon - ein Verlag in der Nomos Verlagsgesellschaft, 69-98. https://doi.org/10.5771/9783956506611-69

UDC Consortium. (2022). About Universal Decimal Classification (UDC). Available from: https://udcc.org/index.php/site/page?view=about