Automatic classification of older electronic texts into the Universal Decimal Classification–UDC

Matjaž Kragelj
Information Technology Office, National and University Library, Ljubljana, Slovenia, and
Mirjana Kljajić Borštnar
Department of Information Systems, Faculty of Organisational Sciences, University of Maribor, Kranj, Slovenia

Received 9 June 2020
Revised 29 October 2020
Accepted 30 October 2020

Abstract

Purpose – The purpose of this study is to develop a model for automated classification of old digitised texts to the Universal Decimal Classification (UDC), using machine-learning methods.

Design/methodology/approach – The general research approach is inherent to design science research, in which the problem of UDC assignment of the old, digitised texts is addressed by developing a machine-learning classification model. A corpus of 70,000 scholarly texts, fully bibliographically processed by librarians, was used to train and test the model, which was used for classification of old texts on a corpus of 200,000 items. Human experts evaluated the performance of the model.

Findings – Results suggest that machine-learning models can correctly assign the UDC at some level for almost any scholarly text. Furthermore, the model can be recommended for the UDC assignment of older texts. Ten librarians corroborated this on 150 randomly selected texts.

Research limitations/implications – The main limitations of this study were unavailability of labelled older texts and the limited availability of librarians.

Practical implications – The classification model can provide a recommendation to the librarians during their classification work; furthermore, it can be implemented as an add-on to full-text search in the library databases.

Social implications – The proposed methodology supports librarians by recommending UDC classifiers, thus saving time in their daily work. By automatically classifying older texts, digital libraries can provide a better user experience by enabling structured searches. These contribute to making knowledge more widely available and useable.

Originality/value – These findings contribute to the field of automated classification of bibliographical information with the usage of full texts, especially in cases in which the texts are old, unstructured and in which archaic language and vocabulary are used.

Keywords Digital library, Artificial intelligence, Machine learning, Text classification, Older texts, Universal Decimal Classification

Paper type Research paper

We are grateful to the librarians from the National and University Library for contributing their expert knowledge in classification of older texts to make this research happen. We are also grateful to Jon Škoberne for his help on machine learning process.

Mirjana Kljajić Borštnar was supported by the Slovenian Research Agency, ARRS, through research programmes P5-0018.

Journal of Documentation
Vol. 77 No. 3, 2021
pp. 755-776
Emerald Publishing Limited
0022-0418
DOI 10.1108/JD-06-2020-0092

1. Introduction

Written sources are a cornerstone of cultural heritage and provide evidence of human creativity, development and culture in specific times and spaces. They are kept and cared for at libraries.

With the digitisation of written sources, books, periodicals, serials and other kinds of written sources have become more easily accessible to scientists and the public. National governments have invested much effort in the digitisation of complete libraries. In Europe, the digital library project Europeana (www.europeana.eu) works with thousands of European archives, libraries and museums to share cultural heritage for enjoyment, education and research. Europeana Collections provides access to over 57m digitised items (books, music, artworks and more), which means that users can access all the available knowledge online through digital libraries by full-text or categorised search. While newer texts and articles are typically equipped with metadata, such as subjects, keywords and the Universal Decimal Classification (UDC), older texts are not. The amount of archive texts and articles available through digital libraries is enormous, and it cannot be expected that librarians could perform the task of UDC classification on their own. It is estimated that several hundred thousand texts, published in the 19th and 20th centuries, will not be manually processed, nor will librarians produce bibliographic records in the library catalogue for those. Because these types of sources will probably not be catalogued (in contrast to scholarly articles, for which a large set of metadata is available), it will be difficult or even impossible to offer filters and faceted navigation (because the lack of available metadata, including classification, such as UDC).

The problem addressed in this paper is the assistance in bibliographic text processing of older digitised texts, which remain in the hands of human experts. These are currently mostly classified only to the general UDC group through the classification of the entire journal. The thesis of this paper is that machine-learning (ML) methods make it possible to automatically assign or propose texts to the appropriate UDC group or several of them. Based on this thesis, we have developed two research questions:

RQ1. Can the UDC classification of the new scholarly texts, assigned by human experts, be used to build the UDC classification model?

and

RQ2. Can a classification model, built on scholarly texts, be used to classify older (unstructured) texts?

For this purpose, we have developed an ML classification model on the newer, correctly classified texts to use it for the classification of the old texts. This research aims to develop a methodology for automatic classification of electronic articles or texts into UDC. The structure of the paper is as follows: first, we start with the problem description and related work. We proceed with the methodology of the research, and the methods used for data collection and analysis. In the results section, we present the classification model, its validation and some use cases. Finally, we conclude with a discussion of the results and their implication for theory and practice.

1.1 Problem description

Like other digital libraries (e.g. Europeana, Open Library (www.openlibrary.org), Library of Congress (www.loc.gov/collections/) and others), there are vast numbers of free digital resources in the web. In the Republic of Slovenia, one of the richest and most complete free electronic sources is the Digital Library of Slovenia (www.dlib.si) containing more than 850,000 electronic (digitised and digital) publications. Digitised sources are those that were

originally only printed and later transformed into a digital format, whereas digital sources are those that are originally created and made accessible in digital format (and may also be printed). In the following text, we will use the term "digital publications" for all written sources available through the Digital Library of Slovenia. Usually, items included in a library would have a bibliographical record. Scholarly publications are systematically bibliographically processed, which means they have a bibliographical record in the library catalogue, and therefore one or more classification numbers from the UDC system. In contrast, most older texts and sources have not been bibliographically processed (e.g. articles and texts from older printed magazines and newspapers from the field of culture) and therefore are not classified into UDC system.

On the UDC Consortium's website, the classification is described as one of the first universal classification systems and remains one of the most widely used international classification systems in librarianship. It was developed at the end of the 19th century by the Universal Bibliographic Repertory, based on the Dewey Decimal Classification (DDC; 1876), by Melvin Dewey (Dale, 1978; Kendall, 2014). Other well-known systems are also the Library of Congress Classification (LCC), Colon Classification (CC) and Bibliographic Classification (BC) (Miksa, 2017). The UDC classification is based primarily on the numerical ("decimal") labelling of the contents of articles with line item sequencers in the form of a sequence of numbers and symbols; it is universally portable, highly structured, the notation is precisely representative of the subject content and understandable between languages. It is built so that it can be expanded and upgraded with new classifications. The classification is used worldwide and is currently translated into more than 50 languages. According to Slavic (2008), it is the second most used classification in the world. It allows unlimited assembly of classification attributes (e.g. master table, place, time, etc.) and relations between them to describe the subject (in our case) of the publication. Because publications in the union library catalogue (the COBIB union bibliographic/catalogue database) in Slovenia use the UDC classification scheme, we used this one. It contains nine basic groups:

(1) Science and knowledge. Organisation. Computer science. Information. Documentation. Librarianship. Institutions. Publications

(2) Philosophy. Psychology

(3) Religion. Theology

(4) Social sciences

(5) Mathematics. Natural sciences

(6) Applied sciences. Medicine. Technology

(7) The arts. Recreation. Entertainment. Sport

(8) Language. Linguistics. Literature

(9) Geography. Biography. History

Group 4 is empty. It originally consisted of linguistics, which was later transferred to Group 8.

An example of the classification of a record in a library catalogue can be made for Ivan Cankar's drama, "Hlapci". In the COBIB (union library catalogue), the following UDC numbers are identified for it:

821.163.6-2 (Slovenian literature, dramatics) 821.163.6.09 (Slovenian literature, literary publication)

For newspapers, they are usually classified as 070 –"Newspapers. Printing. Journalism" or 050 –"Serials. Periodicals".

On the website of the Digital Library of Slovenia, it is possible to search the content of the articles only through the full text. It is currently the best tool for discovering older texts. However, using and researching articles and other publications in such a way only does not offer good user experience, due to optical recognition deficiencies (poor quality of text recognition in newspapers and serials of the older type, use of old Slovene script like "metelčica", "dajnčica", "bohoričica", "gajica", etc.) and too many returned search results. For the majority of the texts and copies of serials, there is only one bibliographic record in the library catalogue. Examples of this include "The Laibacher Zeitung", a newspaper, with more than 58,000 issues and many more articles, Ljubljanski zvon-"The Bell of Ljubljana", with more than 11,000 articles, or Dom in svet-"The Home and World" with over 16,000 articles, etc. The easiest way to illustrate the present situation is the following example: all the articles of the serial "The Home and World" that originated between 1888 and 1944 are placed in the UDC classification sub-group Slovene Literature and Culture (821,163.6 and 008) which means that every article within the magazine "Home and World" is classified as "Slovene Literature" and "Culture" and nothing more than that. If we mention some well-known magazines from those times and describe their content, these would be: "Home and World"–it was a Slovenian literary monthly, which was created as an entertaining and educational magazine for Roman Catholic readers and later developed into a literary magazine. It was founded by the philosopher and theologian Frančišek Lampe and edited until his death (1900). The magazine was initially distinctly Catholic but represented a more tolerant and, above all, the most artistically creative part of Catholic culture. Another one is "The Bell of Ljubljana"–the central Slovenian literary newspaper was published monthly. In addition to literature, it also contained art criticism as well as discussions and essays on the arts. At first, it was more scientifically oriented but later limited to the humanities, and from 1931 also with articles on current social issues. We can also mention "Agricultural and handicraft news"; it was first intended to help farmers and artisans, but later also carried articles in the fields of literature, conservative politics, culture and correspondence from various places. They were important mainly for the consolidation of the Slovene literary language, the general acceptance of the "gajica" and, in general, for the all-round cultural development of the Slovene nation. Gajica is a Latin script developed by the Croatian linguist Ljudevit Gaj. It was first used to write Croatian, but later, with some adaptations, it was also used to write Slovene.

2. Related work

Much research has been done on the classification of data in various fields. Data are everywhere, and its quantity is growing rapidly. Notably, the rise of data created on social media and the growth of online business transactions, which are expected to grow to 450bn transactions a day by 2020 (Khatri, 2016), is contributing to the expansion of the digital universe.

Text mining, as a subfield of data mining, has become increasingly important in recent years due to the wide range of resources that generate huge amounts of data such as social networks, blogs/forums, websites, emails and online libraries that publish research articles (Altınel and Ganiz, 2018). The main goal is to process and use the raw information in texts using ML algorithms (Bhushan and Danti, 2018). One of the problems encountered when analysing texts is that a text is usually in free form, unstructured but machine-learning algorithms usually need structured input. Text mining thus refers to extracting interesting patterns, clusters and high-quality information from large text corpora using ML and statistical learning (Kaushik, 2013). Text classification using machine-learning techniques, as an important tool for managing vast amounts of texts, is described in by Ikonomakis *et al.* (2005).

Text mining is gaining much attention in different scholarly fields, for example, analysing user evaluation of tourism services, especially hotel and tourism services (Jimenez-Marquez

et al., 2019). Similarly, in the field of medicine, with more than 27m articles currently in PubMed database, it is increasingly difficult for researchers and healthcare professionals to efficiently search extract and synthesise knowledge from a variety of publications. Yi (2005) addresses the classification of bibliographic data from the library catalogue in MARC format and doctoral dissertation abstracts. The study seeks to achieve two goals, the use of the Markov Hidden Model to categorise the text, and the use of the Washington Library of Congress classification in conjunction with the Markov Model and data mining. For the set of publications, author used extracts from ProQuest's dissertation database. These are already categorised by librarians and offer the ideal test set to test the model.

Similarly to our proposed research, their goal was to classify the texts, but their corpus consisted of fully described texts in the library catalogue, although in another type of classification (Library of Congress Classification (LCC) -https://www.loc.gov/catdir/cpso/lcco/. There are, of course, other ways to classify texts, publications and articles. Erbs *et al.* (2013) present a hybrid approach to index term assignment, with a combination of key phrase extraction and multi-label classification, which is an extension of the automatic tagging of documents with the use of multi-label classification, which assigns labels to documents, which can be clustered by their labels and which are similar to tags or categories. Another study presents an approach to the assignment of the Library of Congress (LOC) subject headings with the usage of bibliographic records, DDC and abstracts of publications (Wartena and Franke-maier, 2018).

The interest in the topic of scholarly text classification and recommendation has grown in recent years. Regarding the classification of scholarly texts according to the UDC (Romanov *et al.*, 2016), texts are classified by peers based on their keywords. Similarly, bibliographic metadata (title, description and subject tags) can be used to equip texts with DDC to supplement bibliographic records of publications (Khoo *et al.*, 2015). Also, for e-news, support can be developed for automatic inclusion into pre-defined groups, such as art, function, news, reviews, sports, world (Asy'arie and Pribadi, 2009; Ramdass and Seshasai, 2009). One of the research studies on extracting the meaning of vocabulary from sentences, not only by categorising words and finding semantic connections is the one taken by Karras and Mertzios (2002), using DDC.

The spread of digital resources and their integration into the traditional library environment has created the need for an automated tool that organises publications into library classification schemes. Yi (2007) asserts that the search for automatic text classification is a research area for the development of tools, methods and models for use and operation in this field that author describes the currently popular approach for text sorting and lists some projects in the area of classification: sorting publications in libraries, most notably the LCC and DDC. A view on the other aspect, considering the fast development of digital repositories and growth of data and information, the utilisation of semi-automatic metadata generation may be unavoidable in the future (Park and Brenza, 2015). Some techniques include meta-tag and content extraction, automatic indexing, text and data mining, extrinsic data auto-generation, social tagging, among others.

A survey of methods, such as content-based, collaborative filtering, graph-based and hybrid methods can be found in the work of Bai *et al.* (2019). They identified main approaches to recommender systems and commonly used performance evaluation metrics (Precision, Recall, F-measure, NDCG, MAP, MRR, MAE, and UCOV). They outline the open questions that are relevant to all kinds of recommender systems (cold start, sparsity, scalability, privacy, serendipity) and unified scholarly data standards. Analysis of the use of recommendation-as-a-service for academia is presented in the study by Beel *et al.* (2017). Porcel *et al.* (2009) propose a model of a fuzzy linguistic recommender system to help the University Digital Library users accessing research content and collaborate. All the presented recommender systems address predominantly the user side, where our proposed model is primarily aimed to help librarians and consequently improve user experience.

There are several reasons for our choosing the UDC system as a target classification system in our research. The primary reason is that in the COBIB database (https://plus.cobiss.si/opac7/help/cobib), the results of shared cataloguing of more than 680 Slovenian libraries participating in the COBISS (Co-operative Online Bibliographic System and Services). In the SI system, all the newer material is already tagged with this classification system (UDC). Therefore, it is pragmatic to use the same classification system for equipping the older material as well. Second, UDC covers all knowledge sciences (Salah *et al.*, 2012) and offers the following contributions as suggested in the study by Colillas (2011):

(1) Classification codes (the identifier, number) can be used as a key to overcoming the problems experienced due to multi-lingualism (for example, the UDC number 61 has the same meaning, namely "medicine" in all languages).

(2) Harmonisation can help organise the entire architecture from a semantic perspective. Objects (in our case, records, articles, texts, as well as images, maps, etc.) are grouped by subjects, not by alphabetical relationships.

(3) Each new classification serves (like the others) as a code list and can be reused.

(4) It allows composite keys as a kind of descriptor set (e.g., 821'18/19' Literatures of individual languages and language families, 18th–19th century)

(5) It is scalable.

3. Methodology

The approach used in this research falls under the Design Science Research (DSR) (Hevner *et al.*, 2004), in which an IT artefact (ML classification model) is being developed to solve a real-life problem (classification of old texts). The process of model development both builds on prior knowledge and contributes to new knowledge. Efficiency, quality and usefulness must be demonstrated through evaluation, and detailed and verifiable results are provided, considering clear and rigorous methods (Kuechler and Vaishnavi, 2008).

The identification of the problem, which represents the first activity in the DSR methodology, and the objectives were defined in the introductory chapters, in which we described the problem and objectives of the research. Based on the research question stated in the introduction section, we developed the following hypothesis:

H1. It is possible to build a classification model on the corpus of scholarly texts that would assign any randomly chosen publication from the test set (of scholarly texts) to at least one appropriate UDC number with the probability of at least 0.8.

We will test Hypothesis 1 using the performance measures commonly used in text classification: classification accuracy (CA), Recall, Precision and F1.

H2. The classification model built on a corpus of scholarly articles can assign at least 80% of publications to one suitable UDC number for each popular, unstructured, old text.

We will test Hypothesis 2 with human experts (librarians) evaluating the performance of the classification model on 150 randomly selected classified texts.

3.1 Data preparation

At the core of the design cycle, as described by Hevner *et al.* (2004), is the classification model building process. The research process of data collection and data analysis methods, which are presented in Figure 1, are described in detail as follows.

We exported the attributes described below for each text into a .json file to be further processed from the MSSQL database in which we store the data. In total, we exported more than 70,000 scholarly articles and more than 200,000 old texts. In order to uniquely distinguish content and input attributes for model construction, the database included the "Title", "Full-text of article", "Identifier" and "UDC numbers". For the old texts, the UDC numbers were those defined for whole magazines, journals or newspapers, as already stated.

This phase can also be called the "pre-processing" of data. We cleaned out the words that were not useful and only introduced stop words into the model. It is a process of cleaning the text (deleting non-alphanumeric characters, blank lines, etc.).

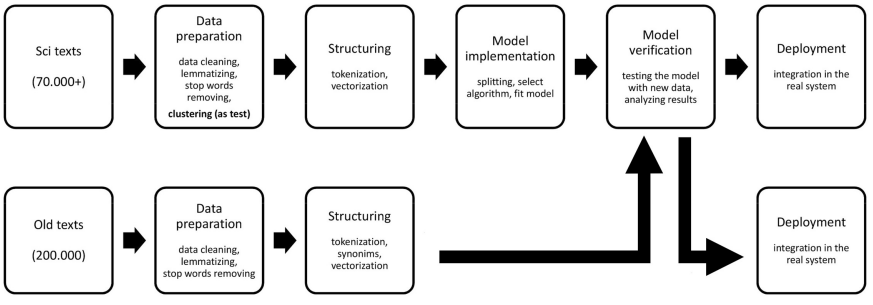
The process of text classification model building starts with preparing a corpus of texts. In our case, we have two corpora: the first corpus of more than 70,000 scholarly texts that are all bibliographically processed, meaning they are assigned with UDC number or numbers by a human expert. The second corpus consists of 200,000 old and bibliographically not processed texts (articles, short notices, popular texts) that were mostly published between 1850 and 2000 in journals and newspapers in the Slovene language used at that time. Slovene is morphologically one of the more difficult/rich languages and, in its history, it has had several variations (i.e. bohoričica, meteljčica, dajnčica) which makes it difficult to compare with language in scholarly articles written in the present day. Consequently, we had to pay extra attention when cleaning the texts, using "lemmatisation"–acquiring the base of the words and replace old vocabulary with that used today. We used two approaches in the cleaning process, as well as one already built model–FastText.

(1) *Minimal processing* – MP (we only retained words that were alphabetic (containing only alphabetical characters) and are lemmatised using a file that contains the root form (lemma) and all the words that share this root. If the word does not exist in the root dictionary, we left it in its original form.)

(2) *Regular processing*–RP (similar as above, plus we also removed "stop words" and words that we do not want to include in the list of words; mostly irrelevant words). We also removed all words smaller than three characters and those whose root or lemma was not in our dictionary.

(3) *FastText*–FT. The FastText model can be seen as a shallow neural network that derives its capabilities by scaling up the number of learnable vector embeddings of n-gram features that are fed into the network (Agibetov *et al.*, 2018). Its database contains data from Facebook and is translated into more than 150 languages.

The ML phase can be done with different programming languages, such as Python, R and similar. Like Jimenez-Marquez *et al.* (2019), we used the NTKL (Natural Language Toolkit)



Source(s): Own source

Figure 1. The process of automatic text classification

with Python, which offers easy-to-use interfaces and language resources, such as WordNet, along with a collection of word-processing libraries for sorting, tokenisation, perception, marking, parsing, semantic reasoning and similar (Bird *et al.*, 2009).

3.1.1 Process of clustering (unsupervised learning). First, we conducted a clustering analysis (Figure 1, before structuring phase), using a k-means algorithm (Colavizza and Franceschet, 2016; Jain, 2010), to test whether UDC classification of the scholarly texts adequately represents the natural groups identified within the texts. Since the unsupervised learning works on unlabelled and uncategorised data, we could take this approach before structuring the data. The goal was to find useful insights from the data. For clustering, we used the complete data set of more than 70,000 scholarly texts, which is presented in detail in the following text.

Clustering is a type of unsupervised ML, in which an algorithm seeks for similarities in a data set without the supervisor assigning or disregarding labels. As can be seen in the work of Romanov *et al.* (2016), the distribution of academic articles in UDC classifications in the UDC top-level library catalogue (0–9) may vary. Some areas of the UDC are better represented in the articles than others, there is a greater number of articles in a category and some categories are less represented. In order to avoid complications in model building and testing for imbalance, we ensured equal representation, equal distribution of texts across all UDC groups at the basic level. We randomly picked 900 scholarly texts, specifically 100 texts for each main UDC class.

We used this set to test for clusters distribution. We set the parameter k (as the desired number of clusters) to 73. We set the parameter k to this number because the sum of different UDC numbers in the set of 900 articles when we shortened the UDC to the second digit (821.16-Literature in Slavic languages become 82-Literature) was equal to 73.

Unsupervised learning identifies common characteristics of data and connects members with the same characteristics to groups, thus creating clusters. As part of the research, we used this type of ML to verify the grouping of input data into sets. We checked if the algorithms grouped related articles into the same groups. The results are presented in Section 4 and Figure 2.

3.2 Structuring

First, the vocabulary had to be "separated" into tokens or words. This process is called "tokenisation", which is a step in a process which that longer strings of text into smaller

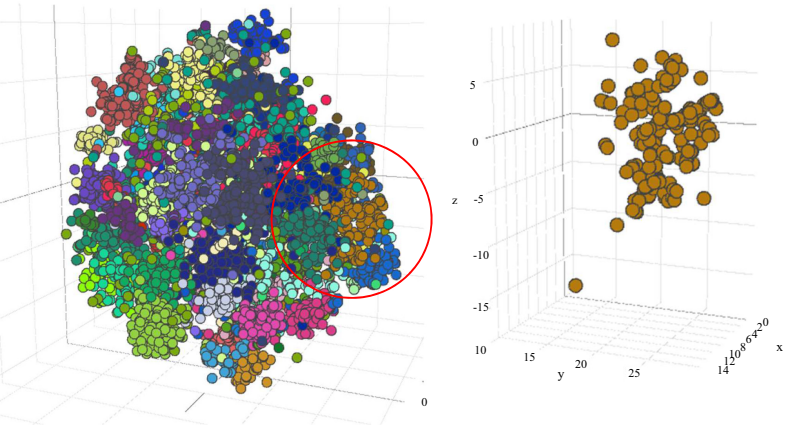


Figure 2.
3D view on 73 clusters (left) and zoom-in on one cluster, UDC 27–Christianity (right)

Source(s): Own source

pieces, called "tokens". Larger pieces of text can be tokenised into sentences and sentences into words. Tokenisation is enabled in Python programming language using the NLTK library and the word_tokenize function (Jimenez-Marquez *et al.*, 2019; Ramdass and Seshasai, 2009). After that, we apply lemmatisation to obtain the dictionary form of words. For the old text corpus, we implemented another step: synonyms replacement. Since the old corpora also use anachronisms, we had to "translate" them into their current form. We used a dictionary with old words and lemmas to convert old words into their current forms. To make tokens useful for the classification, we had to convert them into numbers, which is the process of vectorisation. Similar to the work of Farkas *et al.* (2010) and Zhang *et al.* (2016), we used the Tf-idf method to model the matrix of word vectors appearing in texts.

When building feature vectors from texts, we did not use term frequency but inverse document frequency. According to Aggarwal and Zhai (2012), in general, a common representation used for text processing is the vector-space based TF-IDF (term frequency-inverse document frequency) representation. In the TF-IDF representation, the term frequency for each word is normalised by the IDF. The IDF normalisation reduces the weight of terms, which occur more frequently in the collection, which reduces the importance of common terms in the collection, ensuring that the matching of texts is more influenced by that of more discriminative words that have relatively low frequencies in the collection. The result is sometimes called "the semantic vector" (Jalil *et al.*, 2016). The result of vectorisation is the m*x*n array, which stores a dictionary of words appearing in texts. After the vectorisation step, the vocabulary (an array with vectors representing the articles in n-dimensional space) was ready for the model implementation.

3.3 Model implementation

In this step, we had to perform three key tasks: splitting the data set to train and test set, selecting an algorithm and fitting the model. For this purpose, we split the data set, consisting of more than 70,000 scholarly texts, to train and test subsets in the ratio of 80/20 (57,039 instances used for training the classifier, and 14,299 instances used for testing the classifier). After the step of dividing the data set into learning and testing, we used algorithms to build ML models. Each algorithm built its model on the training data set, which was later tested with test data.

The data set of old texts consisted of more than 200,000 articles. Validation of the performance of the classification model (Figure 1, fifth step) was done by human experts: 15 librarians who assessed the automatic UDC classifier on 150 randomly selected texts, since they were never classified by librarians.

In the research, among the algorithms of supervised learning of which are many (e.g. Bayes classifier, Decision trees, Support vector method, Neural networks, Genetic algorithms, Latent semantic analysis, k-nearest neighbours, etc.) (Abdul *et al.*, 2015; Baharudin *et al.*, 2010; Bhalla and Kumar, 2016; Chowdhury *et al.*, 2019; Du, 2017), we chose Linear Classifiers (*Logistic Regression* (LR) (Healthy and Survey, 2014; Ng and Jordan, 2001), *Naive Bayes Classifier* (NB) (Asy'arie and Pribadi, 2009; Kononenko, 1993)), *Support Vector Machines*, (SVM), (Bhalla and Kumar, 2016; Cortes and Vapnik, 1995), Kernel Estimation (*K-Nearest Neighbours*, (k-NN) (Baharudin *et al.*, 2010; Rawat and Choubey, 2016)), Neural Network (*Multilayer Perceptron* (MLP) (Collobert and Bengio, 2004; Na *et al.*, 2019)).

ML algorithms are described as learning a target function (*f*) that best maps an input variable (*X*) to an output variable *(Y): Y = f (X)*. The main goal of ML is to learn the mapping *Y = f (X)* of *Y* prediction for a new input *X*. This can be called predictive modelling, for which the goal is to obtain the most accurate prediction possible. Fitting the model is making the algorithm learn the relationship between predictors and outcome so that prediction for the future values of the outcome is possible.

To assess the performance of the trained text classification model, for which the target variable can have two or more classes, the measures of Precision, Recall, and F1 score are used. The following Equations (1)–(4) are required to determine the metric values of the confusion matrix (Abdelaziz *et al.*, 2018; Joo *et al.*, 2013).

3.4 Model verification

Normally supervised learning techniques are used for automatic text classification, in which pre-defined category labels are assigned to articles based on the likelihood suggested by a training set of labelled documents (Baharudin *et al.*, 2010). In supervised ML, our task was divided into two phases. Learning and testing 70,000 scholarly articles (UDC was available in the bibliographic catalogue) and use the trained model to classify 200,000 old non-scholarly articles. In the first phase, we tested 20% of the scholarly texts (test set), which amounted to 14,299 articles. We used classification accuracy (CA), Recall, Precision and F1 measures. The following Equations (1)–(4) are required to determine the metric values (Abdelaziz *et al.*, 2018; Joo *et al.*, 2013).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Precission = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1\,score = \frac{2TP}{(2TP + FP + FN)} \tag{4}$$

Where

(1) Accuracy–the ratio of correctly classified cases to all cases of the observed set.

(2) Precision–the ratio of correctly classified positive cases to all positively classified cases of the observed set.

(3) Recall–the ratio of correctly classified positive cases to all positive cases of the observed set.

(4) F1 score–harmonic mean of precision and recall.

4. Results and discussion

In the following, we report the results of the clustering analysis, classification model building and testing on newer scholarly data and use on the older texts.

4.1 Clustering of the scholarly articles data set

In the bibliographic catalogue, all the articles are classified with the usage of UDC; therefore, it was straightforward to check whether the naturally occurring clusters in the scholarly corpus are aligned with the assigned UDC class. Figure 2 shows articles in 73 different clusters (articles represented by dots, colours representing clusters) (abbreviated to the first two digits of UDC, there were 73 different UDC in 900 articles). Articles with similar content tend to be closer to one another and thus form clusters, which are graphically represented by dots in 3D space.

Zoomed-in detail on the right of Figure 2 shows the cluster (or a group), one of the 73 groups. This is an example of a "clean" group, since elements tend to stick together. In this group, the articles contain content about Christianity (UDC = 27). Most significant (TF-IDF) words from articles in this group are "Church, man, life, God, faith, council, God's, question, etc". More than 90% of all articles/elements in this group have UDC = 27 (in the bibliographic catalogue). Another example of a very clean group is the one in which the articles have UDC 82 (literature). Most calculated significant words in this group are: "song, time, sky, eyes, heart, Earth, sun, wind, water, children, love, night, etc.".

We used the clustering method to explore the input data and assess whether we can use this data to conduct the classification training. In so doing, we checked the relationship between publications and their UDC rows on the one hand and clusters on the other on the operation of the k-means algorithm. As can be seen in the example on the right side of Figure 2, clusters were homogeneous. Also, it is true that each publication can have more than one UDC number and should therefore appear in multiple clusters (there is high relation between religion and architecture, so the article should be in both clusters, but it is only in one). The disadvantage of using only unsupervised learning methods is that the system operates only with unmarked data during learning and has no insight into the correctness of the results; its objective is to identify hidden structures in unlabelled data (Vakharia *et al.*, 2015). The use of unsupervised learning methods on a small set of articles (900) served in a research analysis.

4.2 Classification model building for scholarly texts

The complete corpus of 70,000 and more scholarly articles was divided into training and testing sets, 80% instances used for training (57,039 texts) and the remaining 20% instances used for testing (14,299 texts). Hypothesis 1 states that it is possible to build such a classification model on the corpus of scholarly texts, which would assign any randomly chosen publication from the test set to at least one appropriate UDC group with the probability of at least 0.8; it was tested using CA, Recall, Precision and F1 measures. The results of classification algorithms performance on the test set of scholarly texts are presented in Table 1.

Tf-idf MP	SVM	MLP	LogReg	NB
Statistic				
Accuracy	0.963	0.944	0.940	0.873
Precision	0.894	0.842	0.739	0.721
Recall	0.824	0.783	0.630	0.802
F1	0.849	0.804	0.656	0.742
Tf-idf RP	SVM	MLP	LogReg	NB
Statistic				
Accuracy	0.959	0.942	0.940	0.884
Precision	0.883	0.833	0.783	0.693
Recall	0.811	0.788	0.658	0.781
F1	0.837	0.802	0.691	0.720
FASTTEXT	SVM	MLP	LogReg	NB
Statistic				
Accuracy	0.943	0.931	0.840	0.753
Precision	0.853	0.841	0.517	0.562
Recall	0.794	0.751	0.400	0.710
F1	0.814	0.782	0.420	0.590

Table 1. Results for the test data set (14,299 articles) for the second level of UDC using minimal processing, regular processing and FastText

The results obtained for the classification of scholarly texts corroborate Hypothesis 1. As evident from Table 1, at least 80% of articles are accurately classified into an appropriate UDC group. The best performing classifier, according to the classification algorithm, is SVM using Tf-idf (CA = 0.963). When the number of features (i.e. individual measurable characteristics of a subject being observed) is low, SVM, LogReg and MLP algorithms perform better than NB or KNN (Colas and Brazdil, 2006; Musa, 2013; Zanaty, 2012).

4.3 Using the classification model to classify old texts

Next, we tested Hypothesis 2, stating that the classification model trained on the corpus of scholarly texts can be used to classify the corpus of old texts. The old texts are not fully bibliographically processed (texts are merely assigned the UDC of the parent's bibliographic record–magazine or newspaper), often poorly structured, written in archaic Slovene language and vary in length compared to the scholarly texts. The vocabulary of the language in the 19th and early 20th centuries, compared to the present-day language is quite different because the vocabulary itself and language of each nation are changing over time. The complexity of the sentences and the choice of words are also different than in the academic literature. By reviewing the article, the librarian can give an opinion on the correctness of the classification given by a classification model.

Therefore, the main challenges of this task were the fact the language that had changed significantly within one century and the fact that the principle of writing scholarly texts is considerably different from the writing of popular texts. For this reason, we used a dictionary and translated old Slovene words (where possible) into the language used today, in order to make the classification algorithms work better. If we had a large corpus of bibliographically processed old texts, we would have used it for a learning set and probably would have achieved even better results.

Once we had confirmed Hypothesis 1, we used the complete corpus of scholarly texts (more than 70,000) as a training set, following the basic idea that the classification model's performance is better with a larger set. We then used this classification model to classify the 200,000 old texts. The trained classification models served us determining UDC for old texts.

The algorithms have placed the articles in one or more different UDCs. For further analysis, we considered only UDC numbers with at least 10% probability to fit in some UDC group by the classifier and sorted them by probability, for example, "KNN: [("821", 0.59), ("7", 0.19), ("929", 0.12), ("111", 0.1)], where the first number in the bracket is the UDC group, and the second is the probability of correct placement in a group, in this example calculated by KNN algorithm.

A total of 150 randomly selected articles, automatically assigned with UDC classes, were evaluated by 15 librarians (each librarian had the task of evaluating 10 randomly assigned texts). The text and the results of the automatic classification according to three different text processing and vectorisation approaches (minimal processing, regular processing and FastText) were available for each of five classifiers (Naive Bayes classifier, Support vector machines, Multilayer perceptron, Logistic regression and k-nearest neighbours algorithm).

For better understanding the work of the human experts, we present a few examples below. Librarians labelled calculated UDC numbers with green if *the proposed UDC by the classifier is appropriate* and with yellow if *the proposed UDC by the classifier is appropriate in a broader context.* Proposed UDC numbers that were not labelled were not appropriate for the article that was processed. An important difference between UDC numbers assigned by librarians (for the entire journal) and the articles reviewed is that the classifiers mostly "found" and "suggested" UDC numbers that described the article by *content* and not only type (e.g. 070 - Newspapers. The Press. Journalism, 050 - Serial publications, periodicals (as subject)). We explained this in more details by giving three examples.

Examples

Example 1: "Electric Robot" https://www.dlib.si/details/URN:NBN:SI:DOC-Y1E3CHY4 from the newspaper Slovenski Gospodar, volume 72, issue 40, year 1938. It was published mostly weekly in Maribor.

Electric robot (article, translation from Slovenian language): "*In Czechoslovakia, a mechanical device was introduced, which automatically regulates the lighting and switch-off of the light. The robot consists of two cells. The first one is in the building of the city power plant, and the other on the transformer, as soon as it is in the evening, it reacts both cells to the change of light by burning everywhere electric light.*"

Since the article is very short, it seems important not to discard too many words from the corpus. In Tables 2 and 3, we display the results of automatic classifiers where minimal and regular processing was done. In Table 4, we display the results of an article that was processed with FastText.

UDC numbers for the entire publication (from the record in the library catalogue) are:

(1) 070 Magazines. Print. Journalism

(2) (497.4) Slovenia. Republic of Slovenia

(3) "18/19" 19/20. century

Calculated/suggested UDK numbers accepted by the librarians:

(1) 007 Activity and organisation. Communication and control theory in general (cybernetics). "Human Engineering"

Table 2.
Results of classifiers performance for article "Electric Robot". UDCs accepted by the librarian are labelled green and accepted in a broader context are labelled yellow. (minimal processing)

Title	UDC for newspaper	Classifier	UDC1	UDC2	UDC3	UDC4
E. robot	070	SVM	621	53	620	616
E. robot	070	LOG	621	821	616	
E. robot	070	MLP	34	621	620	930
E. robot	070	NB	621			
E. robot	070	K-NN	620	53	544	620

Table 3.
Results of classifiers performance for article "Electric Robot". UDCs accepted by the librarian are labelled green and accepted in a broader context are labelled yellow. (regular processing)

Title	UDC for newspaper	Classifier	UDC1	UDC2	UDC3	UDC4
E. robot	070	SVM	007	621	681	821
E. robot	070	LOG	621	821	007	616
E. robot	070	MLP	621	007	620	
E. robot	070	NB	007			
E. robot	070	K-NN	007	621	681	

Table 4.
Results of classifiers performance for article "Electric Robot". UDCs accepted by the librarian are labelled green and accepted in a broader context are labelled yellow. (FastText)

Title	UDC for newspaper	Classifier	UDC1	UDC2	UDC3	UDC4
E. robot	070	SVM	621	53	537	620
E. robot	070	LOG	621	53	669	638
E. robot	070	MLP	628	621	544	57
E. robot	070	NB	53	537		
E. robot	070	K-NN	620	621	331	579

(2) 53 Physics

(3) 620 Material Investigation. Blessing. Power plants. Energy

(4) 621 Mechanical Engineering General. Nuclear technology. Electrical engineering. Mechanical technology in general

(5) 681 Precision mechanisms and instruments

Interestingly, using FastText, we also obtained highly ranked UDK numbers:

(1) 53 Physics

(2) 537 Electricity. Magnetism. Electromagnetism

(3) 628 Health Tech. Water. Sanitary appliances. Lighting technique

According to librarians, physics (53) is too a broad term, but UDC 537 has been confirmed.

Not appropriate calculated UDC numbers were

(1) 331 Statistics as a science. Statistical theory ($p = 0.19$)

(2) 34 Law. Jurisprudence ($p = 0.96$)

(3) 544 Physical chemistry ($p = 0.19$)

(4) 57 Biological sciences in general ($p = 0.26$)

(5) 579 Microbiology ($p = 0.11$)

(6) 616 Pathology. Clinical medicine ($p_{max} = 0.02$)

(7) 628 Public health engineering. Water. Sanitation. Illuminating engineering ($p = 0.78$)

(8) 638 Keeping, breeding and management of insects and other arthropods ($p = 0.03$)

(9) 669 Metallurgy ($p = 0.06$)

(10) 821 Literatures of individual languages and language families ($p_{max} = 0.18$)

(11) 930 Science of history. Historiography ($p = 0.01$)

Except in some cases, the algorithms have labelled mostly irregular UDC classes as suitable with low probability. If set put a condition (e.g. the minimum probability of placement should be at least $p = 0.2$), most of them would drop out considerably, and librarians would have less work to do with validating appropriate rows. In contrast, UDC 621 was suggested with the following probabilities: [0.15, 0.23, 0.29, 0.40, 0.46, 0.52, 0.57, 0.59, 0.71, 0.90, 0.99, 0.99] which gives us an average of $p = 0.57$ and seems more credible than probabilities below $p = 0.2$.

The classifiers perform surprisingly well in old Slovenian texts. An example from "Kmetijske in rokodelske novice" shows us so. Agricultural and handicraft news was published weekly by Janez Bleiweis. They were first intended to help farmers and craftsmen, and later published articles in the fields of fiction, conservative politics, culture and letters from various places. They were important mainly because of the consolidation of the Slovene literary language, the general acceptance of the "gajica" language and, in general, the comprehensive cultural development of the Slovenian nation. The newspaper continues under the headline "Novice kmetijskih, rokodelnih in narodskih reči".

For example, we took a very short text whose contents are shown in Figure 3.

Example 2: Krajnski Vertnar, https://www.dlib.si/details/URN:NBN:SI:DOC-TYG137JU, from the Journal "Kmetijske in rokodelske novice", Volume 1, Number 1, Year 1843.

Forecast of agricultural books (article/notice, translation from old Slovenian language):

"Forecast of agricultural books, For sale in Ljubljana at the bookstore Mr Lerchar on the big square: Krajnski Vrtnar, or teaching to grow many fruit trees in a short time, to ennoble them by grafting, and plant beautiful gardens with great benefit. The Imperial Royal Society of Agriculture in Carniola was brought to light. Written by Franz Pirc, the pastor at Sv. Jernej in Loče. In Ljubljana 1834–1835. Price 24 crowns."

Tables 5–7 show the calculations of the classification results for the text "Krajnski Vertnar"

Napoved kmetijſkih bukev (knig).

Na prodaj v Ljubljani per bukvarju Gofpodu Lercharju na velikim tergu:

Krajnſki Vertnar, ali Poduzhenje v kratkim veliko ſadnih dreveſ sarediti, jih s zeplenjam poshlahtniti, in lepe verte k velikim pridu safaditi. Na ſvetlobo dala zefarſka kraljeva drushba kmetijſtva na Krajnſkim. Spifal Franz Pirz, fajmofhter per ſ. Jèrneji v Pezbah. V Ljubljani 1834 — 1835. Svesan sa 24 kr.

Source(s): www.dlib.si

Figure 3.
Kranjski Vertnar, KiRN, 05.07.1843

Table 5.
Results of classifiers performance for the article "Krajnski vertnar". UDCs accepted by the librarian are labelled green and accepted in a broader context are labelled yellow. (minimal processing)

Title	UDC for newspaper	Classifier	UDC1	UDC2	UDC3	UDC4
Kranjski Vertnar	070	SVM	930	908	929	94
Kranjski Vertnar	070	LOG	930	821	908	929
Kranjski Vertnar	070	MLP	908	930	398	929
Kranjski Vertnar	070	NB	94	930		
Kranjski Vertnar	070	K-NN	655	34	347	061

Table 6.
Results of classifiers performance for the article "Krajnski vertnar". UDCs accepted by the librarian are labelled green and accepted in a broader context are labelled yellow. (regular processing)

Title	UDC for newspaper	Classifier	UDC1	UDC2	UDC3	UDC4
Kranjski Vertnar	070	SVM	930	908	929	821
Kranjski Vertnar	070	LOG	821	930	929	908
Kranjski Vertnar	070	MLP	930	908	347	94
Kranjski Vertnar	070	NB	316	347	908	
Kranjski Vertnar	070	K-NN	655	34	347	061

Table 7.
Results of classifiers performance for the article "Krajnski vertnar". UDCs accepted by the librarian are labelled green and accepted in a broader context are labelled yellow. (FastText)

Title	UDC for newspaper	Classifier	UDC1	UDC2	UDC3	UDC4
Kranjski Vertnar	070	SVM	930	821	34	908
Kranjski Vertnar	070	LOG	930	821	908	811
Kranjski Vertnar	070	MLP	34	811	617	784
Kranjski Vertnar	070	NB	930			
Kranjski Vertnar	070	K-NN	930	811	908	94

UDC numbers for the entire publication (from the record in the library catalogue) are:

(1) 070.48 Specific types of newspapers

(2) (497.4) Slovenia. Republic of Slovenia

(3) "18" 19th century

Calculated/suggested UDK numbers accepted by the librarians:

(1) 655 Graphic industry. Printing. Publishing. Bookbinding

(2) 821 Literature of individual languages and language families

(3) 908 Examining Areas. Exploring Places [Home Science]

(4) 39 Cultural Anthropology. Ethnography. Customs. Habits. Tradition. Lifestyle

(5) 398 Folklore in the narrow sense

We were interested in the work of classifiers on a short, unrelated, unstructured text. The following is an example from Kmetijske in rokodelske novice (Agricultural and Craft News), the content is shown in Figure 4, and the results of classifiers in Table 8.

(1) 633 Field crops and their production

(2) 339 Trade. Commerce. International economic relations. World economy

Even if the text was very short and unstructured, we still obtained two good UDC numbers, which are a description of the content and not only (as it is in the catalogue for the whole newspaper–070: Newspapers. The Press. Journalism) by type of publication.

The examples shown here indicate the power of automatic, (actually semi-automatic since validation work is still required) of the classification. It is interesting that even though in some cases these are relatively short texts, we received validated UDC numbers suggestions from different main UDC groups: 0, 3, 6, 8.

In Figure 5, we report the analysis of human experts (librarians) of the evaluation of the 150 randomly selected texts. As seen, every bar displays an article tested by librarians.

Žitna cena

v Ljubljani 17. januarija 1880.

Hektoliter: pšenice domače 10 gold. 8 kr. — banaška 12 gold. 3 kr. — turšice 6 gold. 60 kr. — soršice 7 gold. 99 kr. — rží 6 gold. 39 kr. — ječmena 4 gold. 39 kr. — prosa 4 gold. 20 kr. — ajde 5 gold. 20 kr. — ovsa 3 gold. 31 kr. — Krompir 3 gold. 5 kr. 100 kilogramov.

Figure 4.
Very short text passage "Cereal price"

Source(s): dLib.si

The green bar above the identifier of an article presents the count of appropriate UDC number or numbers, confirmed by the human expert (librarian). The yellow bar presents an appropriate UDC number in a broader context, as stated by librarians. As we can see in Figure 5, most of the articles were approved with at least one UDC number by a librarian. We should take into consideration that it is not necessary that every librarian would label or choose the same UDC numbers for the same articles. Even in real life, in the cataloguing process, classification by librarians varies. In the research by Marijan and Leskovar (2015), they showed that the work that includes human decision-making could vary from decision-maker to decision-maker. Librarians are independent in their work, in determining the UDC numbers, in the process of cataloguing. For this reason, they reviewed and assigned UDC numbers for different sets of texts in the present study.

The calculation of the average appropriate or appropriate UDC numbers in a broader context for 150 articles is:

(1) On average 1.8 approved UDC numbers per text were confirmed by librarians

(2) On average 2.55 approved or approved in broader context UDC numbers per text were approved by librarians

Title	UDC for newspaper	Classifier	UDC1	UDC2	UDC3	UDC4
Cereal price	070	SVM	633	930	39	908
Cereal price	070	LOG	908	636	930	39
Cereal price	070	MLP	663	339	633	39
Cereal price	070	NB	656			
Cereal price	070	K-NN	625	656	633	51

Notes(s): "Cereal price", **(** http://www.dlib.si/details/URN:NBN:SI:DOC-5M8MOW7Q)
Cereal price (article/notice, translation from old Slovenian language)
"Cereal price
In Ljubljana on January 17, 1880
Hectoliter: domestic wheat 10 goldinars 8 crowns-banaška (wheat from the province of Banat) 12 goldinars 3 crowns-pickles (white, very old variety of corn) 6 goldinars 60 crowns-soršice (mixture of wheat and rye) 7 goldinars 99 crowns-rye 6 goldinars 39 crowns-barley 4 goldinars 39 crowns-millet 4 goldinars 20 crowns-buckwheat 5 goldinars 20 crowns-oats 3 goldinars 31 crowns-potatoes 3 goldinars 5 crowns for 100 kilograms"

Table 8.
"Kmetijske in rokodelske novice", Volume 38, Number 3, Year 1880. Regular processing

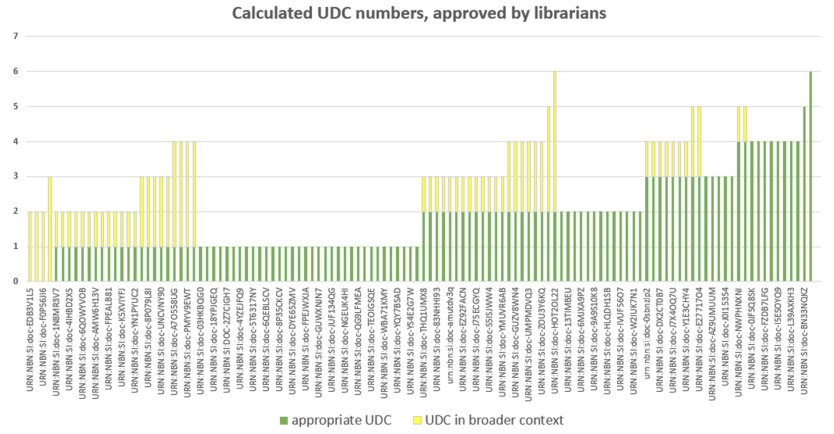


Figure 5.
A view of approved calculated UDC numbers

(3) For four texts, only UDC numbers approved in broader contexts were accepted by librarians

(4) The minimum value for approved UDC per article was 0 (in four cases of articles/texts)

(5) The maximum value for approved UDC per article was 6 (in one case of article/text)

5. Conclusion

We addressed the problem of automatic classification of old texts into UDC classes using a classification model trained on the corpus of newer scholarly texts. For this purpose, we prepared two corpora: the first of the newer scholarly texts and the second of older texts. The newer texts were fully bibliographically processed, well-structured (in an academic manner of writing) and with use of current language. We analysed this corpus using clustering k-means method to confirm the alignment of naturally occurring groups with the UDC groups assigned by human experts: librarians. The corpus was then used for classification model building. The built model was used for classification of older texts, for which labels of UDC classes existed only for the parent publication, meaning all the texts were classified into one category regardless of about what the content of the text was. Therefore, our main goal was to assign categories to these texts automatically according to their content.

We set out to test two hypotheses. The first hypothesis, stating that an efficient classifier trained on the newer scholarly texts, can assign a correct UDC class in more than 80% of cases, was supported by the results (classification accuracy for SVM algorithm was 0.9). The second hypothesis could not be tested by using performance measures such as classification accuracy, as the data set was not labelled. We could only rely on the human expert's opinion, which is the main limitation of this study. We randomly selected 150 automatically classified texts and distributed them to 15 librarians (human experts). The task of each librarian was to evaluate the automatic classification for 10 assigned texts. The results, reported in Figure 5 and explained by examples in Tables 2–8, suggest that the classification models can be used for automatic classification of older texts. Among 150 texts selected for evaluation, there were more than 90% correctly classified into at least one UDC number.

Furthermore, all the 150 texts were accurately classified in at least the broader scope of UDC. The set of automatically assigned UDC numbers, confirmed by human experts, were of such quality that the librarian was able to choose a replacement UDC number (instead of the one assigned for the entire journal) for any given article at least in a broader context. Since the available group of articles was not as big as we would like (e.g. 200,000 scholarly articles evenly distributed across all UDC groups), some areas are much more represented than others, especially if we go in-depth with UDC classifiers. This fact is definitely a limitation of this research. Nevertheless, the research shows a model that can be implemented for other areas and related classification schemes, or related approaches only the main table is used, because the learning set from which we built classification models is too small in our corpus.

In practice, this means that classification models can support the librarians in their daily work as a recommendation system for bibliographical processing. With the help of the research findings, it is possible to create assistance to the cataloguing process and offer the librarians UDC numbers they may have overlooked. As stated by other researchers (Beel *et al.*, 2017; Porcel *et al.*, 2009), there are several approaches for (paper) recommendations for users by librarians, but our research can help librarians themselves in the phase of cataloguing. In addition to this, automatic classification contributes to a better experience for the end-user as well. The retrieval of the categorised old texts through digital libraries and web portals will offer new qualities in accessing the content, categorised by topic, category or

subject of those articles equipped with new information. Thus, additional functionalities can be implemented, such as additional filtering (by topic) and consequently with reducing the time required to search for the data.

This research contributes to the body of knowledge on bibliographic recommendation systems, particularly in the field of old digitised texts that are brought to the public through digital libraries. To the best of our knowledge, there has been no such an attempt described in literature yet. By supporting both the end-user and the librarian, in their work with this accumulated human knowledge, we hope that we serve the society as well.

For the future, we plan to validate the classification models on a data set of bibliographically processed old texts. We also believe that equipping of old texts with metadata (such as UDC numbers) can be further enhanced by implementing the wisdom of the crowds. As concluded by Nguyen *et al.* (2018), the goal is to obtain a sufficient amount of quality data. Existing systems for "mass outsourcing", or.crowdsourcing are usually based on one of three social network structures, as reported by Silvertown *et al.* (2015):

(1) The contributions of all participants have the same weight.

(2) A recognised expert connects and verifies data from the contributions of other users.

(3) The structure is based on the fact that no one can be an expert in identifying all taxonomic groups. Each person's contribution has a different weight, depending on the community's ability to contribute and feedback from the community.

It is probably unlikely to bibliographically process a corpus large enough of old texts to be used for automatic classification with quality more than 99% for not-widespread languages. Most likely, artificial intelligence, more precisely ML, will be the main player that will help us to achieve most of the core in this work.

References

Abdelaziz, A., Elhoseny, M., Salama, A.S. and Riad, A.M. (2018), "A machine learning model for improving healthcare services on cloud computing environment", *Measurement: Journal of the International Measurement Confederation*, Vol. 119, January, pp. 117-128. doi: 10.1016/j.measurement.2018.01.022.

Abdul, N., Mp, P., Hamid, A., Aniza, S. and Shukor, A. (2015), "Homogeneous multi-classifier system for moving vehicles noise classification based on multilayer perceptron", Vol. 29, pp. 149-157.

Aggarwal, C.C. and Zhai, C. (2012), "A survey of text clustering algorithms", in *Mining Text Data*, Springer US. doi: 10.1007/978-1-4614-3223-4_4.

Agibetov, A., Blagec, K., Xu, H. and Samwald, M. (2018), "Fast and scalable neural embedding models for biomedical sentence classification", *BMC Bioinformatics*, Vol. 19, p. 541, doi: 10.1186/s12859-018-2496-4.

Altınel, B. and Ganiz, M.C. (2018), "Semantic text classification: a survey of past and recent advances", *Information Processing and Management*, Vol. 54 No. 6, pp. 1129-1153, ISSN 0306-4573, doi: 10.1016/j.ipm.2018.08.001.

Asy'arie, A.D. and Pribadi, A.W. (2009), "Automatic news articles classification in Indonesian language by using Naive Bayes Classifier method", *Proceedings of the 11th International Conference on Information Integration and Web-Based Applications and Services-iiWAS'09*, p. 658.

Baharudin, B., Lee, L.H. and Khan, K. (2010), "A review of machine learning algorithms for text-documents classification", *Journal of Advances in Information Technology*, Vol. 1 No. 1, pp. 4-20.

Bai, X., Wang, M., Lee, I., Yang, Z., Kong, X. and Xia, F. (2019), "Scientific paper Recommendation: a survey", *IEEE Access*, IEEE, Vol. 7, pp. 9324-9339, doi: 10.1109/ACCESS.2018.2890388.

Beel, J., Aizawa, A., Breitinger, C. and Gipp, B. (2017), "Mr. DLib: recommendations-as-a-service (RaaS) for academia", *Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL)*.

Bhalla, V.K. and Kumar, N. (2016), "*New Review of Hypermedia and Multimedia an efficient scheme for automatic web pages categorisation using the support vector machine*", Taylor & Francis, Vol. 4568, May, doi: 10.1080/13614568.2016.1152316.

Bhushan, S.N.B. and Danti, A. (2018), "Classification of compressed and uncompressed text documents", *Future Generation Computer Systems*, Elsevier B.V., Vol. 88, pp. 614-623.

Bird, S., Klein, E. and Loper, E. (2009), *Natural Language Processing with Python*, 1st. ed., O'Reilly Media, Sebastopol, CA, pp. xx+482, paperbound, ISBN 978-0-596-51649-9.

Chowdhury, S.A., Stepanov, E.A., Danieli, M. and Riccardi, G. (2019), "Automatic classification of speech overlaps: feature representation and algorithms", *Computer Speech and Language*, Elsevier, Vol. 55, pp. 145-167.

Colas, F. and Brazdil, P. (2006), "Comparison of SVM and some older classification algorithms in text classification tasks", *IFIP International Federation for Information Processing*, Vol. 217, pp. 169-178.

Colavizza, G. and Franceschet, M. (2016), "Clustering citation histories in the physical review", *Journal of Informetrics*, Elsevier, Vol. 10 No. 4, pp. 1037-1051.

Colillas, M.G. (2011), "Udc on the internet: theory and project in evolution for use of indexing and retrieval systems", *IFLA Journal*, Vol. 37 No. 4, pp. 305-313.

Collobert, R. and Bengio, S. (2004), "Links between perceptrons, MLPs and SVMs", in *Proceedings of the Twenty-first International Conference on Machine Learning (ICML '04)*, Association for Computing Machinery, New York, NY, p. 23, doi: 10.1145/1015330.1015415.

Cortes, C. and Vapnik, V. (1995), "Support-vector networks", Vol. 297, pp. 273-297.

Dale, D.C. (1978), "A nineteenth-century Cameo: Melvil Dewey in 1890", *The Journal of Library History*, University of Texas Press, Vol. 13 No. 1, pp. 48-56, 1974-1987.

Du, J.H. (2017), "Automatic text classification algorithm based on Gauss improved convolutional neural network", *Journal of Computational Science*, Elsevier B.V., Vol. 21, pp. 195-200.

Erbs, N., Gurevych, I. and Rittberger, M. (2013), "Bringing order to digital libraries: from keyphrase extraction to index term assignment", *D-lib Magazine*, Vol. 19 Nos 9–10, p. 2013.

Farkas, R., Berend, G., Hegeds, I., Kárpáti, A. and Krich, B. (2010), "Automatic free-text-tagging of online news archives", *Frontiers in Artificial Intelligence and Applications*, Vol. 215, pp. 529-534.

Healthy, C. and Survey, K. (2014), "Predicting methamphetamine use of homeless youths attending high school: comparison of decision rules and logistic regression classi fi cation algorithms", Vol. 5 No. 2, doi: 10.1086/676830.

Hevner, A.R., March, S.T., Park, J. and Ram, S. (2004), "Design science IN information systems research", *MIS Quarterly*, MIS Quarterly, Vol. 28 No. 1, pp. 75-105.

Ikonomakis, E., Kotsiantis, S. and Tampakas, V. (2005), "Text classification using machine learning techniques", *WSEAS Transactions on Computers*, Vol. 4, pp. 966-974.

Jain, A.K. (2010), "Data clustering: 50 years beyond K-means", *Pattern Recognition Letters*, Elsevier B.V., Vol. 31 No. 8, pp. 651-666.

Jalil, A.M., Hafidi, I., Alami, L. and Khouribga, E. (2016), "Comparative study of clustering algorithms in text mining context", *International Journal of Interactive Multimedia and Artificial Intelligence*, Vol. 3 No. 7, p. 42.

Jimenez-Marquez, J.L., Gonzalez-Carrasco, I., Lopez-Cuadrado, J.L. and Ruiz-Mezcua, B. (2019), "Towards a big data framework for analysing social media content", *International Journal of Information Management*, Elsevier, Vol. 44, May 2018, pp. 1-12.

Joo, R., Bertrand, S., Tam, J. and Fablet, R. (2013), "Hidden Markov models: the best models for forager movements?", *PLoS ONE*, Vol. 8 No. 8, doi: 10.1371/journal.pone.0071246.

Karras, D.A. and Mertzios, B.G. (2002), in McKay, B. and Slaney, J. (Eds), *A Robust Meaning Extraction Methodology Using Supervised Neural Networks BT - AI 2002: Advances in Artificial Intelligence*, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 498-510.

Kaushik, L. (2013), "Text mining-scope and applications", Vol. 5 No. 2, pp. 51-55.

Kendall, J. (2014), "Melvil Dewey, compulsive innovator", *American Libraries*, American Library Association, Vol. 45 Nos 3/4, p. 52.

Khatri, V. (2016), "Managerial work in the realm of the digital universe: the role of the data triad", *Business Horizons*, Vol. 59 No. 6, Kelley School of Business, Indiana University, pp. 673-688.

Khoo, M.J., Ahn, J.W., Binding, C., Jones, H.J., Lin, X., Massam, D. and Tudhope, D. (2015), "Augmenting Dublin core digital library metadata with Dewey decimal classification", *Journal of Documentation*, Vol. 71 No. 5, pp. 976-998.

Kononenko, I. (1993), "Inductive and bayesian learning in medical diagnosis 1 introduction", *Applied Artificial Intelligence*, Vol. 7 No. 4, pp. 317-337.

Kuechler, W. and Vaishnavi, V. (2008), "The emergence of design research in information systems in North America", *Journal of Design Research*, Vol. 7 No. 1, doi: 10.1504/jdr.2008.019897.

Marijan, R. and Leskovar, R. (2015), *A Library's Information Retrieval System (In)effectiveness: Case Study*, Library Hi Tech VO - 33, Emerald Group Publishing, No. 3, p. 369.

Miksa, S.D. (2017), "The relationship between classification research and information retrieval research , 1952 to 1970", *Journal of Documentation*, Vol. 73 No. 6, pp. 1343-1379.

Musa, A.B. (2013), "Comparative study on classification performance between support vector machine and logistic regression", *International Journal of Machine Learning and Cybernetics*, Vol. 4 No. 1, pp. 13-24.

Na, J., Indra, D. and Santony, J. (2019), "An artificial neural network approach for detecting skin cancer", Vol. 17 No. 2, doi: 10.12928/TELKOMNIKA.v17i2.9547.

Ng, A.Y. and Jordan, M.I. (2001), "Regression and naive Bayes", *Advances in Neural Information Processing Systems*, Vol. 14, pp. 841-848.

Nguyen, T.T.N., Le, T.L., Vu, H., Hoang, V.S. and Tran, T.H. (2018), "Crowdsourcing for botanical data collection towards to automatic plant identification: a review", *Computers and Electronics in Agriculture*, Elsevier, Vol. 155 No. October, pp. 412-425.

Park, J.R. and Brenza, A. (2015), "Evaluation of semi-automatic metadata generation tools: a survey of the current state of the art", *Information Technology and Libraries*, Vol. 34 No. 3, pp. 22-42.

Porcel, C., Moreno, J.M. and Herrera-viedma, E. (2009), "Expert Systems with Applications A multi-disciplinar recommender system to advice research resources in University Digital Libraries", *Expert Systems with Applications*, Elsevier, Vol. 36 No. 10, pp. 12520-12528.

Ramdass, D. and Seshasai, S. (2009), *Document Classification for Newspaper Articles*, pp. 1-12.

Rawat, A. and Choubey, A. (2016), "A survey on classification techniques in internet environment", *International Conference on Advanced Computing and Communication Systems (ICACCS)*, Vol. 2 No. 3, pp. 436-443.

Romanov, A.Y., Lomotin, K.E., Kozlova, E.S. and Kolesnichenko, A.L. (2016), "Research of neural networks application efficiency in automatic scientific articles classification according to UDC", *2016 International Siberian Conference on Control and Communications, SIBCON 2016 - Proceedings*, pp. 7-11.

Salah, A.A., Gao, C. and Suchecki, K. (2012), "Need to categorize: A comparative look at the categories of universal decimal classification system and Wikipedia", *Leonardo*, Vol. 45 No. 1, pp. 84-85. available at: http://scholar.google.com/scholar_lookup?hl=en&volume=45&publication_year=2012&pages=84-85&issue=1&author=AA+Salah&author=C+Gao&author=K+Suchecki&title=Need+to+categorize%3A+A+comparative+look

+at+the+categories+of+universal+decimal+classification+system+and+Wikipedia, https://journals.sagepub.com/servlet/linkout?suffix=bibr54-0306312717692172&dbid=16&doi=10.1177%2F0306312717692172&key=10.1162%2FLEON_a_00344, https://journals.sagepub.com/servlet/linkout?suffix=bibr54-0306312717692172&dbid=128&doi=10.1177%2F0306312717692172&key=000299864900022.

Silvertown, J., Harvey, M., Greenwood, R., Dodd, M., Rosewell, J., Rebelo, T., Ansine, J. and Mcconway, K. (2015), "Crowdsourcing the identification of organisms: a case-study of iSpot", *ZooKeys*, Vol. 480, pp. 125-146, doi: 10.3897/zookeys.480.8803.

Slavic, A. (2008), "Use of the universal decimal classification: a worldwide survey", *Journal of Documentation*, Vol. 64 No. 2, pp. 211-228.

Vakharia, V., Gupta, V.K. and Kankar, P.K. (2015), "Ball bearing fault diagnosis using supervised and unsupervised machine learning methods", *The International Journal of Acoustics and Vibration*, Vol. 20 No. 4, doi: 10.20855/ijav.2015.20.4387.

Wartena, C. and Franke-maier, M. (2018), "A hybrid approach to assignment of library of congress subject headings 1 introduction 2 related work", Vol. 4 No. 1, pp. 1-13.

Yi, K. (2005), "Text classification using a hidden Markov model", PhD Dissertation, 1 February.

Yi, K. (2007), "Automated text classification using library classification schemes: trends, issues, and challenges", *International Cataloguing and Bibliographic Control*, Vol. 36 No. 4, p. 78.

Zanaty, E.A. (2012), "Support vector machines (SVMs) versus multilayer perception (MLP) in data classification", *Egyptian Informatics Journal, Ministry of Higher Education and Scientific Research*, Vol. 13 No. 3, pp. 177-183.

Zhang, Y., Zhang, G., Chen, H., Porter, A.L., Zhu, D. and Lu, J. (2016), "Topic analysis and forecasting for science, technology and innovation: methodology with a case study focusing on big data research", *Technological Forecasting and Social Change*, Elsevier, Vol. 105, pp. 179-191.

About the authors

Matjaž Kragelj graduated in University of Ljubljana, Faculty of computer and information science. Post graduate student-Organization and Management of Information Systems at the Faculty of Organizational Sciences, University of Maribor.

Since 2006 employed at the National and University Library in Ljubljana. Position: Head of the Information Technology and Digital Library. Work Areas: Development of the Digital Library of Slovenia, the development and management of a national aggregator for Culture, manage problems of long-term preservation of digital resources, the introduction of the recommended ISO standards for long-term preservation of digital resources, the provision of permanent identifiers, capturing and archiving the web for permanent preservation of the cultural heritage, text mining. Matjaž Kragelj is the corresponding author and can be contacted at: matjaz.kragelj@nuk.uni-lj.si

Mirjana Kljajić Borštnar received her Ph.D. in Management Information Systems from the University of Maribor. She works as an Associate Professor at the Faculty of Organizational Sciences, University of Maribor and is a member of Laboratory for Decision Processes and Knowledge-Based Systems. Her research work covers decision support systems, data mining, multi-criteria decision-making, and organizational learning. She co-authored several scientific articles published in recognized international journals, including Expert Systems with Application, PLOS ONE, Industrial Management and Data Systems.