UNIWA Diavgeia: Automated Subject and Type Categorization on Organizational Records

Ioannis Triantafyllou
Archival, Library & Information Studies – University of West Attica
Athens, Greece
triantafi@uniwa.gr

Vassilis Vallianos
Archival, Library & Information Studies – University of West Attica
Athens, Greece
v.vallianos@gmail.com

Christos Chrysanthopoulos
History & Archaeology, University of Patras
Patras, Greece
cchrysan@eie.gr

Yannis Stoyannidis
Archival, Library & Information Studies – University of West Attica
Athens, Greece
ystoyannidis@uniwa.gr

Markos Dendrinos
Archival, Library & Information Studies – University of West Attica
Athens, Greece
mdendr@uniwa.gr

Themis Panagiotopoulos
Department of Informatics, University of Piraeus
Piraeus, Greece
themisp@unipi.gr

***Abstract*—This paper researches the intentions and the potential benefits associated with the integration of deep and machine learning technologies into archival and records management practices. With the escalating volume and intricacy of digital records, conventional methods of organizing, categorizing, and administering records confront modern-day challenges. Deep learning (DL) technologies offer prospects to revolutionize the maintenance, accessibility, and utilization of records. This research proposes a case study implementation of deep learning methodologies for thematic and type categorization of records within the University of West Attica (UNIWA). Findings highlight the necessity of deepening the standardization of governmental records management processes in the new big data era. By delving into this subject, the paper endeavors to contribute to a deeper comprehension of the transformative potential of deep and machine learning technologies in archives and records management, aiming to guide future practices and decision-making in the field. Additionally, it represents the initial practical segment of an ongoing research endeavor concerning the computational archival science of records at UNIWA.**

Keywords—Deep learning (DL), Machine learning (ML), Subject classification, Computational archival science, University archives, Archives and records management

I. INTRODUCTION

Central to archival theories and practices lies the concept of archival provenance, which renders it a focal point within archival description [1]. According to this theory, the significance of records is heavily shaped by the circumstances of their creation, emphasizing that the organization and description of these materials should closely align with their original context [2]. The classification of records and archives management, serving as a means to identify and arrange records generated or received during business operations, plays a pivotal role in archival description, particularly in today's landscape where the proliferation of born-digital records has necessitated new requirements in recordkeeping and archives management [3]. Computational archival science represents a burgeoning field that intersects the traditional practices of archival science with technological advancements in the realm of computing and data analytics. Within the context of university record management, this emerging discipline holds the potential to revolutionize the way educational institutions handle, process, and preserve their vast repositories of administrative, academic, and historical data [4]. By linking the power of computational methods, like machine-learning (ML) and deep-learning (DL), natural language processing, data mining, universities can modernize records management processes, enhance information retrieval, and facilitate the efficient analysis of their complex archival collections.

In the context of university records management, the integration of computational archival science can offer several tangible benefits. Firstly, it enables the automated classification and categorization of diverse document types, thereby simplifying the intricate task of organizing and indexing voluminous and born-digital records. Through the application of advanced algorithms, computational archival science facilitates the identification of patterns, trends, and correlations within the university's data landscape, enabling administrators and stakeholders to make informed decisions based on comprehensive insights derived from the archival records. This paper aims to delve into the possible advantages, obstacles, and consequences of incorporating machine learning technologies into the archival and records management systems of universities, by a pilot experiment with the university's administrative records uploaded on the Diavgeia portal (https://diavgeia.gov.gr). Our goal is to discover how these technologies can enhance the efficiency, precision, and accessibility of archival practices, ultimately providing benefits to academic and other institutions. The findings of this research hold practical implications for the University of West Attica (UNIWA) and similar institutions and organizations. If the study demonstrates compelling positive effects, machine learning technologies could significantly streamline and automate the subject categorization process, reducing manual effort and enhancing the overall productivity of records management personnel.

II. MATERIALS AND METHODS

A. Diavgeia

In 2010, the Greek government initiated the Diavgeia project, also known as the Transparency Program Initiative,

The publication of this article was funded by UNIWA

with the aim of restoring trust in the democratic system and providing online visibility into government expenditures. Diavgeia has played a crucial role in advancing transparency and accountability within the Greek government. By making decisions of public administration easily accessible, the platform has fostered a culture of openness, empowering citizens to scrutinize government actions and hold public officials responsible. This transparency has contributed to a reduction in corruption, an increase in public trust, and the promotion of a more accountable government [5]. Records published on Diavgeia portal encompass a broad spectrum of subjects and topics, reflecting the diverse activities of public administration. The content can range from legal and regulatory matters to infrastructure projects, public procurement, and personnel issues. Categorizing such diverse content poses challenges, as decisions may involve multiple subjects or fall into ambiguous categories. Through this approach, a portion of the university's administrative function is categorized and accessible. However, the manually assignment of a document's subject may not always align perfectly with its content or scope. Interpreting and assigning relevant subject categories can be subjective, potentially leading to confusion or misclassification [3].

B. Datasets Description

The datasets used in our research were created by the metadata of the record documents (hereafter simply mentioned as document) that UNIWA has uploaded on the portal of "Diavgeia", as all government institutions are obliged to upload their acts and decisions in it. Each document is digitally signed and assigned a unique Internet Uploading Number (IUN) certifying that it has been uploaded at the portal [6]. The initially engaged dataset, hereafter mentioned as the training dataset (TrDS), consisted of 82,561 documents, from 22/03/2018 to 23/06/2023. The variables/fields of the dataset include the document's IUN, issue date, subject title, thematic categories, type, the unit of the university (department or service) that had issued the document and the text that had been extracted from the document's PDF. Supplementary, a second dataset was used as a test dataset (TeDS) for validation purposes, containing 2,890 documents, from 24/06/2023 to 23/08/2023. For our current research goals, we selected the text of the documents, the type and the thematic categories that each document is assigned to. Ongoing research may also consider the unit of the university that had issued the document, and an improved categorization mechanism for upgraded records' purpose identification.

Diavgeia uses 25 thematic categories, borrowed from EuroVoc, the EU's multilingual and multidisciplinary thesaurus (https://eur-lex.europa.eu/browse/eurovoc.html). Only 14 are currently used by UNIWA: science, education and communications, production/technology/research, trade, economics, etc [6]. Documents can be assigned to more than one thematic category, hereafter called thematic assignments. In our datasets documents were assigned two the most categories: 1st is the basic assignment, while 2nd is the complementary thematic category. Hence, the number of documents was 82,561, however assigned categories were 132,134 in the training dataset. Likewise, the number of documents was 2,890 in the testing dataset, while 4,799 were the assigned categories.

Additionally, Diavgeia supports 35 type categories, but only 19 are currently used by UNIWA: regulatory act, projects/supplies/services assignment, budget approval, contract, etc. Each document can be assigned only one type category, hereafter called type assignment. Hence, the number of documents matches exactly type assignments in both training and testing datasets.

C. Text Pre-Processing and Word-Grouping

The preliminary text pre-processing and grouping of words are crucial for enhancing the semantic representation of both the training and testing datasets. Recognizing morphological variations of word features can be achieved through the suggested statistical technique [6][7][8], which proves more cost-effective compared to linguistic tools. This approach reduces the overall complexity of the procedure and makes it less dependent on specific languages. The initial training corpus contains 30,714,149 tokens. After removing tokens with only one symbol and all stop-words the remaining unique words (unique tokens) were 237,039. Unique words can appear more than one time in the corpus, hence the major difference between the two numbers: millions of tokens to thousands of unique words. Text preprocessing after the word-grouping phase concluded with 69,011 dominant words as possible features for text representation as described in the section below.

D. Text Representation

The inspected research schemes for possible text representation resolutions are a combination of the following methods: Bag-of-Word, χ^2, and DevMax [6][7][8][9]. The first two are well established techniques for text representation, while DevMax is a recently proposed technique for selecting more expressive word-features for text representation. Word-vector representations are generated using the dominant words identified in the preceding word-grouping phase. The presence of words is denoted by binary control: 1 if the document contains any word from a particular word-group, and 0 otherwise [6][7][8]. In the Bags-of-Words approach, a metric is needed to prioritize words in the representation process to select the most indicative ones. In our experiments, two primary ranking metrics, DevMax and χ^2, were utilized based on previous experiences [6][7][8]. Additionally, the size of the representation vector is also scrutinized [6][7][8].

E. EML Methods and Development Environments

Specific major ML methods proclaimed to fit better to similar studies are mainly examined: Decision Tree, Random Forest and SVM. The selected methods are yielding promising results and seem to perform better than others (LogisticRegression, kNN, NeuralNetworks, Bayes, etc), since they can handle very large numbers of features more efficiently [7][8][10]. Deep Learning (DL) neural networks are also recognized as suitable methodology for mining very large numbers of features and data, so we further introduced a simple DL topology in our experiments in section F. PyCharm python's environment was deployed for text preprocessing supported by NLTK (Natural Language ToolKit) and PyMuPDF libraries. Classical ML classifiers were incorporated by scikit-learn library, while Keras library was deployed for DL implementation.

F. Deep Learning Architecture

Neural networks (NN), inspired by the structure of neurons in the human brain, consist of a multitude of nodes, also called neurons, which are placed in different layers and are connected to each other. The layers between the input and output layers are called hidden layers, while the total number of layers is called the depth of the network. Dimensions of

the neural network is a crucial feature of performance, especially the number/depth of total layers. Nodes (neurons) found in each layer expresses the dimension (width) of the layer [11]. Another important feature of the structure of a NN is the activation functions. The purpose of these predefined functions is to transform the data they receive as input (from a previous layer) into a form suitable for further processing by the next layer. Sigmoid is a mature activation function used in early NN, still in use especially in binary classification decisions to determine last layer verdicts of class assignments. Sigmoid is defined as: $f(x) = 1/(1+e)^{-x}$. ReLU (Rectified Liner Unit) is also a very popular, non-linear function used in NN, especially in DL topologies, as it is computationally efficient and easy to apply. Its increased performance (efficiency, economy of computational resources) compared to other functions also lies in the fact that ReLU does not activate all neurons simultaneously [12]. It is defined by the formula: $f(x) = \max(0,x)$.

The DL-topology deployed is a simple feedforward NN with 3 hidden layers, with Adam optimizer and binary cross-entropy as loss-function [11]. The first hidden layer has inputX256 neurons, the second has 256X128 neurons and the third 128X64. The ReLU activation function was used on every hidden layer. There was also a dropout layer between the layers, with a 0.5 dropout rate to avoid overfitting [11][12]. The output layer has 14-thematic or 19-type sigmoid decision neurons, one for each class, depending on the classification case as described in the following section.

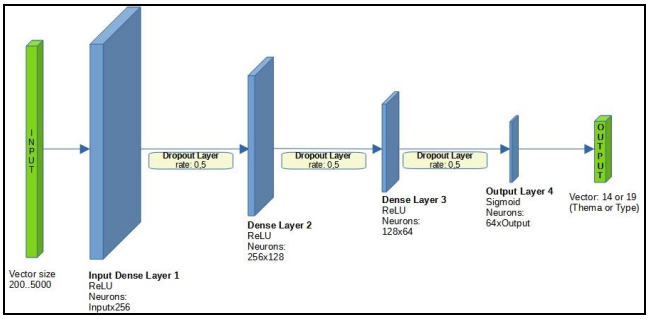

Fig. 1. Deep Learning topology

G. Studied Experiments Architecture

Initially, we conduct data acquisition, text pre-processing, and text tokenization. Subsequently, we employ two main metrics to identify the most significant dominant words and determine the most efficient word-feature selection method. We then conduct experiments using different vector sizes of dominant words, ranging from 200 to 5,000 with increments of 200. Our aim is to determine an optimal vector size for word features that balances representation competency: an insufficient number of features can lead to underfitting of the models, while an excessive number can result in overfitting. The two feature extraction techniques explored are DevMax and χ^2, aiming to ascertain which one provides the most effective representation control. Furthermore, it is essential to identify if DL outperforms conventional ML classification methods. Evaluate phase investigates all methods performance score by the well-established "10-fold cross-validation" technique [6]. 10-folds randomly divides original dataset into 10 parts using stratified sampling. Nine parts are utilized for training and modeling, while one tenth is reserved for testing. This process is repeated 10 times, each time using a different tenth for testing. The overall performance is then calculated as the mean value of the performances of the 10 different models. Validation phase is the next necessary step to ensure that the selected classifiers and their parameters are retaining their robust behavior as new data arrives. The experiments are divided in two classification cases: one for the thematic classification case, hereafter called the thema case, and one for the type classification case, hereafter called the type case.

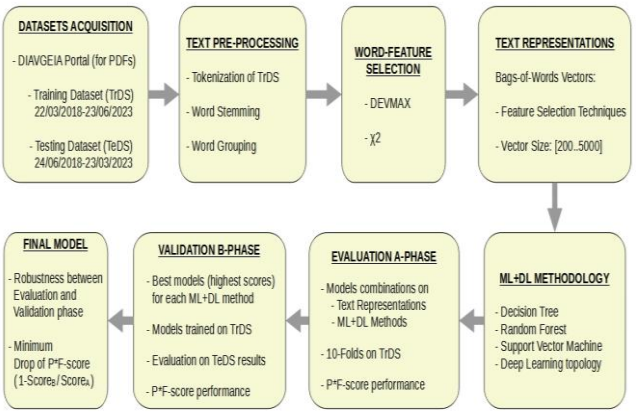

Fig. 2. Proposed modelling and experiments architecture

H. Evaluation Metrics

To evaluate the performance of a classification model, the widely accepted metric utilized is the F1 (or F-measure or F-score), augmented with an additional Precision factor, denoted as P*F1, as Precision holds extra significance for the predictive behavior of the system [6][8]. The F1 represents the harmonic mean of Precision (P) and Recall (R) and is computed as: $F1 = ((P^{-1} + R^{-1}) / 2)^{-1}$. Precision denotes the percentage of accurate positive predictions made by the model relative to all its positive predictions (correct ones or not). Recall represents the percentage of accurate positive predictions relative to the total expected correct predictions (the total population for that category) [13].

III. Results

A. Evaluation (10-folds cross-validation on TrDS)

During this phase (A-phase), we perform 10-folds evaluation using the TrDS. The best results are met as a combination of DevMax and DL classification method. This combination schema produces the highest performance with a 98.89% P*F1 prediction score in the thema case and a 99.08% P*F1 prediction score in the type case. Best results of each classifier are further validated during the next phase to ensure the robust behavior of the concluded schemas.

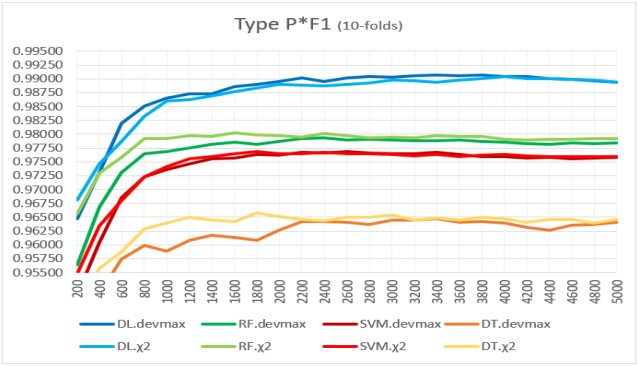

Fig. 3. Thema case: 10-fold performance results

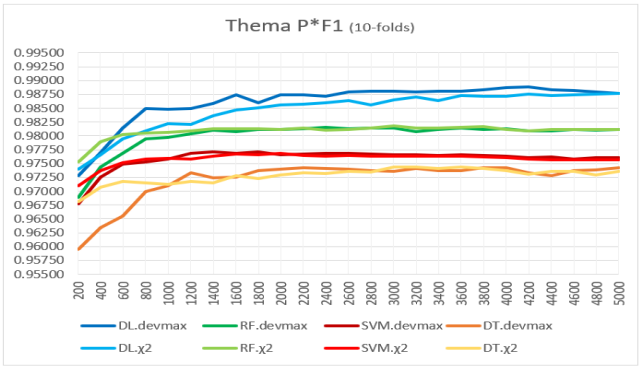

Fig. 4. Type case: 10-fold performance results

B. *Validation (modelling on TrDS & testing on TeDS)*

During this phase (B-phase), best selected classification schemas for each case (thema or type) have to be further tested to validate their efficiency in classifying the TeDS documents. In the following table the overall performances of the best classification schemas for each case are presented. In addition, we use the ScoreDrop metric ($1\text{-}Score_B/Score_A$) to check the drop of performance's predictions between the two phases (A&B), as a metric of schemas' robustness [6].

TABLE I. RESULTS OF A AND B PHASES BASED ON SCOREDROP

Case	Method	Features	A-phase	B-phase		P*F1
			P*F1	P	P*F1	ScoreDrop
thema	**DL**	**4200-DevMax**	**98,89**	**98,86**	**97,52**	**1.38%**
thema	RF	3000-χ^2	98,18	98,21	96,12	2.10%
thema	SVM	1400-DevMax	97,71	98,46	96,09	1.66%
thema	DT	3200-χ^2	97,44	98,02	95,81	1.68%
type	**DL**	**3800-DevMax**	**99,08**	**98,93**	**97,68**	**1.41%**
type	RF	1600-χ^2	98,02	95,13	92,40	5.73%
type	SVM	2600-DevMax	97,69	96,53	93,22	4.58%
type	DT	1800-χ^2	96,58	90,43	84,38	12.63%

IV. CONCLUSION

The primary finding of this research indicates that the most accurate and resilient prediction models result from the combination of DL (outperforming classical ML methods) and the DevMax word-feature selection technique. The predictive efficacy of the recommended models, specifically the DL model with 4200-DevMax word-features in thema case, and 3800-Devmax features in type case, demonstrates remarkably high and robust precisions, scoring at 98.86% and 98,93% respectively with robust (lowest) ScoreDrops. The remarkably high scores emphasize the systems' capacity for highly accurate predictions, implying that automating thematic and type categorization can be confidently pursued with favorable outcomes in academic archival practices.

However, it is important to take a step back and evaluate the different records management systems in the public sector and the needs they address for the operation of the organizations aiming to design a well integration policy for computational archival science at its core. In the context of computational archival science, the records continuum model serves as a guiding principle for understanding the complex interplay between technological advancements and the multifaceted dimensions of records management within academic or other organizational institutions [14]. By acknowledging the fluidity and interrelation of records throughout their lifecycle, computational archival science endeavors to obtain innovative computational tools and methodologies that facilitate the seamless integration, analysis, and preservation of records across various stages of their existence. By leveraging computational approaches such as ML, DL, natural language processing, and data mining, this interdisciplinary field empowers universities to effectively manage the creation, classification, retrieval, and preservation of records in alignment with the principles espoused by the record continuum model. The ongoing and dynamic relationship between archival science and artificial intelligence has the potential to reshape theoretical and methodological frameworks. Organizations can efficiently manage born-digital archives, ensuring their long-term preservation and enhancing their contribution to organizational memory by integrating these technologies within the record continuum model [15]. As the field of archival science evolves, further research and collaboration among archivists, data scientists, and stakeholders will be crucial to fully unlock the potential of ML and DL in transforming university records management practices.

REFERENCES

[1] Sweeney, S. (2008). The ambiguous origins of the archival principle of "provenance". Libraries & the Cultural Record, 43(2), 193-213.

[2] Hensen, S. L. (1993). The first shall be first: APPM and its impact on American archival description. Archivaria, 35, 64-70.

[3] Triantafyllou, I., Chrysanthopoulos, C., Stoyannidis, Y., Tsolakidis, A. (2023). Archives and records management in machine learning technologies context: a research hypothesis on university records. Journal of Integrated Information Management, Vol 8, No1, 7-13.

[4] Cushing, A.L. and Osti, G. (2023), ""So how do we balance all of these needs?": how the concept of AI technology impacts digital archival expertise", Journal of Documentation, 79(7), pp. 12-29.

[5] Karamagioli, E., Staiou, E. R., & Gouscos, D. (2014). Government spending transparency on the internet: an assessment of Greek bottom-up initiatives over the Diavgeia Project. International Journal of Public Administration in the Digital Age (IJPADA), 1(1), 39-55.

[6] Triantafyllou, I. (2023). Thematic Categorization on University Records. IEEE 11th International Conference on Systems and Control (ICSC), Tunisia, 384-389.

[7] Triantafyllou, I., Vorgia, F., & Koulouris, A. (2019). Hypatia Digital Library: A novel text classification approach for small text fragments. Journal of Integrated Information Management, 4, 16-23.

[8] Triantafyllou, I., Drivas, I. C., & Giannakopoulos, G. (2020). How to Utilize my App Reviews? A Novel Topics Extraction Machine Learning Schema for Strategic Business Purposes. Entropy, 22(11), 1310.

[9] Zhang, Y., Jin, R., & Zhou, Z. H. (2010). Understanding bag-of-words model: a statistical framework. International journal of machine learning and cybernetics, 1, 43-52.

[10] Vorgia, F., Triantafyllou, I., & Koulouris, A. (2017). Hypatia Digital Library: A text classification approach based on abstracts. In Strategic Innovative Marketing: 4th IC-SIM, Mykonos, Greece 2015 (pp. 727-733). Springer International Publishing.

[11] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.

[12] Sharma, S., Sharma, S., & Athaiya, A. (2020). Activation functions in neural networks. International Journal of Engineering Applied Sciences and Technology, 04(12), 310–316.

[13] Feurer, M., Klein, A., Eggensperger, K., Springenberg, J., Blum, M., Hutter, F. (2015). Efficient and robust automated machine learning. Advances in neural information processing systems 2015, pp. 2962-2970.

[14] Matlala, M. E., & Maphoto, A. R. (2020). Application of the records life-cycle and records continuum models in organization in the 21st century. ESARBICA Journal, 39(1). Pp/ 79-98.

[15] Colavizza, G., Blanke, T., Jeurgens, C., & Noordegraaf, J. (2021). Archives and AI: An overview of current debates and future perspectives. ACM Journal on Computing and Cultural Heritage (JOCCH), 15(1), 1-15.