https://doi.org/10.2352/issn.2168-3204.2023.20.1.14

AI Powered Tools for Improving Usability in Digital Archiving

Tuomo Räisänen, South-Eastern Finland University of Applied Sciences (Xamk) / Digitalia research center, Mikkeli, Finland; Mikko Lipsanen, Atte Föhr, The National Archives of Finland, Helsinki

Abstract

A software package and PoC to help usability in various cases like quality issues and metadata collecting after digitization is introduced. Here usability means the improvement of the process using AI based automation as much as possible and providing easy-to-use interfaces for the end user. This is done with the help of existing open source tools. Below we present some preliminary results of the ongoing DALAI project[1]*.*

1. *Motivation*

No matter whether we are talking about museums, libraries or archives, they all share several common problems concerning quality assurance and content recognition of the data after digitization. Quality of the digitized data must be ensured. On the other hand, usability of the documents requires collecting as much metadata as possible. Especially in mass digitization projects, manual validation of the quality of the digitized documents is a time consuming and costly process. Therefore, automatic processing of these documents is essential.

Since the early days of digitization of the archived material optical character recognition (OCR) engines have been used to extract text content from digitized documents. The use of AI in document analysis is a relatively new, but rapidly developing research area. Starting from rule based analysis, research is more and more focused on deep learning models. See surveys [7],[11] and references therein.

One method for automatic processing of documents is document classification. In this approach, the input image is classified into predetermined classes, usually with the help of a machine learning model. The recent popularity of deep learning can also be seen in the selection of convolutional networks as the preferred models in many of the applications in the field [7]. In addition, both out-of-domain [8],[9] and in-domain [10] transfer learning has found its way into document image classification. Another interesting method for document classification is using multimodal models [8] and combining both image and text information to improve classification accuracy.

On the other hand, named entity recognition (NER) and automatic subject indexing play an important role in improving usability in digital archiving. Both methods already have an extensive history of use in the field of document analysis [12], [13].

Often the different approaches to document analysis are implemented in separate applications, requiring the users to manually build pipelines according to current needs. Assume, however, that you have a tool to investigate the quality of the digital image, collect the metadata and perform automatic subject indexing in a single interface. Developing such a tool by identifying, adapting and implementing a variety of specialized components is a key goal of the DALAI project, which aims to make digital preservation as simple and straightforward as possible.

One of the project partners is the National Archives of Finland, which aims to improve quality analysis and metadata collection in mass digitization context. Central Archives for Finnish Business Records (ELKA) participates also in the project. The data and feedback provided by the project partners and other stakeholders enables the development of the components and the user interface to be based on a variety of real world cases.

2. *Problems*

What is the quality of the scanned document and how can we improve the usability of the content? On the other hand if the document is digital born, can we utilize the same tools as much as possible?

2.1 Quality analysis

Verifying the condition of digitized material is a truly important factor for developing tools that help to automate various validation procedures.

Document content can be hidden or obscured for instance by folded corners, torn edges and sticky notes, which are not always detected in the digitization phase. Automatic detection of these faults in the digitized documents aims to reduce the need for manual validation as well as to improve the quality of the digitized documents.

Automatic detection of empty document pages helps to speed up the subsequent processing steps, as there is no need to perform fault detection, optical character recognition (OCR), or metadata collection to documents that lack content. It can also improve the experience of the end user of the digitized documents, who can choose to view only images that were not classified as empty.

In all the above cases, the key challenge is to make the automatic detection process accurate and reliable, as well as generalizable to different kinds of digitized documents.

2.2 Collecting metadata

Next phase in improving the usability of digitized documents focuses on the document content. After the text content of the document is extracted (if needed) by an OCR tool, it is passed to named entity

[1] https://kansallisarkisto.fi/en/the-national-archives-2/projects-2/dalai-en

recognition (NER) and automated subject indexing components. These components assemble essential metadata that helps to make the documents more accessible for researchers, public institutions and other potential users by enabling them to find and filter relevant information based on the document contents.

2.3 Ease of use

Improving usability can be a pivotal time factor for digitization procedures. In small archives document by document processing with the help of automation can be a solid choice. In the mass digitization context, large scale processing is needed. Files represent various different formats. How to serve these needs simultaneously as much as possible?

3. *Approaches*

3.1 Quality analysis

3.1.1 Folded corner

When automating image classification problems, typically the best results are achieved with convolutional neural network (CNN) models. Due to specific design of the neural network architecture, CNNs are often able to efficiently recognize patterns and features in the images that are relevant for the classification task at hand [1].

Experiments showed that transfer learning, where a neural network is pre-trained for a classification task with one dataset and subsequently fine-tuned to perform classification for different data, produces superior results compared to training a model from scratch. A Densenet [2] model pre-trained with ImageNet data was used to train the models for both folded corner and post-it marker detection.

The data used in fine-tuning the models consisted mainly of document images from the 1970s onwards, digitized by the National Archives of Finland. The training data containing folded corners and torn edges, in total around 5000 images, were identified by manual annotation. In addition, around 30 000 digitized document images without folded corners were used in the training.

3.1.2 Sticky note

The model and training for sticky note detection followed a similar process. Principal difference relates to the training data, which was partly collected from actual cases of digitized images containing one or several sticky notes, and partly created manually by the annotators. In this case data generation was required since sticky notes occur relatively rarely in the real data. Data generation also made it possible to generate examples that had specific targets, like teaching the model to distinguish sticky notes from other visually similar elements (stamps, colored rectangular text boxes, images etc.), thus helping to improve classification results.

3.1.3 Empty page

The empty page detection model was inherited from a previous project and retrained with additional data. It is also a classification model, but contrary to the previous models it uses a Resnet18 [3] architecture. Our contribution to the development of the model was to retrain it on poorly detected additional data. This data was collected in a test environment in the mass digitization process. Additionally, part of the original training data was also used. The training data consisted of roughly 38500 empty images and 51000 non-empty images.

During training, a TrivialAugment [4] inspired data augmentation was used. Some of the augmentations from TrivialAugment were pruned out, in order to better preserve the original class of the image. The augmentations used were rotating the image, color transformations, sharpness adjusting, blurring, scaling, affine transformation and erasing part of the image.

3.2 Collecting metadata

3.2.1 Automatic subject indexing

Annif software[2] is utilized for automatic subject indexing in the DALAI project. Annif has been mainly developed in the National Library of Finland. This open source software combines several natural language and machine learning algorithms [6]. It is under active development and is already used, for example, by Finnish Broadcasting Company and National Library of Germany. Annif needs a pure text format as input. To achieve this, either the document is OCRed, or if the document already contains data in text format it is extracted with Apache Tika[3] software.

3.2.2 Named entities

Named entity recognition component of the project is developed in cooperation with the FIN-CLARIAH consortium[4]. As a preliminary step, a questionnaire was sent to some of the institutional partners and providers of the digitized documents for the National Archives of Finland as well as to a group of Social Sciences and Humanities researchers, asking them to identify named entities whose automatic recognition would benefit them the most.

Based on the results of the questionnaire, nine entity types (persons, nationalities and religious and political groups, organizations, locations, products, events, dates, Finnish business IDs, journal numbers) were selected. Currently these entities are annotated in the training data that consists of digitized documents from the 1970s onwards.

First version of the named entity recognition component was trained using a dataset containing in total over 500 000 tokens. FinBERT [5], a Finnish version of BERT, is used as the base model that is fine-tuned for the named entity recognition task with the annotated dataset. As the data annotation work proceeds, the model will be retrained with new data to provide a good coverage of all of the included named entity categories.

3.3 Ease of use

[2] https://annif.org/

[3] https://tika.apache.org/

[4] https://www.kielipankki.fi/organization/fin-clariah/

Separate APIs for quality checks and metadata were created from the above mentioned software. For small scale users this makes it possible to develop web UI. On the other hand, large scale users have their own resources to integrate these APIs into their own environment. The scripts were first tested one by one. Using Flask[5], we then collected and modified scripts to two separate APIs. This is relatively simple at least at the development level. At the production level Flask can also be used.

4. *Results*

The data used in testing the components consisted of a wide variety of digitized documents. The documents were mostly typeset, but some also included handwritten text, for example signatures.

4.1 Quality analysis

4.1.1 Folded corner

With a test dataset of close to 10000 images, the model currently correctly classifies 97% of the input images that contain a folded corner, while 98% of the other images are allocated to the right class. Classification accuracy can be further improved by enlarging the training dataset. The current results are already promising when considering the goal of reducing the need for manual validation of the quality of digitized documents.

4.1.2 Sticky note

The results of sticky note detection are very similar to those achieved with folded corner detection. Using a test dataset of equal size, the model correctly detects 98% of the sticky notes in the data, and the images that do not contain sticky notes are classified correctly with similar accuracy. Further improvements in detection accuracy can be achieved for instance by reducing the false classification of rectangular text boxes as post-its.

4.1.3 Empty page

Empty page detection model was tested using an internal test set from National Archives of Finland and an external test set provided by Central Archives for Finnish Business Records. For the internal test set, consisting of roughly 8 000 empty images and 10 500 non-empty images, the model achieved a balanced accuracy of 99.8%. As detecting a non-empty image as an empty image increases the chance of removing documents containing information, the sensitivity of finding non-empty images was prioritized. For the internal test set the sensitivity of finding non-empty images was 99.7%. Due to ELKA not having empty images in their archives, they could only test the model with non-empty images. Their test set consisted of roughly 12 500 images and the model achieved a sensitivity of finding non-empty images of 99.0%.

4.2 Collecting metadata

4.2.1 Automatic subject indexing

In Xamk, Annif is tested for some of our students' thesis. We found that Annif is fast and relatively accurate. It takes only seconds to find subject indexes from a single thesis. Thus it is suitable also for mass digitization contents. From ELKA it has been reported that the choice of the OCR engine can impact the results. Tesseract is used in the project and compared to the previously used OCR engine, which is fixed in the scanning machine.

4.2.2 Named entities

The first version of the model for named entity recognition is included in the web UI. It is able to recognize the following named entities from a text in Finnish: person names, organizations, locations, geopolitical locations, products, events, date, Finnish journal number, Finnish business identity code, and nationality, religious and political groups.

4.3 Ease of use

A web UI is under testing. The feedback is collected from possible end users, like Helsinki City Archives. Our service is called Arkkiivi and it has been tested simultaneously by dozens of end users. UI development has been supported by two organized workgroup meetings. These meetings have included multiple different organizations like museums, digitizing companies and private/government funded archives. Main scope for workgroups has been UI testing and every participant has been given access to a platform for giving feedback to developers.

Also the needs of the mass digitization contents are collected and are under active development.

5. *Conclusions*

The benefit of using open source software, with the collaboration of the different actors and end users, provides a cost efficient way to improve the usability in digital archiving. If the amount of data is 'small' the UI is sufficient. On the other hand the same APIs can be integrated to large scale systems as well, several actors have stated their interest for integrating these components in their own products workflow.

References

[1] Ian Goodfellow et al., Deep Learning. MIT Press. (2016).

[2] Gao Huang et al., Densely connected convolutional networks. Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4700-4708. (2017).

[3] Kaiming He et al. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770-778. (2016).

[4] Müller, Samuel G., and Frank Hutter, Trivialaugment: Tuning-free yet state-of-the-art data augmentation. Proceedings of the IEEE/CVF international conference on computer vision. pp. 774-782. (2021).

[5] Antti Virtanen et al., Multilingual is not enough: BERT for Finnish. arXiv preprint arXiv:1912.07076. (2019).

[6] Suominen, O., Inkinen, J. and Lehtinen, M. Annif and Finto AI: Developing and Implementing Automated Subject Indexing, *JLIS.it*, 13(1), pp. 265–282 (2022). doi: 10.4403/jlis.it-12740.

[7] Lei Cui, et al. Document ai: Benchmarks, models and applications. arXiv preprint arXiv:2111.08609. (2021).

[8] Souhail Bakkali, et al. Visual and textual deep feature fusion for document image classification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. pp. 562-563 (2020).

[9] Adam W Harley, Alex Ufkes, Konstantinos G Derpanis. Evaluation of deep convolutional nets for document image classification and retrieval. In: 2015 13th International Conference on Document Analysis and Recognition (ICDAR). IEEE, pp. 991-995 (2015).

[10] Muhammad Zeshan Afzal, et al. Cutting the error by half: Investigation of very deep cnn and advanced training strategies for

[5] https://flask.palletsprojects.com/en/2.2.x/

document image classification. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). IEEE, pp. 883-888 (2017).

[11] Francesco Lombardi, Simone Marinai. Deep learning for historical document analysis and recognition—A survey. *Journal of Imaging*, (2020), 6.10: 110.

[12] Li, J., Sun, A., Han, J., & Li, C. (2020). A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, *34*(1), 50-70.

[13] Koraljka Golub (2021) Automated Subject Indexing: An Overview, Cataloging & Classification Quarterly, 59:8, 702-719, DOI: 10.1080/01639374.2021.2012311

Author Biography

Tuomo Räisänen has a PhD (2014) from the University of Jyväskylä, Finland. His current interests are in AI, large scale computing and usability, using open source tools.

Mikko Lipsanen has a MSc in Data Science from the University of Helsinki, Finland. His work in the National Archives of Finland focuses on the use of Machine Learning to improve the processing and accessibility of digitized data.

Atte Föhr has a MSc in Bioinformatics from Aalto University, Finland. His work in the National Archives of Finland focuses on machine learning and especially computer vision.