Proceedings of the 4th Financial Narrative Processing Workshop @ LREC 2022, pages 95–99
Marseille, 24 June 2022

swapUNIBA@FinTOC2022: Fine-tuning pre-trained Document Image Analysis model for Title Detection on the Financial Domain

Pierluigi Cassotti*, Cataldo Musto*, Marco de Gemmis*, Georgios Lekkas†, and *Giovanni Semeraro
*University of Bari
Italy
{name}.{surname}@uniba.it

†Objectway SpA
Italy
{name}.{surname}@objectway.com

Abstract

In this paper, we introduce the results of our submitted system to the FinTOC 2022 task. We address the task using a two-stage process: first, we detect titles using Document Image Analysis, then we train a supervised model for the hierarchical level prediction. We perform Document Image Analysis using a pre-trained Faster R-CNN on the PublyaNet dataset. We fine-tuned the model on the FinTOC 2022 training set. We extract orthographic and layout features from detected titles and use them to train a Random Forest model to predict the title level. The proposed system ranked #1 on both Title Detection and the Table of Content extraction tasks for Spanish. The system ranked #3 on both the two subtasks for English and French.

Keywords: keyword1, keyword2, keyword3

1. Introduction

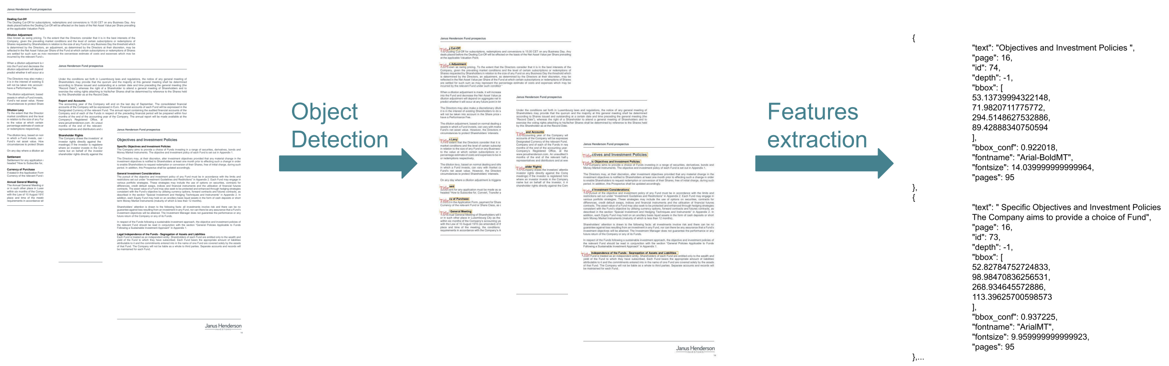
Financial prospectuses contain relevant information about financial funds. These documents are typically released as PDF documents, which can feature very different layouts. Often these documents miss the Table Of Content (TOC) which can help the reader to focus on relevant content. Most of the existing datasets for Table Of Content extraction are domain-specific. The FinTOC task aims to fill the gap, proposing a TOC task specifically for financial documents.

In this work, we address the FinTOC task using a Document Image Analysis approach, exploiting the graphical layout for the Title Detection task. Page Layout Analysis is a long-studied task in the field of Computer Vision. We focus on approaches that exploit Convolutional Neural Networks (CNN) for Object Detection. R-CNN (Girshick et al., 2014) is an object detector that involves three stages: Regions Proposal, Feature Extraction and Classification. The Region Proposal is implemented using the Selective Search (van de Sande et al., 2011) algorithm and aims to find the Regions of Interest (ROI). R-CNN use a Convolutional Neural Network to extract the features from each ROI. The extracted features are used in a Support Vector Machine (SVM) classifier to predict the object class. Fast R-CNN (Girshick, 2015) improves R-CNN, avoiding the feature extraction for each ROI. Fast R-CNN computes a feature map using CNN on the image and extracts ROI from the feature map. Both R-CNN and Fast R-CNN use Selective Search as algorithms for the ROI extraction. The Selective Search algorithm can results inexpensive in terms of computation time. Faster R-CNN (Ren et al., 2015) is a neural network for object detection which jointly train the three object detection stages, implementing the Region Proposal Network (RPN). The RPN is a Convolutional Neural Network that tunes the Region Proposals according to the specific object detection task. While these models offer high performance and efficiency, they require large datasets to be trained. The PubLayNet (Zhong et al., 2019) is an automatically annotated dataset consisting of more than 360,000 pages of scientific articles. Each page is annotated with typical layout elements: title, table, list and text. In particular, it contains more than two million title instances. Since the layout structure of financial documents can diverge in a significant way from those of scientific articles we finetuned a pre-trained model on the PubLayNet dataset for the FinTOC 2022 task. In Section 2 we report the related works. In Section 4 we introduce the proposed TOC extraction pipeline including the Title Detection module and the module for the Level classification. Finally, in Section 5 we report the results on the FinTOC 2022 task.

2. Related Work

(Bourez, 2021) ranked first on the FinTOC 2021 task (Maarouf et al., 2021) on the subtasks of Title Detection and TOC extraction for both English and French. (Bourez, 2021) relies on the commercial software ABBYY[1] for the blocks and tables extraction, then the XGBoost classifier (Chen and Guestrin, 2016) is

[1] `https://www.abbyy.com/`



```
{
        "text": "Objectives and Investment Policies ",
        "page": 16,
        "id": 74,
        "depth": -1,
        "bbox": [
        53.137399943221 48,
        71.9820711775772,
        294.5148627532886,
        89.42888340750594
        ],
        "bbox_conf": 0.922018,
        "fontname": "Arial-BoldMT",
        "fontsize": 14.039999999999964,
        "pages": 95
},
{
        "text": " Specific Objectives and Investment Policies
        The Company aims to provide a choice of Fund",
        "page": 16,
        "id": 73,
        "depth": -1,
        "bbox": [
        52.82784752724833,
        98.98470836256531,
        268.934645572886,
        113.39625700598573
        ],
        "bbox_conf": 0.937225,
        "fontname": "ArialMT",
        "fontsize": 9.959999999999923,
        "pages": 95
},...
```

Figure 1: An overview of the pipeline.

trained on style features such as font color, name, size and weight and the involved text.

(Hercig and Kral, 2020) focuses on the Title Detection task for the FinTOC 2020 task (Bentabet et al., 2020). (Hercig and Kral, 2020) perform an ablation analysis on the features using the leave-one-out cross-validation. From the results emerged that character bigrams, orthographical features and font type represents relevant features. On the contrary, the Title Detection task seems to take no advantages from binary features such as is_bold, is_italic, is_all_caps. A prior work on Title Detection using Document Image Analysis is represented by (Gupta et al., 2021). (Gupta et al., 2021) fine-tune a pre-trained Faster R-CNN on the PubLayNet and filter the detected titles using a Gradient Boosting Classifier. The system proposed by (Gupta et al., 2021) achieves the highest precision with respect all the other systems submitted in the FinTOC 2021 task.

3. Task

The FinTOC 2022 task is the fourth edition of the shared task on Table of Contents extraction from financial documents. FinTOC 2022 extends the FinTOC 2021 task (Maarouf et al., 2021) including spanish documents. In particular, the training data consist of a set of pdf documents for each language, namely English, French and Spanish. For each document, the table of contents is provided. The table of contents includes the text of the title and the related page on which the title appears and the depth of the title. The FinTOC 2020 shared task involves two subtasks. The former is the Title Detection (TD) task, which is a binary task expecting the positive label for text blocks representing a title and a negative label for non-title text blocks. The latter is the Table Of Content (TOC) task, which requires extracting the hierarchical structure of the headers.

4. TOC extraction pipeline

In this section, we introduce the TOC extraction pipeline (Figure 1). It consists of two modules: the Title Detection module and Level classification module. The Title Detection module aims to detect titles in the pdf documents. On the other side, the Level classification module extracts the features from the detected titles and predicts for each title the respective hierarchical level.

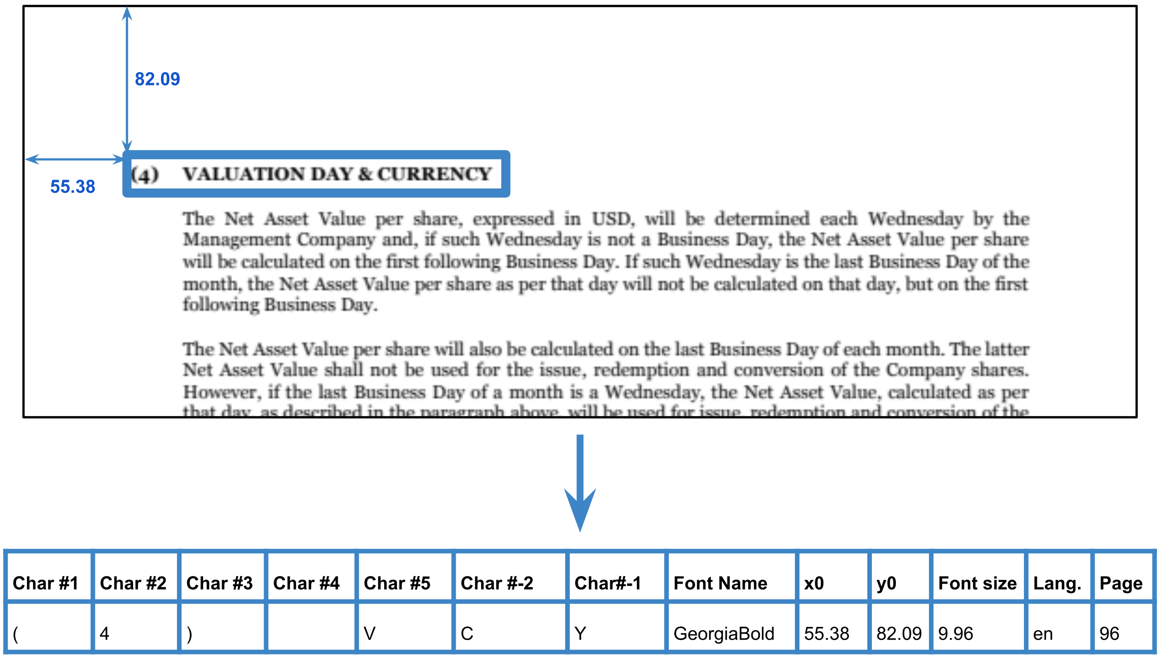
4.1. Title Detection

To model the Title Detection task as a Document Image Analysis task, we extract the bounding boxes associated with each title. We use the Python library pdfplumber [2] for the processing of the pdf documents. We search the text occurrences of titles reported in the training set on the specific page. It is important to state here that the same text of the title can occur multiple times on the same page. Consequently the training set we build can be affected by noise due to title text ambiguities.

Once we find an exact match with the text of the title we extract the related bounding box. For each character belonging to the extracted text pdfplumber provides the char coordinates $(c_{x_0}, c_{y_0}, c_{x_1}, c_{y_1})$ which represent respectively the distance of the left side of character from the left side of the page, the distance of the top of character from the top of the page, the distance of right side of character from the left side of the page, and the distance of the bottom of the character from the top of the page. We extract the bounding box (x_0, y_0, x_1, y_1) coordinates of the overall title text as follows:

- x_0 is computed as the minimum distance from the left page border $\min_{\forall c \in T} c_{x_0}$
- y_0 is computed as the minimum distance from the top page border $\min_{\forall c \in T} c_{y_0}$

[2] https://github.com/jsvine/pdfplumber



(4) VALUATION DAY & CURRENCY

The Net Asset Value per share, expressed in USD, will be determined each Wednesday by the Management Company and, if such Wednesday is not a Business Day, the Net Asset Value per share will be calculated on the first following Business Day. If such Wednesday is the last Business Day of the month, the Net Asset Value per share as per that day will not be calculated on that day, but on the first following Business Day.

The Net Asset Value per share will also be calculated on the last Business Day of each month. The latter Net Asset Value shall not be used for the issue, redemption and conversion of the Company shares. However, if the last Business Day of a month is a Wednesday, the Net Asset Value, calculated as per that day, as described in the paragraph above, will be used for issue, redemption and conversion of the

Char #1	Char #2	Char #3	Char #4	Char #5	Char #-2	Char#-1	Font Name	x0	y0	Font size	Lang.	Page
(	4	)		V	C	Y	GeorgiaBold	55.38	82.09	9.96	en	96

Figure 2: Extracted features.

- x_1 is computed as the maximum distance from the left page border $\max_{\forall c \in T} c_{x_1}$
- y_1 is computed as the maximum distance from the top page border $\max_{\forall c \in T} c_{y_1}$

The extracted bounding boxes are arranged in the COCO format (Lin et al., 2014) for the Object Detection task. Each pdf document d is converted into images $i_1, i_2, ..., i_N$, where N is the number of pages. We train the pretrained model Faster R-CNN included in the Model Zoo [3] of the LayoutParser library (Shen et al., 2021). Specifically, the model uses the Feature Pyramid Networks (FPN) (Lin et al., 2017) as backbone model. We finetuned the model for 80,000 iterations using Detectron [4].

Once the titles of a specific page are extracted, we filter those for which the bounding box has a confidence level greater than 0.5 and we sort them. First, we check for titles that appear in the second column (for double-column documents). A title that has the x_0 coordinate greater than the page width is considered to belong to the second column. Then, we sort titles belonging to the first column in decreasing order sorted by the y_0 coordinate. If there are titles in the second column they are sorted in decreasing order by the y_0 coordinate and appended to the titles in the first column.

Lang.	Precision	Recall	F1
FR	0.728	0.672	0.695
EN	0.802	0.885	0.838
SP	0.462	0.827	0.569

Table 1: Results on TD subtask.

4.2. Level classification

The level classification module attempt to predict the hierarchical level of the title. The hierarchical level of a title is strongly dependent on the overall TOC structure, i.e. the level of a single TOC entry depends on the previous and next titles features. (Bentabet et al., 2019) model the level classification task as a sequence labelling task representing the document hierarchy as a sequence. For simplicity, we propose an element-wise approach that takes into account only the features of a single TOC entry. For the level classification, we train a multi-class Random Forest classifier (Breiman, 2001) that takes in input the features of a single TOC entry extracted by the module of Title Detection and predict the title hierarchical level. The classes correspond to the hierarchical level that goes from 1 to 10 for the FinTOC 2022 task. We use the default hyper-parameters provided by the Scikit-learn library[5], i.e. 100 estimators, and the gini function to measure the quality of the split. The input features (Figure 2) of the Random Forest classifier are:

- First five Characters: one-hot encoding of the first

[3] https://layout-parser.readthedocs.io/en/latest/notes/modelzoo.html

[4] https://github.com/facebookresearch/Detectron

[5] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

Lang.	Inex08-P	Inex08-R	Inex08-F1	Inex08-Title acc.	Inex08-Level acc.	harmonic mean
FR	40.0	37.0	38,3	43.8	30,7	34,08
EN	61.4	66.4	63,6	71.4	42,9	51,24
SP	31.8	59.0	40	65.5	46,5	43,01

Table 2: Results on TOC subtask. Namely Precision, Recall and F1 measure of Inex08 score, Inex08-Title accuracy, Inex08-Level accuracy and the harmonic mean computed over the Inex08 F1 and the Inex08-Level accuracy.

five characters of the text title

- Last two Characters: one-hot encoding of the last two characters of the text title
- Bounding box x_0 normalized by the document width
- Bounding box y_0 normalized by the document height
- Page number normalized by the number of pages of the document
- Language, one-hot encoding of language class: English, French and Spanish
- Font name, pre-processed by removing punctuation and foundry names (i.e., LT, MT, FF, EF) by the font name.
- Font size

We use the same special ID for padding the character sequences and for out-of-dictionary characters. Financial documents can be grouped based on several different aspects. In particular, the language can be discriminative since in some countries the financial documents have to follow specific templates (e.g., EDGAR SEC [6] or AMF[7]). Previous works show that documents belonging to the same class often share the same specific page layout pattern (Esposito et al., 1990). For this reason, we argue that the use of the document class can represent a relevant feature in the TOC task.

5. Results

Results on the Title Detection and Table of Content tasks are reported in Table 4.1 and Table 4.1, respectively. The Title Detection task is evaluated using the F-measure computed on the predicted titles that match the ground truth titles. The TOC task instead evaluates the systems against the harmonic mean computed over the Inex08 F1 score and the Inex08-Level accuracy. In particular, for the Inex08 F1 score the predicted TOC entries are considered correct if match the ground truth TOC entries and have the same page number. The Inex08-Level accuracy evaluates the number of predicted titles with the correct page number and the correct hierarchical level.

We perform a qualitative analysis on the three document classes, i.e. English, French and Spanish documents. The English fund documents are simple and of regulatory nature. The French fund documents are also regulatory but more oriented to investors with graphical elements and colour. The Spanish documents are annual reports with a strong emphasis on creative communication with a large variety in form, colour, text flow and photographs, which makes them less predictable. Our system ranked #1 on the Spanish TD subtask with an F1 score of 0.569 and the TOC subtask with a harmonic mean of 43.01. The system performs better for the Title Detection task in English achieving an F1 score of 0.838. On the other side, for the level classification, the system performs better in Spanish, achieving 46.5 of Inex08-Level accuracy.

6. Conclusion

In this work, we presented our system submitted to the FinTOC 2022 task. Our system ranked #1 on the Spanish subtask and #3 on the English and French subtasks, achieving high recall performance. The Title Detection module is language independent and can be extended to a wider scope of documents written in other languages than English, Spanish and French.

In future developments, we plan to fine-tune hyperparameters, such as the level of confidence of the Title Detection model to improve the system performance.

7. Acknowledgements

This research has been partially funded by Objectway SpA under the “intelligent Business Platform for Wealth & Asset Management - iBWAM” project. The authors would like to thank Dr. Sciancalepore and Dr. Marvulli for supporting the preliminary development of the Title Detection pipeline.

8. References

Bentabet, N., Juge, R., and Ferradans, S. (2019). Table-of-Contents Generation on Contemporary Documents. In *2019 International Conference on Document Analysis and Recognition, ICDAR 2019, Sydney, Australia, September 20-25, 2019*, pages 100–107. IEEE.

Bentabet, N.-I., Juge, R., El Maarouf, I., Mouilleron, V., Valsamou-Stanislawski, D., and El-Haj, M. (2020). The financial document structure extraction shared task (FinToc 2020). In *Proceedings of the 1st Joint Workshop on Financial Narrative Processing*

[6] https://www.sec.gov/edgar.shtml

[7] https://www.amf-france.org/

and MultiLing Financial Summarisation, pages 13–22, Barcelona, Spain (Online), December. COLING.

Bourez, C. (2021). FINTOC 2021 - Document Structure Understanding. In *Proceedings of the 3rd Financial Narrative Processing Workshop*, pages 89–93, Lancaster, United Kingdom, 15-16 September. Association for Computational Linguistics.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Chen, T. and Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In Balaji Krishnapuram, et al., editors, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 785–794. ACM.

Esposito, F., Malerba, D., Semeraro, G., Annese, E., and Scafuro, G. (1990). An experimental page layout recognition system for office document automatic classification: an integrated approach for inductive generalization. In *Proceedings 10th International Conference on Pattern Recognition*, volume 1, pages 557–562.

Girshick, R. B., Donahue, J., Darrell, T., and Malik, J. (2014). Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 580–587. IEEE Computer Society.

Girshick, R. B. (2015). Fast R-CNN. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 1440–1448. IEEE Computer Society.

Gupta, A., Akl, H. A., and de Mazancourt, H. (2021). Not All Titles are Created Equal: Financial Document Structure Extraction Shared Task. In *Proceedings of the 3rd Financial Narrative Processing Workshop*, pages 86–88, Lancaster, United Kingdom, 15-16 September. Association for Computational Linguistics.

Hercig, T. and Kral, P. (2020). UWB@FinTOC-2020 Shared Task: Financial Document Title Detection. In *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*, pages 158–162, Barcelona, Spain (Online), December. COLING.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Lin, T., Dollár, P., Girshick, R. B., He, K., Hariharan, B., and Belongie, S. J. (2017). Feature Pyramid Networks for Object Detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 936–944. IEEE Computer Society.

Maarouf, I. E., Kang, J., Azzi, A. A., Bellato, S., Gan, M., and El-Haj, M. (2021). The Financial Document Structure Extraction Shared Task (FinTOC2021). In *Proceedings of the 3rd Financial Narrative Processing Workshop*, pages 111–119, Lancaster, United Kingdom, 15-16 September. Association for Computational Linguistics.

Ren, S., He, K., Girshick, R. B., and Sun, J. (2015). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In Corinna Cortes, et al., editors, *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 91–99.

Shen, Z., Zhang, R., Dell, M., Lee, B. C. G., Carlson, J., and Li, W. (2021). LayoutParser: A Unified Toolkit for Deep Learning Based Document Image Analysis. *arXiv preprint arXiv:2103.15348*.

van de Sande, K. E. A., Uijlings, J. R. R., Gevers, T., and Smeulders, A. W. M. (2011). Segmentation as selective search for object recognition. In Dimitris N. Metaxas, et al., editors, *IEEE International Conference on Computer Vision, ICCV 2011, Barcelona, Spain, November 6-13, 2011*, pages 1879–1886. IEEE Computer Society.

Zhong, X., Tang, J., and Jimeno-Yepes, A. (2019). PubLayNet: Largest Dataset Ever for Document Layout Analysis. In *2019 International Conference on Document Analysis and Recognition, ICDAR 2019, Sydney, Australia, September 20-25, 2019*, pages 1015–1022. IEEE.