The FinTOC-2019 Shared Task: Financial Document Structure Extraction

Rémi Juge[*]
Fortia Financial Solutions
Paris, France

Najah-Imane Bentabet[*]
Fortia Financial Solutions
Paris, France

Sira Ferradans
Fortia Financial Solutions
Paris, France

name.surname@fortia.fr

Abstract

In this paper, we present the results and findings of The FinTOC-2019 Shared Task on structure extraction from financial documents. This shared task was organized as part of Second Financial Narrative Processing Workshop, collocated with the 22nd Nordic Conference on Computational Linguistics (NoDaLiDa'19) Conference. The shared task aimed at collecting systems for extracting table of contents from Financial prospectuses by detecting the document titles and reorganizing them in a hierarchical way. The FinTOC shared task is the first to target the task of Table of content extraction in the domain of Finance.

1 Introduction

Long document comprehension is still an open problem in Natural Language Processing (NLP). Most of the corporate information or academic knowledge is locked in long documents (> 10 pages) with complex semantic and layout structure. Documents are generally converted into plain text and processed sentence by sentence, where the only structure that is easily identified are the paragraphs, thus loosing the internal organization of the document. Despite the importance long document analysis, there are few available resources and none in a low resource domain such as the finance.

In this shared task, we focus on extracting the table-of-contents (TOC) of financial prospectuses that are official pdf documents in which investment funds precisely describe their characteristics and investment modalities. The majority of prospectuses are published without a TOC which is of fundamental importance for sophisticated NLP tasks such as information extraction or question answering on long documents. Although the content they must include is often regulated, their format is not standardized and displays a great deal of variability ranging from plain text format, towards more graphical and tabular presentation of data and information, making the analysis of the discourse structure even more complicated.

In this paper, we report the results and findings of the FinTOC-2019 shared task.[1] The Shared Task was organized as part of Second Financial Narrative Processing Workshop, co-located with the 22nd Nordic Conference on Computational Linguistics (NoDaLiDa'19) Conference.[2]

A total of 6 teams submitted runs and contributed 4 system description papers. All system description papers are included in the FNP 2019 workshop proceedings and cited in this report.

2 Previous Work on TOC extraction

We find mostly two approaches. The goal of the first approach is to parse the hierarchical structure of sections and subsections from the TOC pages embedded in the document. Most of the research developed in this area has been linked to the INEX [1] and ICDAR competitions [2, 3, 4] which target old and long OCR-ised books instead of small papers. These documents are very different from the documents that we target in this shared task, charaterized by having complex layout structure (see Fig. 1 for some examples). Outside these competitions, we find the methods proposed by El-Haj et al [5, 6, 7], based also in TOC page parsing.

In the second approach, we find methods that detect headings using learning methods based on layout and text features. The set of titles are hierarchically ordered according to a predefined rule-based function [2, 8, 9]. Recently, we find methods

[*]Both authors contributed equally to this work

[1] http://wp.lancs.ac.uk/cfie/fnp2019/
[2] http://wp.lancs.ac.uk/cfie/

Figure 1: Random pages from the investment document data set. We observe that the title organization and, in general, the layout is complex.

that address TOC extraction as a sequence labelling task to which deep learning methods can be applied [10].

3 Task Description

As part of the Financial Narrative Processing Workshop, we present a shared task on Financial Document Structure Extraction.

Systems participating in this shared task were given a sample collection of financial prospectuses with a wide variety of document structures and sizes. The goal was to automatically process them to extract their document structure. In fact the task was decomposed into two subtasks:

- **Title detection**: The document is splitted into text blocks (a text block regroup lines that have the same layout and that are spatially close to each others) extracted from financial prospectuses by our in-house parser. Each text block needs to be classified as a 'title' or 'non-title'. As shown in Fig. 2 the titles can have different layouts (marked with red and green boxes) and they have to be distinguished from the regular text ('non-title' with grey boxes).

- **TOC extraction**: In this subtask, the goal is to (i) identify the hierarchical level of the titles, for instance, in Fig. 2, the text in green bounding boxes are hierarchically at the same level and at a different level than the title in red, (ii) organize the titles of the document according to the hierarchical structure to produce the final TOC. Again in Fig. 2, the system needs to identify that the red tagged heading is hierarchically above than the green ones.

  It is important to note that two titles, with the same layout and the same text can have different hierarchical levels depending on their location in the document.

All participating teams were provided a common training data set for subtask 1 which included the original pdfs, the xml versions of the pdfs obtained using the Poppler[3] library, and a csv file containing, for each text block, a set of layout features and binary labels indicating if the text block is a title or not. For the second subtask, the training data set also included a TOC of the documents in the xml format proposed by ICDAR competitions[2]. A blind test set was used to evaluate the output of the participating teams.

As stated in Section 2, most of the previous research on TOC generation has been confined to short papers such as research publications (*Arxiv* database), or standard documents such as digitalized books. However, the task of extracting the TOC of commercial documents with a complex layout structure in the domain of finance is not much explored in the literature.

4 Shared Task Data

Next, we discuss the corpora used for the title detection and TOC extraction subtasks.

4.1 Corpus annotation

Financial prospectuses are available online in a pdf format and are also made available from asset managers. We compiled a list of 58 prospectuses from Luxembourg written in English to create the

[3] poppler.freedesktop.org

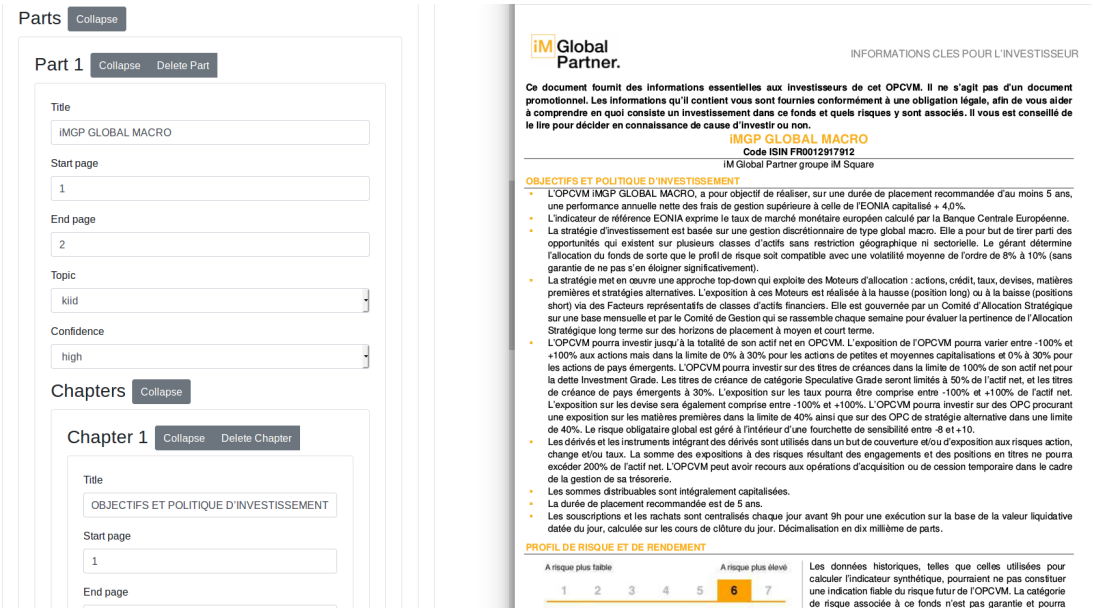


Figure 2: Screenshot of the annotation tool developed internally.

data sets of the subtasks. We chose prospectuses with a wide variety of layouts and styles.

	Xerox F1	Inex08 F1
tagger 1 & tagger 2	93.8%	87.5%
tagger 1 & reviewer	96.7%	92%
tagger 2 & reviewer	96.8%	93.5%

Table 1: Agreement scores between different annotators of the investment document data set.

We provided three annotators with the original pdfs and an internally developed web tool that produces a hierarchical *json* file containing each TOC-entry together with some features (title, starting-page, ending-page and children). Each annotator was ask to:

1. Identify the title: Locate a title inside the pdf document.

2. Associate the entry level in the TOC: Every title must have an entry level in the TOC of the document with the following constraints 1) high level entries cannot be inside lower level entries (i.e. a *Part* cannot be inside a *Chapter*), 2) the entries levels must be successive (i.e. after a *chapter* we have a *section* not a *subsection*).

3. Add title: Copy-paste the title text directly into a web form, see Fig. 2 label *Title*

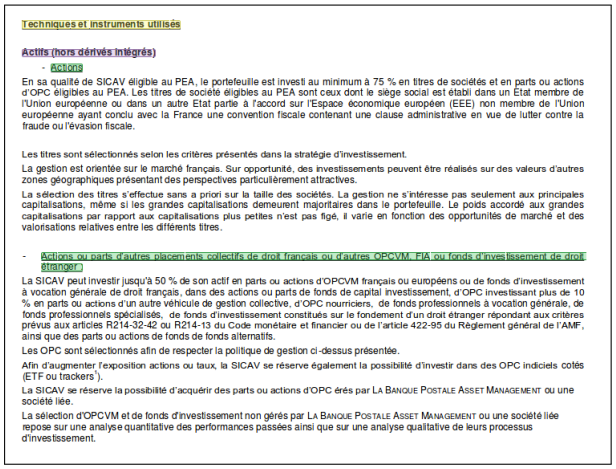
The predefined type of entry levels for the TOC were *Part*, *Chapter*, *section*, *subsection* or *paragraph*, that could be inside the *Front matter*, *Body matter* or *Back matter*. Therefore, the maximum TOC level was 5.

Each document was annotated independently by two people and a third person would review the annotations to resolve the possible conflicts. The agreement scores between annotators are depicted in Table 1. We can observe high agreement scores, allowing us to be confident enough about the quality of our data set.

Annotation Challenge: Headings identification Investment prospectuses are commercial documents whose complex layout aims at highlighting specific information such that a potential investor can identify it quickly. Hence, annotating a title and its level in the TOC hierarchy is a difficult task as one cannot rely on the visual appearance of the title to do so. Some examples can be observed in Fig. 3 and Fig. 4.

Annotation Challenge: Tagging pdf documents. The annotation of pdf documents is not evident since they are meant to be used for display. The tool we developed for the annotations does not allow annotators to directly annotate on the pdf and thus they had to manage two different platforms at the same time. Working this way is prone to mistakes.

Annotation Challenge: Matching annotations and text blocks. Our internal tool uses a copy-paste mechanism to create the TOC entries, intro-


Techniques et instruments utilisés

Actifs (hors dérivés intégrés)

- Actions

En sa qualité de SICAV éligible au PEA, le portefeuille est investi au minimum à 75 % en titres de sociétés et en parts ou actions d'OPC éligibles au PEA. Les titres de société éligibles au PEA sont ceux dont le siège social est établi dans un État membre de l'Union européenne ou dans un autre Etat partie à l'accord sur l'Espace économique européen (EEE) non membre de l'Union européenne ayant conclu avec la France une convention fiscale contenant une clause administrative en vue de lutter contre la fraude ou l'évasion fiscale.

Les titres sont sélectionnés selon les critères présentés dans la stratégie d'investissement.

La gestion est orientée sur le marché français. Sur opportunité, des investissements peuvent être réalisés sur des valeurs d'autres zones géographiques présentant des perspectives particulièrement attractives.

La sélection des titres s'effectue sans a priori sur la taille des sociétés. La gestion ne s'intéresse pas seulement aux principales capitalisations, même si les grandes capitalisations demeurent majoritaires dans le portefeuille. Le poids accordé aux grandes capitalisations par rapport aux capitalisations plus petites n'est pas figé, il varie en fonction des opportunités de marché et des valorisations relatives entre les différents titres.

- Actions ou parts d'autres placements collectifs de droit français ou d'autres OPCVM, FIA ou fonds d'investissement de droit étranger :

La SICAV peut investir jusqu'à 50 % de son actif en parts ou actions d'OPCVM français ou européens ou de fonds d'investissement à vocation générale de droit français, dans des actions ou parts de fonds de capital investissement, d'OPC investissant plus de 10 % en parts ou actions d'un autre véhicule de gestion collective, d'OPC nourriciers, de fonds professionnels à vocation générale, de fonds professionnels spécialisés, de fonds d'investissement constitués sur le fondement d'un droit étranger répondant aux critères prévus aux articles R214-32-42 ou R214-13 du Code monétaire et financier ou de l'article 422-95 du Règlement général de l'AMF, ainsi que des parts ou actions de fonds de fonds alternatifs.

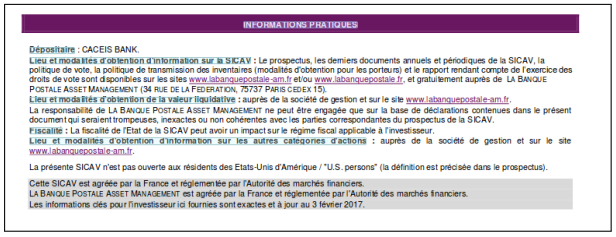
Les OPC sont sélectionnés afin de respecter la politique de gestion ci-dessus présentée.

Afin d'augmenter l'exposition actions ou taux, la SICAV se réserve également la possibilité d'investir dans des OPC indiciels cotés (ETF ou trackers').

La SICAV se réserve la possibilité d'acquérir des parts ou actions d'OPC érés par LA BANQUE POSTALE ASSET MANAGEMENT ou une société liée.

La sélection d'OPCVM et de fonds d'investissement non gérés par LA BANQUE POSTALE ASSET MANAGEMENT ou une société liée repose sur une analyse quantitative des performances passées ainsi que sur une analyse qualitative de leurs processus d'investissement.

Figure 3: Page from a prospectus with the titles selected coloured boxes. An example where title identification is not evident because titles can have the same style as regular text.


INFORMATIONS PRATIQUES

Dépositaire : CACEIS BANK.

Lieu et modalités d'obtention d'information sur la SICAV : Le prospectus, les derniers documents annuels et périodiques de la SICAV, la politique de vote, la politique de transmission des inventaires (modalités d'obtention pour les porteurs) et le rapport rendant compte de l'exercice des droits de vote sont disponibles sur les sites www.labanquepostale-am.fr et/ou www.labanquepostale.fr, et gratuitement auprès de LA BANQUE POSTALE ASSET MANAGEMENT (34 RUE DE LA FÉDERATION, 75737 PARIS CEDEX 15).

Lieu et modalités d'obtention de la valeur liquidative : auprès de la société de gestion et sur le site www.labanquepostale-am.fr.

La responsabilité de LA BANQUE POSTALE ASSET MANAGEMENT ne peut être engagée que sur la base de déclarations contenues dans le présent document qui seraient trompeuses, inexactes ou non cohérentes avec les parties correspondantes du prospectus de la SICAV.

Fiscalité : La fiscalité de l'Etat de la SICAV peut avoir un impact sur le régime fiscal applicable à l'investisseur.

Lieu et modalités d'obtention d'information sur les autres catégories d'actions : auprès de la société de gestion et sur le site www.labanquepostale-am.fr.

La présente SICAV n'est pas ouverte aux résidents des Etats-Unis d'Amérique / "U.S. persons" (la définition est précisée dans le prospectus).

Cette SICAV est agréée par la France et réglementée par l'Autorité des marchés financiers.
LA BANQUE POSTALE ASSET MANAGEMENT est agréée par la France et réglementée par l'Autorité des marchés financiers.
Les informations clés pour l'investisseur ici fournies sont exactes et à jour au 3 février 2017.

Figure 4: Page from a prospectus with the titles selected coloured boxes. An example where title identification is not evident because titles may expand a part of a line.

ducing some noise at the string level. On the other hand, we extract text from the pdf using an automatic pdf to xml process. For the data set creation, each title annotation had to be matched to a text block. This pipeline introduces noise in the final csv.

4.2 Corpus Description

In the following, we provide an analysis of the data used for the shared task. For both subtasks, the released training sets were the same excepted that for subtask 2, an additional xml file (*groundtruth[...]max_depth=5.icdar2013.xml*) at the ICDAR format was given. The reason for it is twofold: it gave to the participants the output format that they had to respect for the submissions and allowed them to participate in subtask 2 without having a title extraction system from subtask 1.

In the csv files available to the participants, each text block came with a set of layout features: *is_bold, is_italic, is_all_caps, begins_with_cap, begins_with_numbering and page_number* and its source xml file. Some statistics on this data set are presented in Table 2.

number of documents	58
average number of pages	90
number of text blocks	90441
number of titles (% of text blocks)	14%
begin_with_numbering (% of text blocks)	20%
is_bold (% of text blocks)	18%
is_italic (% of text blocks)	1.3%
is_all_caps (% of text blocks)	20%
begins_with_cap (% of text blocks)	68%
level 1 (% of titles)	7%
level 2 (% of titles)	26%
level 3 (% of titles)	33%
level 4 (% of titles)	30%
level 5 (% of titles)	4%

Table 2: Statistics on the investment document data set.

5 Participants and Systems

	# teams	# std runs
subtask 1	5	10
subtask 2	2	3
papers	5	-

Table 3: Statistics on the participation in the two subatsks.

A total of 24 teams registered in the shared task from 18 different institutions, and 6 teams participated with standard runs and 5 submitted a paper with the description of their method, see Table 4 for more information about their affiliation. In Table 3, we show the details on the submissions per task. It is important to note that not all the participants that submitted a standard run, sent a paper describing

Team	**Affiliation**	Tasks
Daniel [11]	STIH, Sorbonne Université	1 and 2
FinDSE [12]	Faculdade de Engenharia da Universidade do Porto	1
UWB [13]	University of West Bohemia	1
YseopLab [14]	Yseop	1
IHSMarkit	IHS Markit	2
Aiai	OPT, Inc	1

Table 4: List of the 6 teams that participated in Subtasks of the FinTOC Shared Task.

their approach.

Participating teams explored and implemented a wide variety of techniques and features. In this section, we give a brief description of each system, more details could be found in the description papers appearing in the proceedings of the FNP 2019 Workshop.

Daniel [11]: The only team to submit to both subtasks and a paper. Their approach for the FinTOC title detection task assumed the presence of a TOC page which they detect by identifying the page numbers that are aligned at the right of the page. Then, they extract each TOC entry using regular expressions and construct the hierarchical structure of the TOC with a rule-based method based on indentation and multi-level numbering.

FinDSE [12]: They addressed the FinTOC Title detection as a sentence classification task. They added to the provided features (see Section 4.2 for more details) some others such as morphological (number of characters distributed into categories), semantic (contains date) and linguistic features (predetermined tokens such as 'appendix', 'annex', etc, part-of-speech of the first word, ...). Their best performing model used an extra-tree classifier. It is interesting to note that, according to their experiments, adding predetermined tokens actually reduced the performance of the final method.

UWB [13]: Only the FinTOC Title detection was addressed in this paper. As for the other methods, they state the problem as a binary classification of text sentences, for which they use a Maximum entropy classifier, on top of a diverse set of features. In addition to the provided characteristics, they add others related to style (font size, font type size), orthographic descriptors, and char n-grams.

YseopLab [14]: The authors tackle only the FinTOC Title detection task. Similarly to other participants, they first try to design an additional set of features to feed an SVM classifier. Then, unlike previous methods, they run two separated experiments where they use a character-level CNN and a word-level BiLSTM with attention to extract semantic features from text blocks and classify them.

Aiai [15]: This team proposes the use of word2vec word-embeddings followed by a LSTM and BiLSTM, respectively for run 1 and 2, see Table 5. Then, they add an attention layer. Finally, they train several times the same model and do ensembling as a last step.

6 Results and Discussion

Evaluation Metric For the first subtask, the participating systems are ranked based on the weighted F1 score obtained on a blind test set (official metric). Table 5 reports the results obtained on FinTOC title detection task by the teams detailed in the previous section.

Team	F1 score
Aiai_2	**0.982**
Aiai_1	0.98
UWB_2	0.972
YseopLab_2	0.9716
FinDSE_1	0.970
FinDSE_2	0.968
UWB_1	0.965
Daniel_1	0.949
Daniel_2	0.942
YseopLab_1	0.932

Table 5: Results obtained by the participants for the first FinTOC task. The teams are ordered by the weighted F1 score.

Regarding the FinTOC TOC extraction subtask, the metric is based on the official title-based measure of the ICDAR 2013 competition on book structure extraction [2] (ICDAR'13 measure from now on). More specifically, the final F1 score is the mean of the InexF1 score and the Inex level accuracy. For the results on this task, please check Table 6.

Team	ICDAR'13 measure
Daniel_1	**0.427**
IHSMarkit_1	0.39
IHSMarkit_2	0.388

Table 6: Results obtained by the participants for the FinTOC TOC extraction task. The teams are ordered by the ICDAR'13 measure (see the text for more details).

Discussion. A surprising fact of the reported methods is that the best performing methods (Aiai_1 and Aiai_2 with 0.98, UWB-2 with 0.972 and YseopLab_2 with 0.9716 F1 score) have radically different approaches. Team UWB-2 does not use deep learning methods. Instead, they add

meaningful features to a maximum entropy classifier and they show through ablation tests that all features are important to attain their result. On the other hand, the best performing system of YseopLab_2 implements character-level CNN with no hand-engineered features. Finally, both methods of Aiai use (Bi-)LSTMs with attention mechanisms on top of word embeddings. For the second task, only one paper was submitted describing Daniel_1 team's method, which proposed a rule-based approach to title hierarchization.

In their paper [12], the team FinDSE performs a set of experiments with a wide variety of features. An interesting conclusion is that the usage of common first words from titles such as *Annex* or *Appendix* can be counter productive. This contradicts the methods commonly used in the literature [16, 17, 18, 19]. Moreover, it shows the difficulty of transferring state-of-the-art methods trained on public datasets to commercial documents such as financial prospectuses.

7 Conclusions

In this paper we presented the setup and results for the FinTOC-2019 Shared Task:Financial Document Structure Extraction, organized as part of the Second Financial Narrative Processing Workshop, collocated with the 22nd Nordic Conference on Computational Linguistics (NoDaLiDa'19) Conference. A total of 24 people registered from 18 different institutions. 6 teams participated in the shared task with a wide variety of techniques.

We introduced a new data set on the TOC extraction problem in text automatically extracted from pdf files in English. This scenario is very realistic in everyday applications which may explain the participation of public universities and profit organizations from the financial domain.

Acknowledgments

We would like to thank our dedicated annotators who contributed to the building of the corpora used in this Shared Task: Anais Koptient, Aouataf Djillani, and Lidia Duarte.

References

[1] Bodin Dresevic, Aleksandar Uzelac, Bogdan Radakovic, and Nikola Todic. Book layout analysis: Toc structure extraction engine. In Shlomo Geva, Jaap Kamps, and Andrew Trotman, editors, *Advances in Focused Retrieval*, pages 164–171, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg.

[2] Antoine Doucet, Gabriella Kazai, Sebastian Colutto, and Günter Mühlberger. Icdar 2013 competition on book structure extraction. In *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*, pages 1438–1443. IEEE, 2013.

[3] Thomas Beckers, Patrice Bellot, Gianluca Demartini, Ludovic Denoyer, Christopher M. De Vries, Antoine Doucet, Khairun Nisa Fachry, Norbert Fuhr, Patrick Gallinari, Shlomo Geva, Wei-Che Huang, Tereza Iofciu, Jaap Kamps, Gabriella Kazai, Marijn Koolen, Sangeetha Kutty, Monica Landoni, Miro Lehtonen, Véronique Moriceau, Richi Nayak, Ragnar Nordlie, Nils Pharo, Eric Sanjuan, Ralf Schenkel, Xavier Tannier, Martin Theobald, James A. Thom, Andrew Trotman, and Arjen P. De Vries. Report on INEX 2009. *Sigir Forum*, 44(1):38–57, June 2010. Article disponible en ligne : http://www.cs.otago.ac.nz/homepages/andrew/papers/2010-4.pdf.

[4] Thi Tuyet Hai Nguyen, Antoine Doucet, and Mickael Coustaty. Enhancing table of contents extraction by system aggregation. In *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, 2018.

[5] Mahmoud El Haj, Paul Rayson, Steven Young, and Martin Walker. *Detecting document structure in a very large corpus of UK financial reports*. LREC'14 Ninth International Conference on Language Resources and Evaluation. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014) . European Language Resources Association (ELRA), Reykjavik, Iceland, pp. 1335-1338, 2014.

[6] Mahmoud El Haj, Paul Edward Rayson, Steven Eric Young, Paulo Alves, and Carlos Herrero Zorita. *Multilingual Financial Narrative Processing: Analysing Annual Reports in English, Spanish and Portuguese*. World Scientific Publishing, 2 2019.

[7] Mahmoud El-Haj, Paulo Alves, Paul Rayson, Martin Walker, and Steven Young. Retrieving, classifying and analysing narrative commentary in unstructured (glossy) annual reports published as pdf files. *Accounting and Business Research*, pages 1–29, 2019.

[8] Caihua Liu, Jiajun Chen, Xiaofeng Zhang, Jie Liu, and Yalou Huang. Toc structure extraction from ocred books. In *International Workshop of the Initiative for the Evaluation of XML Retrieval*, pages 98–108. Springer, 2011.

[9] Abhijith Athreya Mysore Gopinath, Shomir Wilson, and Norman Sadeh. Supervised and unsupervised methods for robust separation of section titles and prose text in web documents. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 850–855, 2018.

[10] Najah-Imane Bentabet, Rémi Juge, and Sira Ferradans. Table-of-contents generation on contemporary documents. In *Proceedings of ICDAR 2019*.

[11] Gael Lejeune Emmanuel Giguet. Daniel fintoc-2019 shared task: Toc extraction and title detection. In *The Second Workshop on Financial Narrative Processing of NoDalida 2019*, 2019.

[12] Henrique Cardoso Carla Abreu and Eugénio Oliveira. Findsefintoc-2019 shared task. In *The Second Workshop on Financial Narrative Processing of NoDalida 2019*, 2019.

[13] Tomas Hercig and Pavel Král. Uwbfintoc-2019 shared task: Financial document title detection. In *The Second Workshop on Financial Narrative Processing of NoDalida 2019*, 2019.

[14] Anubhav Gupta Hannah Abi-Akl and Dominique Mariko. Fintoc-2019 shared task: Finding title in text blocks. In *The Second Workshop on Financial Narrative Processing of NoDalida 2019*, 2019.

[15] Ke Tian and Zi Jun Peng. Finance document extraction using data augmented and attention. In *The Second Workshop on Financial Narrative Processing of NoDalida 2019*, 2019.

[16] Anoop M Namboodiri and Anil K Jain. Document structure and layout analysis. In *Digital Document Processing*, pages 29–48. Springer, 2007.

[17] Alan Conway. Page grammars and page parsing. a syntactic approach to document layout recognition. In *Document Analysis and Recognition, 1993., Proceedings of the Second International Conference on*, pages 761–764. IEEE, 1993.

[18] Florendia Fourli-Kartsouni, Kostas Slavakis, Georgios Kouroupetroglou, and Sergios Theodoridis. A bayesian network approach to semantic labelling of text formatting in xml corpora of documents. In *International Conference on Universal Access in Human-Computer Interaction*, pages 299–308. Springer, 2007.

[19] Koji Nakagawa, Akihiro Nomura, and Masakazu Suzuki. Extraction of logical structure from articles in mathematics. In *International Conference on Mathematical Knowledge Management*, pages 276–289. Springer, 2004.