11th MALAYSIA STATISTICS CONFERENCE 2024
Data and Artificial Intelligence: Empowering the Future
Sasana Kijang, Bank Negara Malaysia
19th September 2024

Application of Language Models to Analyse Scanned Textual Compliance Reports

Chun Onn Pam[1]; Nur Aisyah binti Mohd Nasir[1]; Siti Hajar binti Khairul Anwar[1]

[1] Payment Services Oversight Department, Central Bank of Malaysia
(chunonn.pam@bnm.gov.my, aisyah.nasir@bnm.gov.my, hajaranwar@bnm.gov.my)

Abstract:

The use of supervisory technology (SupTech) has enhanced the ability to analyse various structured and unstructured data, facilitating supervisory authorities in ensuring regulatees' compliance to the set standards and requirements. Reviewing large volumes of text-based compliance reports is one of the challenges in identifying non-compliances by regulatees efficiently. Another challenge is the reports are prepared in various formats e.g. image-based PDFs and presentation slides, making it difficult to identify an appropriate tool to automate the analysis. This paper proposes the use of Optical Character Recognition (OCR) and language models (LMs) to address these challenges. The OCR results indicate PyTesseract tool performs well in extracting text from images while the Img2Table feature in the PaddleOCR tool excels in extracting text from tabular formats. This paper also examines transformer-based models, i.e. Text-To-Text Transfer Transformer (T5), Bidirectional and Auto-Regressive Transformers (BART) and Generative Pretrained Transformer 2 (GPT-2), for their effectiveness in summarising text. The results reveal that BART produces high-quality summaries that effectively capture key information while maintaining clarity and coherence as compared to other LMs.

Keywords:

Suptech; payment sector; optical character recognition; language models; text summarisation.

[1] The views expressed herein are those of the authors and do not necessarily reflect those of the Central Bank of Malaysia.

1. Introduction:

In recent years, the use of Suptech has increased globally, assisting supervisory authorities to make more data-driven and timely decisions (Di Castri et al, 2018). Since 2017, the Central Bank of Malaysia (BNM) has adopted Suptech to identify potential suspicious activities relating to money laundering and terrorist financing (ML/TF), and to ensure effective supervision of the large number of regulatees in the payments sector, specifically the Money Services Business (MSB) licensees and E-money Issuers (EMI) (Bank Negara Malaysia, 2020).

On annual basis, MSB and EMI regulatees submit text-based compliance reports prepared by external auditors to BNM, consisting agreed-upon procedures (AUP) for MSB and fund management reports (FMR) for EMI. However, the large number of lengthy reports in various formats, including imaged-based PDF and table-based report slides, has strained supervisors’ resources to analyse this information. The methodology proposed to overcome this issue is two-fold. First, we apply OCR to convert the textual reports submitted by regulatees into machine-readable text. Second, the extracted text is summarised using LM to provide a concise overview of the report contents. With the summarised text, supervisors can identify any findings from the compliance reports more efficiently.

2. Methodology:

2.1 Data Source

This study analyses the scanned text-based reports submitted annually in PDF format by MSB and EMI regulatees, based on the requirements stipulated in the E-Money Policy Document (Bank Negara Malaysia, 2022) and Money Services Business Act (Bank Negara Malaysia, 2011), respectively. The reports cover assessments of various aspects of compliance from 2020 to 2022, involving more than 300 regulatees, with each report ranging from 1 to 40 pages.

2.2 Optical Character Recognition (OCR)

Since the reports are prepared by different external auditors, we observe various report formats. Some reports are produced in letter format, that is easy for data extraction, while others include complex data structures e.g. tables, which can be in scanned slide images. Hence, the reports are further classified based on similar external auditor for the ease of data pre-processing and extraction. Identification of these auditors for the regulatees has been done through simple OCR on the first two pages or by matching the regulatees’ data within our database. For pre-processing steps, we convert the reports to grayscale images, apply thresholding to create binary images, remove noise, and resize the images depending on the format using fitz (PyMuPDF), OpenCV and PIL (Pillow) python libraries. The pre-processed images are then passed to the OCR engine for text extraction (Tuychiev, 2024). We utilise two approaches for the OCR engine: PyTesseract and Img2Table with PaddleOCR. These OCRs are selected due to their open-source availability and their capability to analyse data offline on our local devices. This is crucial for preserving data privacy and confidentiality, as it allows us to extract data without sending it through API-based engine or online services to third parties.

2.2.1 PyTesseract

PyTesseract (Python wrapper for Tesseract), is used to identify the regions of interest (ROIs) in the images that contain text for extraction. We leverage the flexibility of Tesseract, which allows users to extract text using the character pattern engine (Tesseract 3 engine) that uses statistical models, and the LSTM engine (Tesseract 4 engine) that uses neural networks. Tesseract performs spell checking, context correction and format adjustments after the extraction (Zelic & Sable, 2023). In our study, PyTesseract is particularly effective for extracting paragraph text data due to its high accuracy and fast computation. However, it is less effective for text extraction from table structures due to its limitations (Rosebrock, 2022).

2.2.2 Img2Table with PaddleOCR

While PyTesseract performs well with paragraph text data, it struggles with tabular text data. This is where Img2Table with the PaddleOCR engine can fill the gap. Img2Table is a Python library that leverages OpenCV to identify and extract tables from images. It supports various OCR engines, swiftly extracts tables into a data frame format with the implicit rows and borderless table functionality to handle complex tables (xavctn, n.d.). Built on a deep learning framework named PaddlePaddle, PaddleOCR is able to identify the structure of the page, be it tables or paragraph text. It also offers a selection of models to choose from, such as lightweight model that can be executed on normal CPU, model for multilingual detection, etc. In this study, we utilise the PP-OCRv3 model, which consists of three parts: detection, classification, and recognition. Each part has its own model trained within the framework (Czerwinska, 2023).

2.3 Language Models (LMs) and Regular Expressions (Regex)

We then employ LMs to summarise extracted text, guided by the research conducted by Garg et al. (2020) and Rao et al. (2024). We utilise the BART and T5 models, as well as GPT - 2. The transformer-based models are chosen for their capability to process data locally on devices, without the need of GPU. Regex is subsequently applied to the extracted data to identify the keyword for compliance detection. BART pre-trains a standard sequence-to-sequence (seq2seq) model with denoising autoencoder. BART merges bidirectional encoder with left-to-right decoder, making text summarisation effective due to the capability to process text out of order (Lewis, 2019). T5 is trained on a large corpus of text, employing a unified text-to-text approach and is subsequently fine-tuned for every language task (Chen, 2021). GPT-2 is pre-trained on a different type of internet text data but is not fine-tuned for any specific task. It generates coherent text paragraphs, relying entirely on its ability to predict the next word in a sentence (Radford, 2019).

2.4 Performance Metrics to evaluate LMs

This study primarily focuses on Recall-Oriented Understudy for Gisting Evaluation – Longest Common Subsequence (LCS) metric (ROUGE-L), which is used in the research conducted by Rao et al. (2024), along with two additional metrics, Bilingual Evaluation Understudy (BLEU) and Bidirectional Encoder Representations from Transformers (BERT) Score to evaluate the summarisation performance of the LMs. (Mansuy, n.d.)

ROUGE-L measures similarity between automatically summarised text and the extracted text using the LCS. A longer shared sequence indicates higher similarity. The F1-Score is computed using precision (ratio of length of the LCS over the number of unigrams in summarised data) and recall (ratio of length of the LCS over the number of unigrams in extracted data) as given by the equation (1).

$$ROUGE\text{-}L,\, F1 - score = \frac{2\times(precision\times recall)}{precision+recall}\ , \qquad (1)$$

BLEU counts matching n-grams in the summarised text to n-grams extracted text. It is a geometric mean of the n-gram precisions and computed based on *brevity_penalty* (the length of the summarised text divided by the length of the extracted text), w_n (the weight applied to the n-gram accuracy score), and p_n (the n-grams precision). The formula is given in equation (2).

$$\mathrm{BLEU} = brevity\_penalty \times e^{\sum_{n=1}^{N} w_n \log(p_n)}, \qquad (2)$$

BERTScore is a metric leverages the pre-trained BERT LM to measure similarity between two sentences using the cosine similarity formula, as given in equation (3).

$$Similarity\ (A, B) = \cos(\theta) = \frac{A\cdot B}{\|A\|\|B\|}\ , \qquad (3)$$

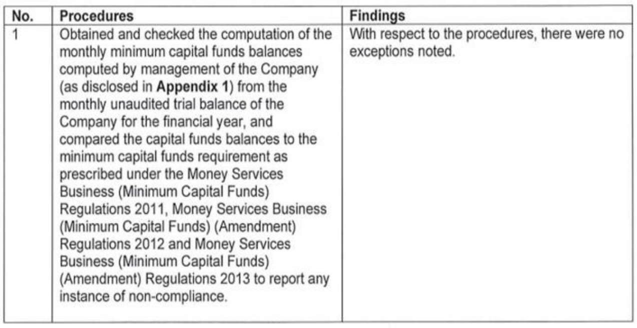
3. Result:

3.1 Optical Character Recognition

For each pre-processed image, tools like PaddleOCR and PyTesseract with regex are used to detect the page structure or keywords to identify the compliance region. This is based on preliminary analysis conducted on different reports from different external auditors. This process is generalised to ensure flexibility in handling different conditions that might exist during OCR extraction and to prevent errors and incorrect extraction, even when deviations from the template occur.

Figure 1: Scanned Compliance Report with Table Structure Example

(a): Scanned Compliance Report



No.	Procedures	Findings
1	Obtained and checked the computation of the monthly minimum capital funds balances computed by management of the Company (as disclosed in **Appendix 1**) from the monthly unaudited trial balance of the Company for the financial year, and compared the capital funds balances to the minimum capital funds requirement as prescribed under the Money Services Business (Minimum Capital Funds) Regulations 2011, Money Services Business (Minimum Capital Funds) (Amendment) Regulations 2012 and Money Services Business (Minimum Capital Funds) (Amendment) Regulations 2013 to report any instance of non-compliance.	With respect to the procedures, there were no exceptions noted.

(b): Extracted Text Data using PyTesseract

```
[No. [Procedures Findings
4 Obtained and checked the computation of the | With respect to
the procedures, there were no
monthly minimum capital funds balances exceptions noted.
computed by management of the Company
(as disclosed in Appendix 1) from the
monthly unaudited trial balance of the
Company for the financial year, and
compared the capital funds balances to the
minimum capital funds requirement as
prescribed under the Money Services
Business (Minimum Capital Funds)
Regulations 2011, Money Services Business
(Minimum Capital Funds) (Amendment)
Regulations 2012 and Money Services
Business (Minimum Capital Funds)
(Amendment) Regulations 2013 to report any
instance of non-compliance.
```

(c): Extracted Text Data using Img2Table with PaddleOCR

No	Procedures	Findings
1	Obtained and checked the computation of the monthly minimum capital funds balances computed by management of the Company (as disclosed in Appendix 1) from the monthly unaudited trial balance of the Company for the financial year, and compared the capital funds balances to the minimum capital funds requirement as prescribed under the Money Services Business (Minimum Capital Funds) Regulations 2011, Money Services Business (Minimum Capital Funds) (Amendment) Regulations 2012 and Money Services Business (Minimum Capital Funds) (Amendment) Regulations 2013 to report any instance of non-compliance.	With respect to the procedures, there were no exceptions noted.

Figure 1(a) displays a scanned document image highlighting the compliance findings in a table structure. Figure 1(b) presents the data extracted using PyTesseract, while Figure 1(c) shows the data extracted using Img2Table with PaddleOCR. While PyTesseract is able to extract the text from the table, it lacks context of the different columns in the table. Whereas Img2Table with PaddleOCR able to extract the table data according to the table format found in the scanned images of the PDF.

3.2 Language Models (LMs) - Summarisation

We then summarise the extracted text data using three LMs namely, BART, T5 and GPT-2. Following this, we further compare the performance of all these LMs using the selected metrics.

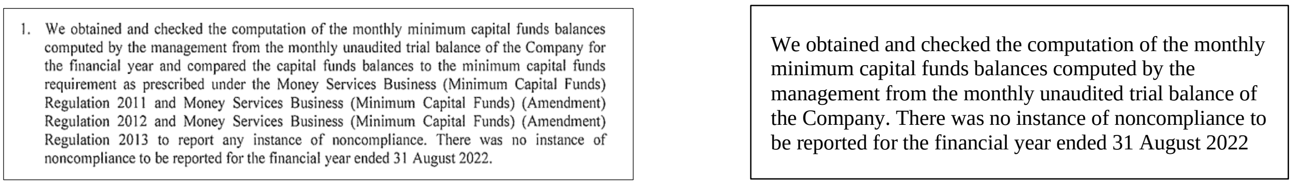
Table 1: Performance Measures Comparison for BART, T5, and GPT-2

Language Models	ROUGE-L	BLEU	BERTScore
BART	**0.7532**	**0.4671**	**0.9362**
T5	0.7331	0.4404	0.9293
GPT-2	0.2238	0.1001	0.7608

Evaluation on these LMs is measured using ROUGE-L, BLEU and BERTScore. These metrics measure the similarity between the extracted text and summarised text. Table 1 shows the performance of models across these metrics. Surprisingly, as shown in Table 1, GPT-2 unexpectedly has the lowest similarity score between the extracted text and summarised text across all metrics, given the popularity of ChatGPT model.

Interestingly, BART and T5 have almost identical similarity scores across all metrics. In line with the findings of Rao et al. (2024), BART achieves the highest similarity score across all metrics, hence chosen as the champion LM for our summarisation task. This allows supervisors to view high-level summaries for better understanding to the regulatees and decision making.

Figure 2 : Extracted Text and Summarised Text from BART model



(a) Extracted Text from Compliance Report

1. We obtained and checked the computation of the monthly minimum capital funds balances computed by the management from the monthly unaudited trial balance of the Company for the financial year and compared the capital funds balances to the minimum capital funds requirement as prescribed under the Money Services Business (Minimum Capital Funds) Regulation 2011 and Money Services Business (Minimum Capital Funds) (Amendment) Regulation 2012 and Money Services Business (Minimum Capital Funds) (Amendment) Regulation 2013 to report any instance of noncompliance. There was no instance of noncompliance to be reported for the financial year ended 31 August 2022.

(b) Summarised Text

We obtained and checked the computation of the monthly minimum capital funds balances computed by the management from the monthly unaudited trial balance of the Company. There was no instance of noncompliance to be reported for the financial year ended 31 August 2022

Figure 2 illustrates that BART effectively generate concise and accurate summaries reflecting its strong performance in matching n-grams with reference summaries. It demonstrates a notable advantage in producing more concise summaries by effectively removing unnecessary words, resulting in more streamlined and readable summaries.

3.3 Compliance analysis

Based on the summarised text of the reports, supervisors can then assess regulatees' compliance based on the areas covered in the reports. To facilitate that, we leverage on regex techniques to identify obvious keywords used by supervisors (e.g. "Exception noted" and "Not met") to evaluate compliance level of the regulatees. We utilise the BART LM for reports without obvious keywords to determine regulatory adherence, specifically focusing on zero-shot classification on the summarised text. By comparing the performance of our compliance check tools with the true results evaluated by humans, we discover that our tool currently achieves an average accuracy of 85%.

4. Discussion and Conclusion:

In this study, BART outperforms other LMs in summarisation and combining different OCR techniques to improves data quality significantly. For reports that are lacking keywords to determine the regulatory adherence, LMs prove to be an effective tool for text analysis. LMs help supervisory authorities enhance their oversight capabilities by offering comprehensive analysis for large volume of reports.

One of the limitations of this study is that we are not leveraging on proprietary OCR tools, like Adobe Acrobat, and Azure Intelligent Document Processing, which generally offer superior performance, but come with associated costs. Furthermore, we do not leverage the multimodal Large LMs (LLMs) that naturally handle multiple data types, due to the current inconsistencies in their output and the ongoing development. For future research, incorporating the proprietary OCR tools could enhance the accuracy of data extraction and leveraging multimodal LLMs could enable simultaneous text extraction and summarisation.

References:

Bank Negara Malaysia. (2011, September 15). Money Services Business Act 2011 (MSBA). *Laws of Malaysia*. Retrieved July 24, 2024, from https://www.bnm.gov.my/documents/20124/815175/en_money_services.pdf.

Bank Negara Malaysia. (2020, April 3). Annual Report 2019. Retrieved July 26, 2024, from https://www.bnm.gov.my/documents/20124/2724769/ar2019_en_book.pdf

Bank Negara Malaysia. (2022, December 30). Electronic Money (E-Money). Retrieved July 24, 2024, from https://www.bnm.gov.my/documents/20124/943361/PD-eMoney-202302.pdf

Chen, Q. (2021, December 14). T5: a detailed explanation - Analytics Vidhya. *Medium*. Retrieved July 25, 2024, from https://medium.com/analytics-vidhya/t5-a-detailed-explanation-a0ac9bc53e51

Czerwinska, U. (2023, March 06). Deep Dive in PaddleOCR inference. *Adevinta Tech Blog*. Retrieved July 24, 2024, from https://medium.com/adevinta-tech-blog/deep-dive-in-paddleocr-inference-e86f618a0937

Di Castri, S., Grasser, M., & Kulenkampff, A. (2018). Financial authorities in the era of data abundance: Regtech for regulators and suptech solutions. *Available at SSRN 3249283*.

Garg, A., Adusumilli, S., Yenneti, S., Badal, T., Garg, D., Pandey, V., ... & Agarwal, R. (2021). NEWS article summarization with pretrained transformer. In *Advanced Computing: 10th International Conference, IACC 2020, Panaji, Goa, India, December 5–6, 2020, Revised Selected Papers, Part I 10* (pp. 203-211). Springer Singapore.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2019, October 29). BART: Denoising Sequence-to-Sequence Pre-training for natural language generation, Translation, and Comprehension. arXiv.org. https://arxiv.org/abs/1910.13461v1

Mansuy, R. (n.d.). Evaluating NLP Models: A comprehensive guide to ROUGE, BLEU, METEOR, and BERTScore metrics. *plainenglish*, Retrieved July 24, 2024, from https://plainenglish.io/community/evaluating-nlp-models-a-comprehensive-guide-to-rouge-bleu-meteor-and-bertscore-metrics-d0f1b1

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9.

Rao, R., Sharma, S., & Malik, N. (2024). Automatic text summarization using transformer-based language models. *International Journal of System Assurance Engineering and Management*, 1-7.

Rosebrock, A. (2022, February 28). Multi-Column Table OCR. *pyimagesearch*. Retrieved July 24, 2024, from https://pyimagesearch.com/2022/02/28/multi-column-table-ocr/

Tuychiev, B. (2024, April). A Comprehensive Tutorial on Optical Character Recognition (OCR) in Python With Pytesseract. *datacamp*. Retrieved July 24, 2024, from https://www.datacamp.com/tutorial/optical-character-recognition-ocr-in-python-with-pytesseract

xavctn. (n.d.). img2table. *github*. Retrieved July 23, 2024, from https://github.com/xavctn/img2table

Zelic, F., & Sable, A. (2023, February 27). OCR using Pytesseract and OpenCV. *Nanonets*. Retrieved July 24, 2024, from https://nanonets.com/blog/ocr-with-tesseract/

Acknowledgements:

This paper has benefited from the critical review and feedback provided by the following individuals.

- Zarifa Izan binti Zainol Abidin, Central Bank of Malaysia, Malaysia
- Mohd. Faizal bin Abdullah, Central Bank of Malaysia, Malaysia
- Vincent Lee Wai Seng, Central Bank of Malaysia, Malaysia
- Ahmad Shazwan bin Mohd Kharip Shah, Central Bank of Malaysia, Malaysia