SULIT

MINISTRY OF ECONOMY
DEPARTMENT OF STATISTICS MALAYSIA

Application of Language Models to Analyse Scanned Textual Compliance Reports*

Chun Onn Pam; Nur Aisyah binti Mohd Nasir; Siti Hajar binti Khairul Anwar
Central Bank of Malaysia
Payment Services Oversight Department

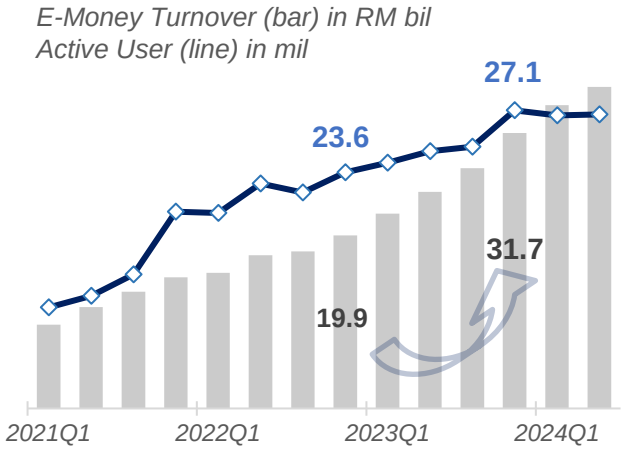
(*) The views and conclusions expressed herein are exclusively those of the author(s) and do not necessarily reflect the position of the Central Bank of Malaysia or of the Board members.

11Th MALAYSIA STATISTICS CONFERENCE
"Data and Artificial Intelligence: Empowering the Future"
19Th September 2024
Organized by:
MINISTRY OF ECONOMY DEPARTMENT OF STATISTICS MALAYSIA
BANK NEGARA MALAYSIA
MALAYSIA INSTITUTE OF STATISTICS
MyStats 2024

Application of language models enhances the utilisation of text-based reports

Supervisory authorities leverage on **Suptech** to support data-driven and timely monitoring for more effective supervision**.** Non-bank payment supervisors in BNM supervise over **50 e-money issuers** (EMI) and **260 money services businesses** (MSB).

The **fund-safeguarding and/or capital adequacy** of the EMIs/MSBs are closely monitored, to ensure the protection of customers' fund and the sustainability of regulated businesses. EMIs and MSBs are expected to operate effectively, as transaction value and number of customers have increased in recent years.



Growing number of large size MSBs

2022: ***83***

2023: ***109***

MSBs and EMIs submit **scanned text-based compliance reports** prepared by external auditors to BNM, consisting of:

- Agreed-Upon Procedures (AUP)
- Fund Management Report (FMR)

Reports are **large** in number, **lengthy** and in **various formats** (imaged-based PDF, table-based report slides)- straining supervisors' resources to analyse this information efficiently.

This paper focuses on the application of:

(i) **Optical Character Recognition (OCR)** to convert the scanned textual reports into machine-readable text**; and**

(ii) **language models (LMs)** and to provide a concise overview of the report contents

This compliance tool is deployed in Q1 2024, currently utilised by supervisors to assess non-compliances detected from the reports.

Extracting and summarising textual reports to monitor compliance with regulatory requirement

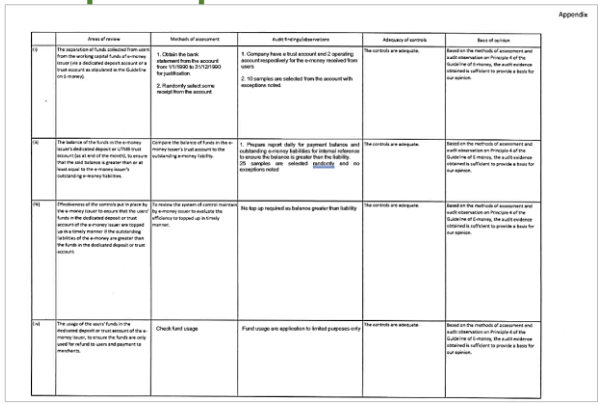
Assessment by auditors

- Separation of funds to Trust account and the usage of funds.
- Trust Account balance and the effectiveness to top up in a timely manner.

Reports

- 123 reports received.
- **10 pages** (min: 1 page, max: 40 pages)

Sample Reports



Appendix

	Areas of review	Methods of assessment	Audit findings/observations	Adequacy of controls	Basis of opinion
(i)	The separation of funds collected from users from the working capital funds of e-money issuer into a dedicated deposit account or a trust account as stipulated in the Guideline on E-money.	1. Obtain the bank statement from the account from 1/1/1900 to 31/12/1900 for justification 2. Randomly select some receipt from the account	1. Company have a trust account and 2 operating account respectively for the e-money received from users 2. 10 samples are selected from the account with exceptions noted.	The controls are adequate.	Based on the methods of assessment and audit observation on Principle 4 of the Guideline of E-money, the audit evidence obtained is sufficient to provide a basis for our opinion.
(ii)	The balance of the funds in the e-money issuer's dedicated deposit or u/Trust trust account (as at end of the month), to ensure that the said balance is greater than or at least equal to the e-money issuer's outstanding e-money liabilities.	Compare the balance of funds in the e-money issuer's trust account to the outstanding e-money liability.	1. Prepare report daily for payment balance and outstanding e-money liabilities for internal reference to ensure the balance is greater than the liability. 25 samples are selected randomly and no exceptions noted	The controls are adequate.	Based on the methods of assessment and audit observation on Principle 4 of the Guideline of E-money, the audit evidence obtained is sufficient to provide a basis for our opinion.
(iii)	Effectiveness of the controls put in place by the e-money issuer to ensure that the users' funds in the dedicated deposit or trust account of the e-money issuer are topped up in a timely manner if the outstanding liabilities of the e-money are greater than the funds in the dedicated deposit or trust account.	To review the system of control maintain by e-money issuer to evaluate the efficiency to topped up in timely manner.	No top up required as balance greater than liability	The controls are adequate.	Based on the methods of assessment and audit observation on Principle 4 of the Guideline of E-money, the audit evidence obtained is sufficient to provide a basis for our opinion.
(iv)	The usage of the users' funds in the dedicated deposit or trust account of the e-money issuer, to ensure the funds are only used for refund to users and payment to merchants.	Check fund usage	Fund usage are application to limited purposes only	The controls are adequate.	Based on the methods of assessment and audit observation on Principle 4 of the Guideline of E-money, the audit evidence obtained is sufficient to provide a basis for our opinion.

Fund Management Report (FMR)

Agreed-Upon Procedures (AUP)

Period:
2020 - 2022

Format include table form, words, with **imaged-based PDF (scanned PDF)**

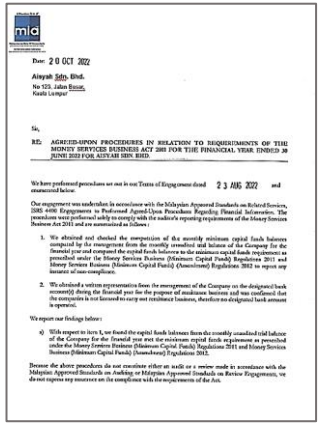
Assessment by auditors

- minimum capital fund, designated account and outstanding customers' monies.

Reports

- 482 reports received.
- **2 pages** (min: 2 pages, max: 14 pages)

Sample Report



mia

Date: 20 OCT 2022

Alayah Sdn. Bhd.
No 123, Jalan Besar,
Kuala Lumpur

Sir,

RE: AGREED-UPON PROCEDURES IN RELATION TO REQUIREMENTS OF THE MONEY SERVICES BUSINESS ACT 2011 FOR THE FINANCIAL YEAR ENDED 30 JUNE 2022 FOR ALAYAH SDN. BHD.

We have performed procedures set out in our Terms of Engagement dated 23 AUG 2022 and enumerated below.

Our engagement was undertaken in accordance with the Malaysian Approved Standards on Related Services, ISRS 4400 Engagements to Performed Agreed-Upon Procedures Regarding Financial Information. The procedures were performed solely to comply with the auditor's reporting requirements of the Money Services Business Act 2011 and are summarized as follows :

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) Regulations 2011 and Money Services Business (Minimum Capital Funds) (Amendment) Regulations 2012 to report any instances of non-compliance.
2. We obtained a written representation from the management of the Company on the designated bank account(s) during the financial year for the purpose of maintenance business and was confirmed that the companies is not licensed to carry out remittance business, therefore no designated bank account is operated.

We report our findings below :

a) With respect to item 1, we found the capital funds balances from the monthly unaudited trial balance of the Company for the financial year met the minimum capital funds requirement as prescribed under the Money Services Business (Minimum Capital Funds) Regulations 2011 and Money Services Business (Minimum Capital Funds) (Amendment) Regulations 2012.

Because the above procedures do not constitute either an audit or a review made in accordance with the Malaysian Approved Standards on Auditing or Malaysian Approved Standards on Review Engagements, we do not express any assurance on the compliance with the requirements of the Act.

Employing OCR to extract image-based reports, and language models for content summarisation

Detailed Methodology

Key steps

Text Extraction

Using **optical character recognition (OCR)** for text extraction from scanned documents.

Text Summarisation

Using **language models (LMs)** to summarise content of the report for its findings.

Performance Evaluation

Evaluate Language Models' performance by comparing the similarity of meaning between the original text and the summarised text.

Detailed steps

…Python library, models and metrics used

Pre-processing
- ✓ PyMuPDF
- ✓ Pillow

Text Extraction
- ✓ PyTesseract
- ✓ Img2table
- ✓ PaddleOCR

Application of 3 LMs were compared for summarisation:

- **T5**
- **BART**
- **GPT-2**

Performance metrics for sentences similarity were evaluated:

- **ROUGE – L**
- **BLEU**
- **BERTScore**

…challenges faced, or highlights

- × Table inside a table or complex table.
- × Small/unrecognizable characters.
- × Inconsistent text format.

- × Inaccurate summarisation due to limited inputs allowed.

- ✓ RegEx is applied to check on the compliance; or
- ✓ NLP classification model is applied for reports lacking keywords to determine compliance status.

Text Extraction from table: Img2Table & PaddleOCR outperforms PyTesseract

Text Extraction (OCR)

PyTesseract, Img2table and PaddleOCR are used to accommodate different report formats.

- Accurate extraction of text paragraph from scanned document.
- Manage to extract tables in proper table format.
- Able to analyse data offline on local devices without sending through API-based engine.

Scanned Compliance Report using different OCR models:

(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 from table 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 from table 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.

Text Summarisation: BART performed the best among three LMs

Performance Evaluation

Evaluation on the 3 LMs is measured using ROUGE - L, BLEU and BERTScore.

BART achieves highest similarity score across all metrics.

- Generate concise and accurate summary.
- Content of the summaries are similar to the original text.

BART result:

(a) Extracted Text from Compliance Report

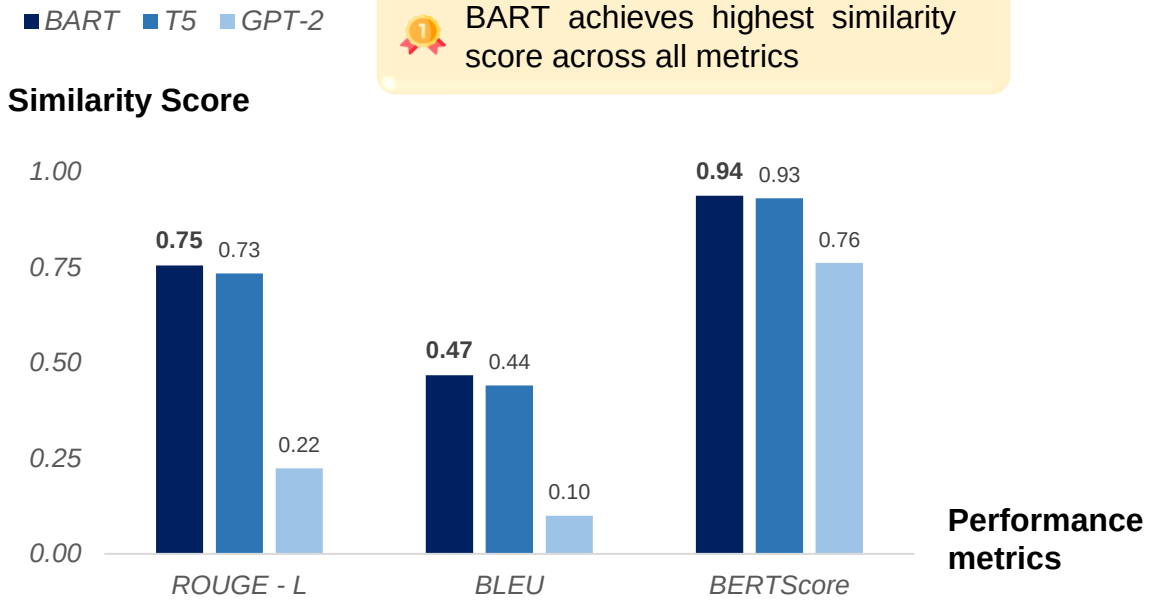
> 1. We obtained and checked the computation of the monthly minimum capital funds balances computed by the management from the monthly unaudited trial balance of the Company for the financial year and compared the capital funds balances to the minimum capital funds requirement as prescribed under the 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

Performance Evaluation on Text Summarisation Model

BART achieves highest similarity score across all metrics



Text analysis output is visualised in dashboard for supervisors' monitoring

Enhanced supervisor efficiency: Reduce need for manual reading of reports.

Facilitates decision-making: Provides quick summaries of past submissions for a holistic view of compliance history of specific entity.

Information extracted from methodology → *Visualised in*

OCR for text extraction

- ✓ **Name of auditors.**
- ✓ **Content of the reports.**

Regex for word search

The keywords such as ***'Exception noted', 'Satisfactory', 'met',*** *and* ***'not met'*** *and others are located for compliance checking.*

- ✓ **Number of regulatees meeting and/or breaching the requirement.**

LLM for text summarisation and classification

- ✓ **Summary of requirement breach, classified by specific act / compliance.**

- **AUP Monitoring Dashboard***
- **FMR Monitoring Dashboard***

Example outcome: Several regulatees which did not meet capital adequacy requirement were imposed monetary penalty.

** The dashboards are not displayed due to privacy and confidentiality.*

Application of language models enhances the utilisation of text-based reports

Summarisation accuracy matters

- BART outperforms other language models in text summarisation.

Flexible in approach

- Combining different OCR techniques for different reporting structures enhance data extraction and improve compliance checking.
- Language models (LMs) serve as an effective tool to help detect compliance, for reports lacking keywords to determine regulatory adherence.

Proprietary OCR Tools

- Paid services like Adobe Acrobat and Azure Intelligent Document Processing offer superior performance.

Multimodal Large Language Models

- Direct image pre-processing for NLP tasks like text summarisation and classification.

Report Standardisation

- Enhancing data extraction by imposing standardised reporting structure.
- Standardised and clearer policy assessment for auditors on compliance.

Thank you

Appendix

	RegEx (Regular Expression)	**Languange Models** (T5, BART, and GPT-2)	**Performance Metrics** (ROUGE – L , BLEU, BERTScore)
What is it?	• Regex is a tool for defining patterns **to search for certain characters or words** inside text-based content • Allows to process text and make **locating information much easier** • Aside from literal characters (like 'abc'), Regex can also search other specifial characters (*,+,?,@ and so on) which may be relevant in the report	• LMs are **pre-trained on a vast amount of text data to generate human-like responses** or perform language-related tasks, including summarisation or classification • An on-premises LM AI model allow LM to **operate within our own infrastructure**, rather than in the cloud ensuring no confidential data shared to other entity. • LMs **are transformer[1]-based models** capable of performing language tasks – ○ **T5 or Text-to-Text Transfer Transformer (Encoder-Decoder),** developed by **Google**, converts all tasks into a text format, making it versatile for various text-related tasks. ○ **BART (Encoder-Decoder)**, developed by **Meta**, corrupts text with noise and learns to reconstruct the original text, making it effective for text generation and summarisation. ○ **GPT-2 (Decoder-Only)**, developed by **OpenAI,** generates relevant and coherent text by predicting the next word sequence.	• Performance metrics are used to calculate the **similarity score** to evaluate the performance of the summarisation models • ROUGE – L formula: $ROUGE\text{–}L, F1-score = \frac{2 \times (precision \times recall)}{precision + recall}$ • BLEU formula: $\text{BLEU} = brevity\_penalty \times e^{\sum_{n=1}^{N} w_n \log(p_n)}$ • BERTScore formula: $Similarity\ (A,B) = \cos(\theta) = \frac{A \cdot B}{\lVert A \rVert \lVert B \rVert}$
How we use it?	• We locate the keywords such as **'Exception noted', 'Satisfactory', 'met',** and **'not met'** and others for compliance checking.	• We use 3 LMs to **summarize text** and identify the one with the highest score as our champion LM. • We use BART to **classify** the non-regex cases.(R*eports that do not have the specific word that regex can detect*)	• We calculate the average of the similarity for each performance metrics for all reports

[1] Transformers use an attention mechanism to understand the context and dependencies between words in a sentence. This mechanism allows the model to focus on different parts of the input sequence when generating the output sequence.