MINISTRY OF ECONOMY
DEPARTMENT OF STATISTICS MALAYSIA

Enabling Enhanced Analytics and AI Applications Through Effective Big Data Processing

Kai Boon Chang[1*]; Mohamad Akmal Izzuddin[2*]

[1]People Department, Bank Negara Malaysia, Kuala Lumpur
[2]Business Technology Department , Bank Negara Malaysia, Kuala Lumpur

*Work completed during rotation in the Financial Intelligence and Enforcement Department

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
CENTRAL BANK OF MALAYSIA
MALAYSIA INSTITUTE OF STATISTICS

MyStats 2024

1.0 Introduction

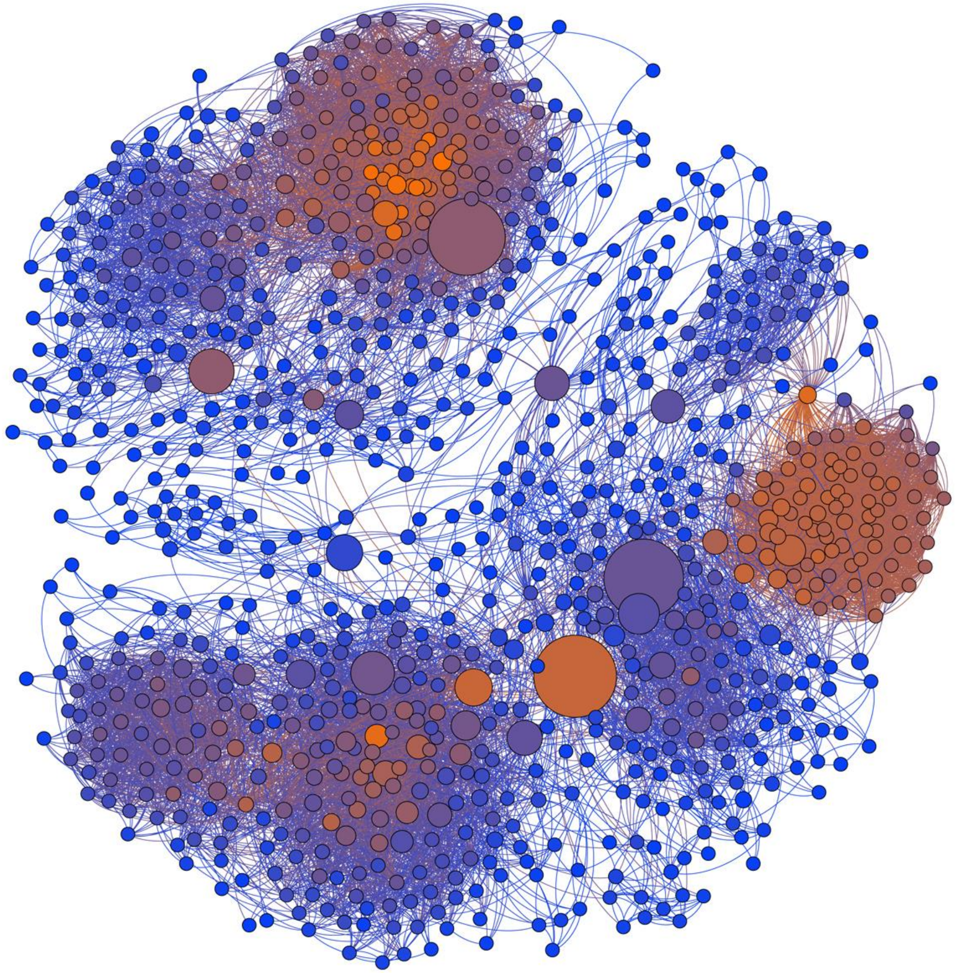
- Integrating AI and big data involves harmonizing technologies from both fields to effectively manage, analyse, and extract insights from large and complex datasets.

- In the realm of big data, harmonization is a crucial preprocessing step that facilitates AI applications and downstream tasks.

- This paper explores the methodologies and findings from a comprehensive street address harmonization project involving 115 million records.

- The objective is to group similar addresses linked to different Suspicious Transaction Reports (STR) to uncover hidden relationships between entities across various reports.

- This process is particularly valuable for fraud detection, law enforcement and market analysis as it helps to identify potential linkages and patterns within vast dataset.



2.0 Methodology

2.1 Data Cleaning and Standardisation

- Expanded common abbreviations to their full forms to maintain uniformity.
- Compiled regular expressions to handle various common patterns in addresses such as separating digits followed by specific keywords, concatenating digits separated by special characters, removing postcodes embedded within addresses and etc..

2.2 Address Normalisation

- Town names embedded within the street addresses were removed to avoid redundancy. Next, Unit numbers were extracted and cleaned.
- Furthermore, postcodes were cleaned and validated against a predefined list of valid postcodes by Pos Malaysia to ensure accuracy.
- Lastly, the addresses were split into primary and secondary components to facilitate more precise comparison.

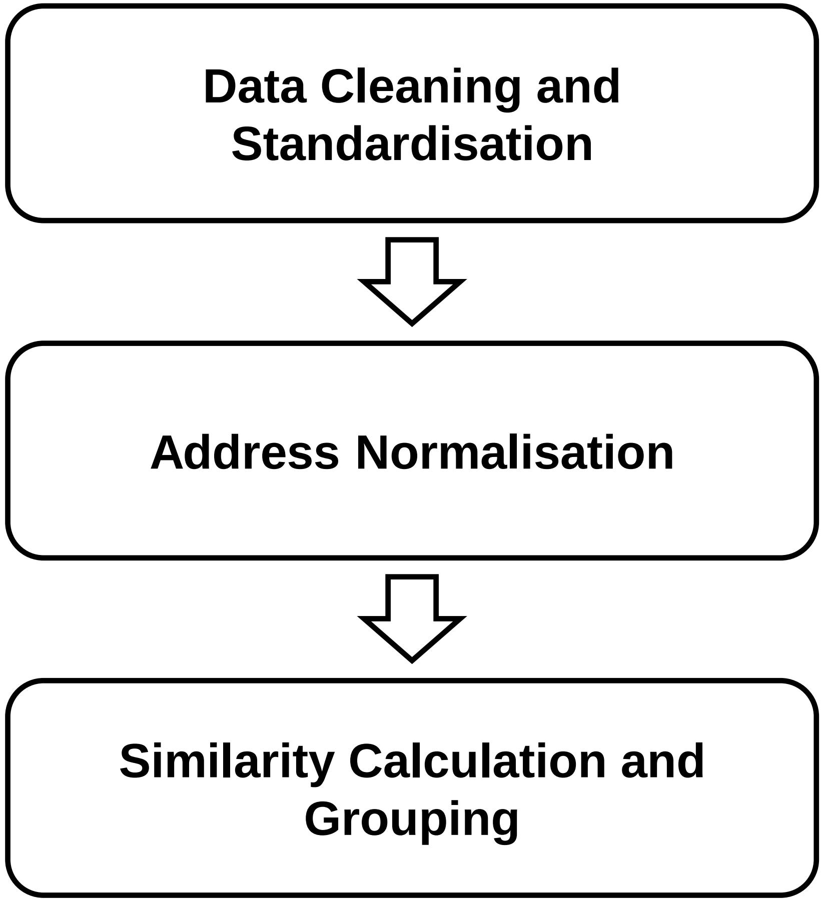

Data Cleaning and Standardisation
Address Normalisation
Similarity Calculation and Grouping

Figure 1: General framework of our harmonization process

2.0 Methodology

2.3 Similarity Calculation and grouping

- Addresses are tokenized and sorted alphabetically to mitigate issues by variations in token order. A similarity score is then computed between two addresses.]

- Similarity Score = $\left(1 - \frac{Levenshtein\ Distance}{Maximum\ Length\ of\ the\ strings}\right) x\ 100$

- The addresses with a similarity score above a specified threshold (**80%** in this study) are considered similar and grouped together.

Example A:

- String 1: Jalan Imbi
- String 2: Jalan Istana
- Score = $\left(1\text{-}\ \frac{5}{12}\right)$ x 100 = **58.3**

Example B:

- String 1: Jalan Hang Tuah
- String 2: Jalan Hang Jebat
- Score = $\left(1\text{-}\ \frac{4}{16}\right)$ x 100 = **75.0**

2.4 Levenshtein Distance

- Levenshtein Distance measures the minimum number of single-character edits required to transform one string into another.
- The edits can include Insertion, deletions and substitutions.

$$d(i,j) = \min \begin{cases} d(i-1,j)+1 \\ d(i,j-1)+1 \\ d(i-1,j-1)+1_{(i \neq j)} \end{cases}$$

- For example, to transform “kitten” into “sitting” we need three edits (distance):

1. Substitute “k” with “s” -> “sitten”
2. Substitute “e” with “i” -> “sittin”
3. Insert “g” at the end -> “sitting”

3.0 Result

- We conducted the analysis on a private dataset comprising of 115 million records. The performance of the address harmonisation process is evaluated based on the similarity scores between the addresses.

- Table 1 shows that 97.5% of the addresses have scores ≥ 80, indicating a high level of similarity and successful harmonisation. On the other hand, 2.5% of addresses have score more than 75 but less than 80, suggesting a moderate level of similarity. Notably, no addresses scored below 75, meaning no addresses were classified with a low similarity score.

- These results demonstrate that our methodologies are highly effective, implying that most of the addresses are considered similar enough to be grouped together — a positive outcome for the address harmonisation process.

Category	Number of records	Percentage (%)
Score ≥ 80	112626013	97.5%
75 ≤ Score ≤ 80	2829900	2.5%
Score < 75	0	0
Total	**115455913**	**100.0%**

Table 1: Similarity score of the addresses

3.0 Result

- Table 2 shows addresses grouped into clusters based on similarity. Each group represents related addresses, demonstrating the effectiveness of the harmonisation process.

- For example, in grouping 9, "Jalan Chan Sow Lin" and "Jalan Chan Sew Lin" were grouped together, while "Jalan SS 242/417 Sri Jaya" and "Jalan SS 242/455 Sri Jaya" were not. This ability to identify and group similar addresses is crucial for uncovering hidden relationships and enhancing data analysis accuracy.

ENTITY ID	Unit Number	Address	Postcode	Grouping
1004	17	Jalan SS 242/417 Sri Jaya	47300	2
1005	17	Jalan SS 242/417 Sri Jaya	47300	2
1008	27	Jalan SS 242/455 Sri Jaya	47300	4
1014	4	Jalan Chan Sow Lin	81800	9
1017	4	Jalan Chan Sew Lin	81800	9
1019	33a	Jalan Segambur 3A Bandar Puteri	41400	11
1020	6a	Jalan Sultan Ismail 1B	41400	12
1021	33a	Jalan Segambur 3A Bandar Puteri	41400	11

Table 2: Grouping of similar addresses (Mock data)

4.0 Discussion

4.1 Benefits and challenges

- Our study successfully demonstrates an effective methodology which ensures data quality and consistency, allowing more reliable insights in advanced data analysis and AI model training.

- We also faced challenges particularly in managing the computational complexity of processing vast dataset. The challenges are consistent with those observed in other domains, such as clinical harmonisation.

Challenge	Proposed Solution
High computational complexity	Optimised algorithms, parallel processing
Data format variability	Regular expression refinement
Large memory requirement	High-RAM servers, efficient data handling
Handling inconsistent data	Data cleaning and standardization procedures

Table 3 : Challenges and solutions

4.2 Recommendations for future research

- Research should aim to enhance computational efficiency for handling complex datasets without compromising performance.

- It is crucial to address the ethical implications of big data processing and AI integration to ensure responsible and transparent technology use.

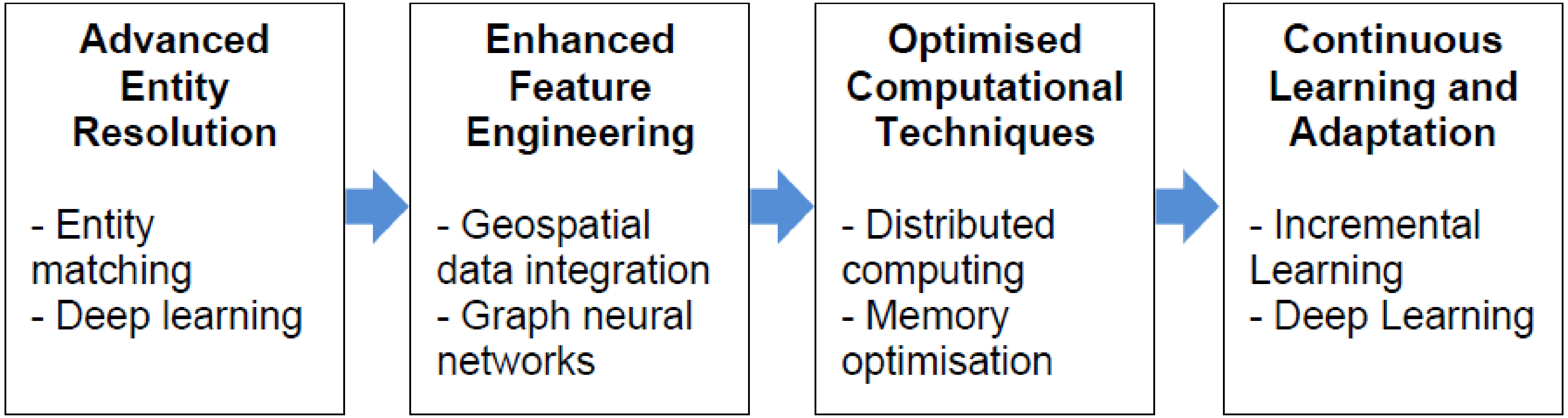
- Collaborative efforts between academia, industry and regulatory bodies can foster the development of best practices and standards for the field.

4.0 Discussion

4.3 Impact of the study

- Our findings provide a robust foundation and can serve as a benchmark for future studies and real-world applications, guiding the development of more advanced and efficient processing methods.

- Our study also underscores the importance of high-quality data in AI applications, highlighting the sequential relationship between big data and AI. By ensuring the quality of input data, we can significantly improve the performance and reliability of AI models, ultimately driving more accurate and actionable insights.



Advanced Entity Resolution

- Entity matching
- Deep learning

Enhanced Feature Engineering

- Geospatial data integration
- Graph neural networks

Optimised Computational Techniques

- Distributed computing
- Memory optimisation

Continuous Learning and Adaptation

- Incremental Learning
- Deep Learning

Figure 2 : Potential advanced workflow diagram for harmonisation process

Conclusion

- Our study demonstrates the significant potential of harmonization in preparing high-quality data for AI applications.

- While challenges remain, particularly in managing computational complexity, the benefit of having robust processing methods are clear.

- Future research and applications should focus on refining existing techniques, exploring advanced algorithms, and addressing ethical considerations.

- By harmonising big data and AI, we can unlock new opportunities for accurate predictions, efficient operations and ultimately driving technological advancement and societal progress.

Thank you

Kai Boon Chang
kaiboon.chang@bnm.gov.my

Mohamad Akmal Izzuddin Bin Mohamad Yazid
akmal.yazid@bnm.gov.my