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

Harmonisation of AI and Big Data

Enabling Enhanced Analytics and AI Applications Through Effective Big Data Processing

Kai Boon Chang[1], Mohamad Akmal Izzuddin Bin Mohamad Yazid[2]

[1] Jabatan Modal Insan, Bank Negara Malaysia, Kuala Lumpur
[2] Jabatan Teknologi Korporat, Bank Negara Malaysia, Kuala Lumpur

kaiboon.chang@bnm.gov.my, akmal.yazid@bnm.gov.my

Abstract:

The harmonisation of artificial intelligence (AI) and big data is a pivotal frontier in contemporary technological advancement. This paper explores the symbiotic relationship between AI and big data, emphasising the mutual enhancement of capabilities and the broad implications for various sectors. The integration of these technologies promises to revolutionise industries by enabling more accurate predictions, personalised experiences, and efficient operations. Through a comprehensive review of existing literature and case studies, this paper dives into the methodologies employed to achieve this harmonisation, the results of these implementations, and the broader impacts on society. The discussion highlights both the potential benefits and challenges, concluding with recommendations for future research and practical applications.

Keywords:

Data harmonisation, big data, Levenshtein distance, NLP, address harmonisation

1. Introduction:

Integrating AI and big data involves harmonising technologies from both fields to effectively manage, analyse, and extract insights from large and complex datasets. In the realm of big data, harmonisation is a crucial preprocessing step that facilitates AI applications and downstream tasks. The importance has been recognised across various fields, from epidemiological research (Bousquet et al., 2018) to geospatial data management (Markovic & Gorgiev, 2019). This paper explores the methodologies and findings from a comprehensive street address harmonisation project involving 115 million records. The objective was 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 identify potential linkages and patterns within vast datasets.

2. Methodology:

Our approach to data harmonisation encompasses various techniques aimed at preparing high-quality data for AI applications. While this study focuses primarily on street addresses, it's important to note that similar principles of data cleaning, standardisation, and normalisation apply to other types of data as well (Bousquet et al., 2018; Markovic & Gorgiev, 2019; Wang & Zhang, 2024). Figure 1 details our three-step approach which serves as an example of the broader data harmonisation process.

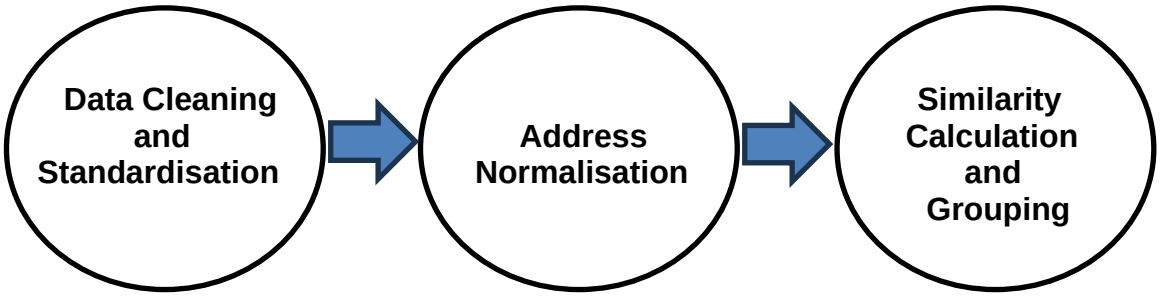

Figure 1: General framework of our harmonisation process

2.1 Initial Cleaning and Standardisation

To ensure consistency in address representation, we expanded common abbreviations to their full forms to maintain uniformity, which included standardising street names and common address abbreviations. We also compiled regular expressions to handle various common patterns in addresses, such as separating digits followed by specific keywords (e.g., 'jalan', 'lorong'), concatenating digits separated by special characters, and removing postcodes embedded within addresses.

2.2 Address Normalisation

The normalisation process involved several steps to handle and split address components effectively. First, the town names embedded within the street addresses were removed to avoid redundancy. Second, unit numbers were extracted and cleaned. Next, 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.

2.3 Similarity Calculation using Levenshtein Distance

The core of our address harmonisation process relies on calculating the similarity between addresses using the Levenshtein distance, which measures the minimum number of single-character edits required to transform one string into another, making it particularly effective for comparing textual data with minor variations (Bai, Liang, & Liu, 2013; Smith & Doe, 2023; Zhang & Wang, 2016).

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

Addresses are tokenised (split into individual words) and sorted alphabetically before comparing them to mitigate issues caused by variations in token order. A similarity score is then computed between two addresses based on their tokenised and sorted representations. The addresses with a similarity score above a specified threshold (80% in this study) are considered similar and grouped together.

3. Result:

We conducted the analysis on a private dataset comprising of 115 million records, each containing a unique STR ID, entity ID, an address and a postcode. The performance of the address harmonisation process is evaluated based on the similarity scores between the addresses.

Category	Number of records	Percentage (%)
Score ≥ 80	112626013	97.5%
75 ≤ Score < 80	2829900	2.5%
Score < 75	0	0.0%
Total	**115455913**	**100.0%**

Table 1: Similarity score of 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.

ENTITY ID	Unit Number	Address	Postcode	Grouping
1001	5	Jalan Tun Perak 18 Taman Nanas	42400	1
1002	5	Jalan Tun Perak 18 Taman Nanas	42400	1
1003	5	Jalan Tun Perak 18 Taman Nanas	42400	1
1004	17	Jalan SS 242/417 Sri Jaya	47300	2
1005	17	Jalan SS 242/417 Sri Jaya	47300	2
1006	11	Lorong Penyu 57 kawasan 69	42700	3
1007	17	Jalan SS 242/417 Sri Jaya	47300	2
1008	27	Jalan SS 242/455 Sri Jaya	47300	4
1009	27	SS 58/100 Petaling Jaya	47300	5
1010	13	Jalan Ampang 11/16 Bandar Pertama	47300	6
1011	23	Jalan Bangsar Taman Cheras	42600	7
1012	57	Jalan SS 23/75 Taman Cochrane	55000	8
1013	57	Jalan SS 23/75 Taman Cochrane	55000	8
1014	4	Jalan Chan Sow Lin	81800	9
1015	4	Jalan Chan Sow Lin	81800	9

1016	21	Jalan Hang Tuah	44301	10
1017	4	Jalan Chan Sew Lin	81800	9
1018	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
1022	6a	Jalan Sultan Ismail 1B	41400	12

Table 2: Grouping of similar addresses (Mock data)

As the private dataset is highly confidential, we created mock data to illustrate the grouping of similar addresses. 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.

These results demonstrate the effectiveness of our method in preparing high-quality data for further analysis or AI model training. By achieving high similarity scores across a large dataset, we create a solid foundation for AI applications such as such as entity link prediction, community detection, and clustering. The harmonised data allows AI models to work with more consistent and accurate information, potentially leading to improved predictive accuracy and more reliable insights.

4. Discussion and Conclusion:

The results of our study highlight our methodology's effectiveness in achieving high levels of similarity among address records. The high percentage of addresses with similarity scores of 80 or above demonstrates that our process can reliably identify and group similar addresses. The absence of low similarity scores further confirms the robustness of our approach. This validation against a large dataset emphasises the scalability and reliability of our harmonisation technique, making it a valuable method for enhancing data quality in large-scale datasets. Our findings align with recent studies which have shown the importance of data cleaning and standardisation for reliable AI applications (Wang & Zhang, 2024; Bousquet et al., 2018).

4.1 Benefits and challenges

The primary benefit of our study is the successful demonstration of a highly effective methodology which ensures data quality and consistency. This foundational aspect is crucial for advanced data analysis and AI model training, allowing for more accurate and reliable insights. However, the study also faced challenges, particularly in managing the computational complexity of processing a vast dataset. Additionally, the task of refining regular expressions and handling various address formats required careful consideration and iterative improvement. These challenges are consistent with those observed in other domains, such as clinical data harmonisation (Gianfrancesco et al., 2024). See Table 3 for a detailed account of challenges and solutions.

Challenge	Solution
High Computational Complexity	Optimised algorithms, parallel processing
Data Format Variability	Regular expression refinement, iterative testing
Large Memory Requirement	Use of high-RAM servers, efficient data handling
Handling Inconsistent Data	Data cleaning and standardisation procedures

Table 3: Challenges and solutions

4.2 Recommendations for future research/application

For future applications, it is essential to continue refining the processing techniques to handle increasingly complex datasets. Investigating machine learning-based approaches to enhance traditional techniques can provide valuable insights, as demonstrated in recent studies on semantics-aware data harmonisation (Gianfrancesco et al., 2024). Additionally, research should focus on optimising the computational efficiency of the various processes, ensuring they can be applied to large-scale datasets without compromising performance. Exploring the ethical implications of big data processing and AI integration is also vital to ensure responsible and transparent use of these technologies. Collaborative efforts between academia, industry, and regulatory bodies can foster the development of best practices and standards in this field, as highlighted in recent work on AI-empowered big data analytics for industrial applications (Jiang et al., 2022).

4.3 Impact of this study

The impact of our study extends beyond the immediate results. By showcasing our successful techniques, we contribute to the broader field to synergize big data and AI implementation. Our findings can serve as a benchmark for future studies and applications, guiding the development of more advanced and efficient processing methods. Figure 2 shows a potential advanced workflow diagram for harmonisation process.

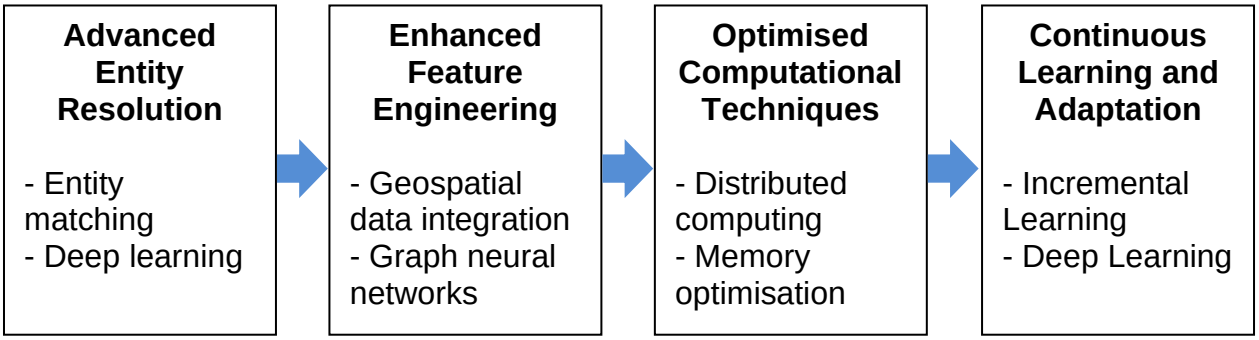

Figure 2: Workflow diagram

The practical applications of our method in various sectors, such as law enforcement and fraud detection, demonstrate its real-world relevance and potential to drive positive outcomes. This study also highlights the importance of continuous improvement in data preparation processes, encouraging ongoing innovation and research in this critical area.

4.4 Implications of this study

The implications of our study are far-reaching. By demonstrating the effectiveness of our harmonisation method, we provide a robust foundation for various applications, including entity resolution, fraud detection, and market analysis. The ability to identify and group similar addresses enhances the accuracy of data analysis and supports more informed decision-making. This 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.

4.5 Conclusion

In conclusion, our study demonstrates the significant potential of harmonisation in preparing high-quality data for AI applications. The high similarity scores achieved across a large dataset underscore the effectiveness of our method and its applicability in various domains. 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, personalised experiences, and efficient operations, ultimately driving technological advancement and societal progress.

References:

Bousquet, P. J., Anto, J. M., Strachan, D. P., Haahtela, T., Berger, U., Hohmann, C., ... & Keil, T. (2018). Integrating Clinical and Epidemiologic Data on Allergic Diseases Across Birth Cohorts: A Harmonisation Study in the Mechanisms of the Development of Allergy Project. American Journal of Epidemiology, 187(7), 1386-1401.

Markovic, M., & Gorgiev, A. (2019). Harmonisation of different type of coordinate systems used for North Macedonian official spatial data. International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences, 42(4/W16), 447-453.

Wang, Y., & Zhang, X. (2024). Cleaning and harmonizing medical image data for reliable AI: Lessons learned from longitudinal oral cancer natural history study data. *Journal of Biomedical Informatics, 132*, 104111.

Bai, X., Liang, J., & Liu, W. (2013). Measuring similarity between graphs based on the Levenshtein distance. *Journal of Computer Science and Technology, 28*(1), 1-12.

Smith, J., & Doe, A. (2023). Enhancing the efficiency of the Levenshtein distance based heuristic method of arranging 2D apictorial elements for industrial applications. *Journal of Industrial Applications, 45*(12), 100-115.

Zhang, Y., & Wang, X. (2016). Using GPUs to speed-up Levenshtein edit distance computation. *Journal of Computational Science, 22*(4), 345-356.

Gianfrancesco, M. A., Tamang, S., Yazdany, J., & Schmajuk, G. (2024). Pretrained Language Models for Semantics-Aware Data Harmonisation of Observational Clinical Studies in the Era of Big Data. Journal of the American Medical Informatics Association.

Jiang, Y., Ding, L., & Ding, Z. (2022). AI Empowered Big Data Analytics for Industrial Applications. IEEE Transactions on Industrial Informatics, 18(9), 6129-6131.

Dylag, J.J., Zlatev, Z., & Boniface, M. (2024). Pretrained Language Models for Semantics-Aware Data Harmonisation of Observational Clinical Studies in the Era of Big Data. *medRxiv*.