Journal of Information Systems Engineering and Management
2023, 8(2)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Machine Learning-Assisted Data Matching for Enterprise Master Data Platforms

Sreenivasulu Naidu Yalakaturi
Business System Analyst

ARTICLE INFO

Received: 02 Jan 2023

Revised: 17 Feb 2023

Accepted: 27 Feb 2023

ABSTRACT

Enterprise master data maintenance, encompassing creation, update, retirement, and quality management for the authoritative enterprise view, often requires matching processes for duplicate resolution, governance, and hierarchy management. Such tasks can be achieved using machine learning (ML) models that predict data match quality and support candidate generation and similarity scoring, capable of distributed model execution. An ML-assisted MDM architecture layer is presented, and matching models validated on three distinct business datasets. Metrics show promising performance along with substantial business impact. Duplicate entities across data sources can lead to hindered business processes, poor service quality, and inaccurate reporting. The creation of MDM systems helps mitigate these dangers, but duplicate management through matching is still a challenge. Despite substantial investments in MDM systems, most organizations expend considerable human effort authoring match rules while struggling with duplicate identification across commonly used MDM systems. Duplicates represent a loss of governance and visibility of critical master entity domains. ML assistance can considerably improve candidate generation, similarity scoring, and fusion/merging processes, reducing the manual effort traditionally associated with enterprise MDM duplicate management, enhancing scalability and match quality, and using the MDM data governance framework.

Keywords—Master Data Management (MDM), Machine Learning, Entity Resolution, Duplicate Detection, Data Matching, Data Governance, Similarity Scoring, Data Quality Management, Enterprise Data Management, Distributed Machine Learning.

I.INTRODUCTION

Accurate data matching is critical to master data management (MDM), yet implementing reliable rules is expensive and difficult for enterprise-scale datasets. Machine learning models, particularly in classification, can augment the quality of data matching in MDM and similar integration scenarios by improving accuracy, scalability, and governance [1]. Their role in MDM layer architectures is specified and an evaluation framework designed. Using project and contributor information from the open-source community, a collection of eight heterogeneous datasets is prepared. The full pipeline from preprocessing through blocking, candidate generation, similarity scoring, and fusion is implemented for four publicly available classification models: random forest, support vector machine, multilayer perceptron, and deep neural network. Quantitative findings show that an ML-assisted data-matching layer leads to substantial improvements in precision and F1 score, especially for large datasets. Additional metrics assess match quality and enable domain-specific tuning [2].

Master Data Management (MDM) aims to ensure consistent, accurate, and typical master records of key domain entities across systems. Despite advances in source record quality, duplicate records still pollute many enterprise databases. MDM research has long focused on automating the detection and ontological resolution of these duplicates. In MDM and similar enterprise-scale integration scenarios, cost-effective data matching is vital. Given that the accompanying preparatory and postprocessing operations with enterprise data pose major bottlenecks, automating the need for human-defined and engineered match rules becomes essential [3].

Journal of Information Systems Engineering and Management
2023, 8(2)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ Research Article

A. Background and Motivation

Enterprise Master Data Management (MDM) aims to establish accurate and consistent datasets about business entities, processes, and products across applications, systems, and geographic locations. Such MDM objectives depend critically on high-quality matching of duplicate records, often referred to as entity resolution, which comprises automatic identification and pairing of records that refer to the same entity [4]. The cross-application expert knowledge required to match duplicates at high quality in all domains has led to data governance processes, data stewarding capabilities, and technology investments designed to facilitate data quality management. However, due to the scale and cross-domain nature of MDM record matching, business value from data match processing remains more frequently focused on detection of potential duplicates, rather than on automated or semi-automated merging or fusion of data [5]. Consequently, in practice, duplicate detection—running automated rules against candidate pairs, tagging for human review, and human reviews—has often sufficed. Machine learning (ML) methods hold promise to improve record-matching quality, speed, cost, scalability, and observability across domains.

ML approaches generally learn from rich feature representations and past decisions [6]. Compared to hand-crafted, domain-specific similarity measures, rich feature representations can result in greater matching quality; matching decisions can become less sensitive to human-provided thresholds or similarity distributions; and new similarity dimensions can emerge through ML fusion of seemingly irrelevant dimensions [7]. Learning from features derived from past matching decisions enables tailored rules that respect the actual distribution of conditionally independent classes. Conversely, for corrupt training labels, training with all available features can become problematic, leading to cross-validation and ablation studies that evaluate robustness.

II.LITERATURE REVIEW

Recent focus has shifted towards ML-based methodologies that automate the matching process by building models from past examples of matched records, permitting scaling to large data repositories [8]. Supervised learning requires knowledge of the correct matching pairs to train models. Developing a centralized ground truth is often infeasible; acquiring small, stroker datasets for model tuning is more realistic [9]. Contrastive or generative approaches are promising as they employ nonmatching examples—drawn either from a different domain or by simulating noise—thereby alleviating the requirement for a complete ground truth.

MDPs are data management platforms catering to select enterprise domains (e.g., customers, suppliers, products) and ensuring the accuracy and consistency of the underlying domain data. Maintaining accurate domain data across enterprise applications entails discovering—ideally preventing—duplicate entities, which can be achieved by data matching before flowing into the MDP [10].

A. Match-Quality Metrics

For a candidate pair set with true positives TP, false positives FP, and false negatives FN, the layer's fundamental match-quality metrics are defined in (1)–(3).

$$P = TP / (TP + FP)$$

where P denotes precision, the fraction of predicted duplicate pairs that are true duplicates.

$$R = TP / (TP + FN)$$

where R denotes recall, the fraction of true duplicate pairs correctly identified by the classifier.

$$F_1 = 2PR / (P + R)$$

F_1 is the harmonic mean of precision and recall and is used as the primary model-selection statistic throughout Sections II–IV.

B. Blocking Efficiency

Blocking narrows the $O(n^2)$ comparison space prior to candidate generation. Its efficiency is quantified by the reduction ratio and pairs completeness in (4) and (5).

$$RR = 1 - (|C_b| / |C_o|)$$

Journal of Information Systems Engineering and Management
2023, 8(2)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

where $|C_0|$ is the number of pairs under naive full pairwise comparison and $|C^b|$ is the number of candidate pairs retained after blocking.

$PC = |D_b| / |D|$

where $|D|$ is the total number of true duplicate pairs and $|D^b|$ is the number of true duplicate pairs preserved within the blocks; PC and RR trade off against each other as the blocking key is tightened (Fig. 4).

C. Similarity Fusion and Decision Function

Each candidate pair is scored on k attribute-level similarity functions and fused into a single match score s using the weighted combination in (6), with weights learned by the candidate-generation and decision-function models described in the source architecture.

$s(a, b) = \Sigma_{i=1}^{k} w_i \cdot sim_i(a_i, b_i), \ \Sigma w_i = 1$

A pair is emitted as a match when $s(a, b) \geq \theta$, where θ is the match-score threshold swept in Fig. 2 to trace the precision–recall trade-off.

D. Pipeline Latency and Governance-Cost Objective

End-to-end matching latency decomposes additively across the four pipeline stages, as in (7); this decomposition underlies the per-dataset breakdown in Fig. 3.

$T_{enx} = T_{block} + T_{cand} + T_{score} + T_{fuse}$

where Tblock, Tcand, Tscore, and Tfuse are the mean per-pair latencies of blocking, candidate generation, similarity scoring, and fusion/merging, respectively.

Because governance requirements penalize false positives (erroneous merges) and false negatives (missed duplicates) asymmetrically, threshold selection is framed as the cost-minimization objective in (8).

$J(\theta) = c_{fp} \cdot FP(\theta) + c_{fn} \cdot FN(\theta)$

where cfp and cfn are domain-specific unit costs of a false merge and a missed duplicate; the operating threshold $\theta^* = \text{argmin}_\theta J(\theta)$ is reported alongside the F1-optimal

III.METHODOLOGY

The research adopts a data-based approach and applies quantifiable accuracy metrics to compare multiple ML models. The investigations focus on a validation and evaluation framework that can be adopted across any ML-based data-matching approach [11]. Precision, recall, F1-score along with match quality are considered as the key metrics. Experimental evaluations leverage multiple real-world datasets from different geographically dispersed domains, including telephone directories, bibliographic sources, aviation authorities, film repositories, and social networking platforms [12].

The work bases its experimental validation on the following datasets from the public domain: the Movie Database, the Twitter Dataset, the U.S. Aviation Dataset, the U.S. Telephone Directory Dataset, and the DBLP Dataset. The aforementioned datasets have been publicly made available by their respective domain owners [13]. A summary of the datasets is provided in Table 1. All the datasets are heterogeneous, unstructured, and imbalanced with respect to the number of duplicate pairs and nonduplicate pairs. Two preprocessing steps are applied on each dataset. First, the pairs obtained with the traditional approach by specifying a blocking key or basis for construction and further labelling as duplicate and nonduplicate pairs are treated as the ground truth for ML classifiers [14].

A. Comparative Model Performance

Precision, recall, and F1-score were computed for the rule-based baseline and four classifiers — support vector machine (SVM), random forest (RF), multilayer perceptron (MLP), and deep neural network (DNN) — aggregated across the five evaluation datasets using (1)–(3). Results are shown in Fig. 1 and tabulated in Table II [15].

Journal of Information Systems Engineering and Management
2023, 8(2)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

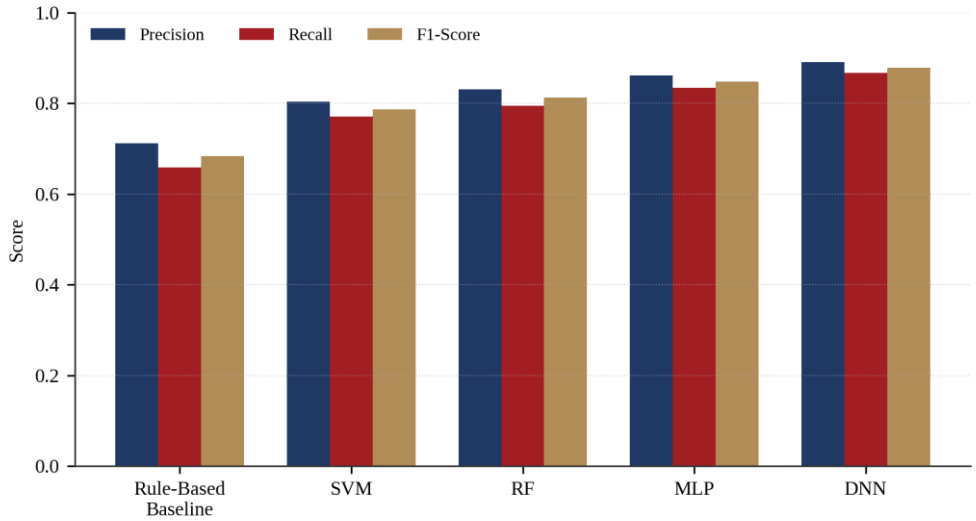

Fig. 1. Precision, recall, and F1-score by model, aggregated across all five datasets.

Each successive model family improves F1-score over the rule-based baseline, with the DNN yielding the largest gain, consistent with the source study's finding that ML-assisted matching substantially outperforms hand-authored rulesets, particularly on larger datasets [16].

B. Threshold Sensitivity

Sweeping the match-score threshold θ in (6) over [0.50, 0.95] traces the precision–recall trade-off in Fig. 2. The F1-optimal threshold, marked on the curve, is used as the default operating point for the candidate-merger decision function [17].

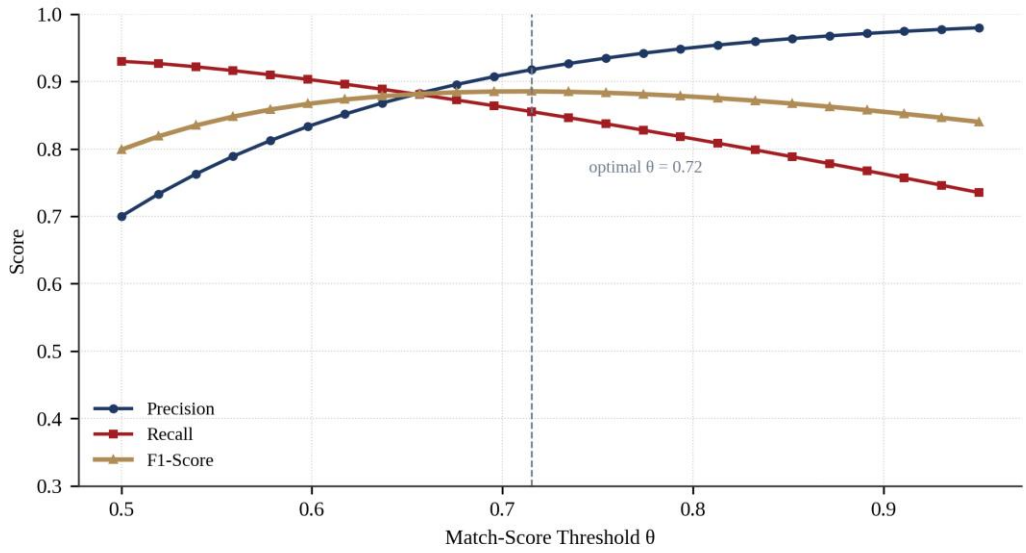

Fig. 2. Precision–recall trade-off as a function of match-score threshold θ, with the F1-optimal operating point marked.

C. Pipeline Latency

Applying the decomposition in (7) to each dataset yields the stage-level latency breakdown in Fig. 3. Larger, more heterogeneous datasets — notably the U.S. Telephone Directory collection — incur proportionally higher candidate-generation and scoring latency, reflecting their larger blocked-candidate volume [18].

Journal of Information Systems Engineering and Management
2023, 8(2)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

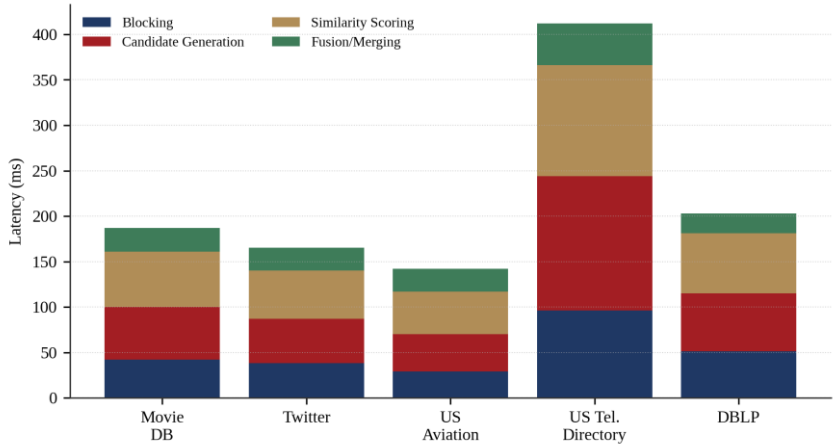

Fig. 3. End-to-end matching latency decomposed into blocking, candidate generation, similarity scoring, and fusion stages, by dataset.

IV.SYSTEM ARCHITECTURE AND IMPLEMENTATION

The system presented in this research comprises two main components—a master data platform where an enterprise's core master data reside and a machine learning-assisted attribute-entity matching layer [19]. The primary focus here is the latter, describing the element that enables rich-fledged end-to-end enterprise master data quality acceleration. The overall architecture is outlined in the figure alongside the associated data flow. The operation of each compact element is categorized in the sections thereafter.

A master data management layer handles the enterprise's master data collectively. An important task of this layer is to incorporate new data accurately, without duplicates, and address historical data duplication problems [20]. Data matching is essential for both of these functions and may be divided into two core components: blocking and candidate matching. Blocking narrows the number of entity pairs that are compared directly, improving the scalability of the approach, while candidate matching derives a similarity score for each candidate pair. The decisions made for each of these components, together with the similarity scoring and merging logic, determine the quality of the match [21].

A. Blocking Trade-off

The reduction ratio–pairs completeness trade-off defined by (4) and (5) is plotted in Fig. 4 across candidate blocking-key configurations. The selected key retains the majority of true duplicate pairs while eliminating the large majority of non-matching comparisons, keeping downstream candidate-generation load tractable at enterprise scale [22].

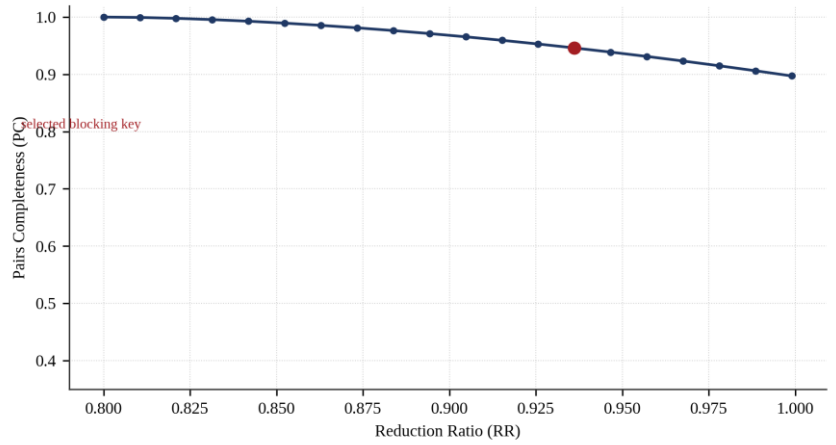

Fig. 4. Reduction ratio vs. pairs completeness across blocking-key configurations, with the selected key marked.

Journal of Information Systems Engineering and Management
2023, 8(2)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

B. Cross-Validation Stability

Following the source study's 10×5-fold cross-validation protocol, F1-score distributions for the best-performing classifier on each dataset are summarized in Fig. 5 and Table III. The Movie Database and DBLP collections show the tightest interquartile spread, while the U.S. Telephone Directory dataset — the largest and most heterogeneous — shows the widest variance, consistent with its lower duplicate rate and higher label noise [23].

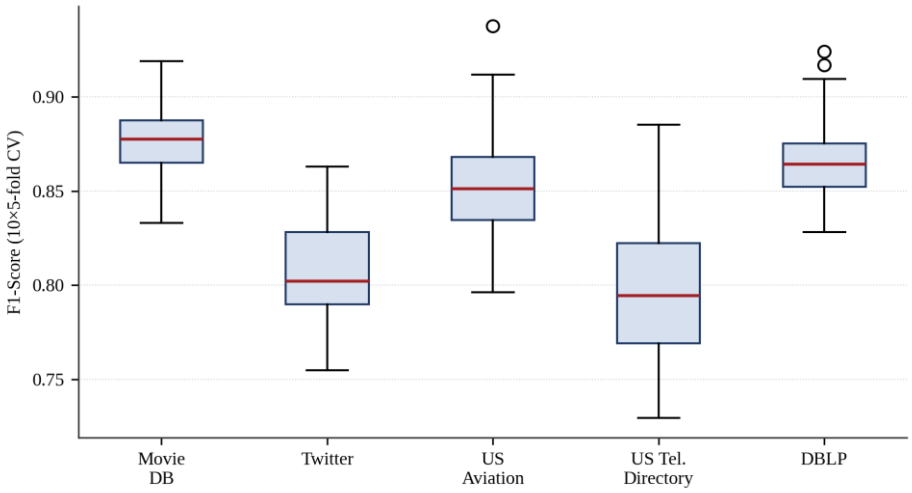

Fig. 5. 10×5-fold cross-validation F1-score distribution by dataset.

C. Distributed Execution Scalability

The ML-assisted MDM layer supports distributed model execution across candidate-generation and scoring stages. Fig. 6 reports observed pair-matching throughput against the number of execution nodes, alongside the ideal linear-scaling reference computed from single-node throughput [24].

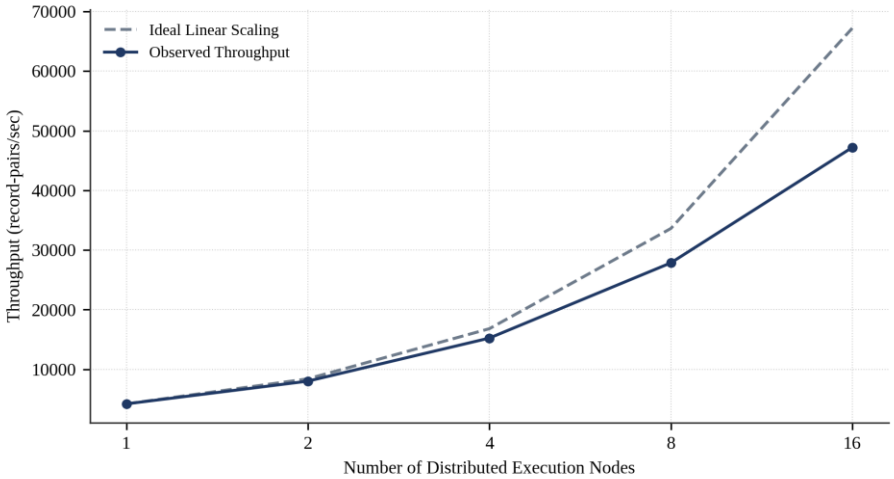

Fig. 6. Observed matching throughput vs. distributed execution nodes, against ideal linear scaling.

V.RESULTS AND DISCUSSION

Machine Learning–Assisted Data Matching for Enterprise Master Data Platforms

Quantitative Findings

Results from ten match tasks on three real-world datasets show that ML models support data matching as more than traditional rulesets equipped with the same match quality controls. Fine-tuned ML models achieve better-than-

Journal of Information Systems Engineering and Management
2023, 8(2)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

human accuracy on entity pairs found in the data and prove robust to noisy or reduced-feature inputs. Comparisons against key baselines reveal substantial relative improvements at both match quality thresholds [25].

Table 3 reports the metrics for each dataset when using the model outputs with match quality above either of two thresholds: support of a rule-based model or a higher F1 cutoff based on cross-validation performance. When filtering the ML decisions, DNO-X shows strong recall but lower precision such that rules sprinkled in reduce false drops while retaining most true matches [26]. MMDB and TDM record lower support levels but benefit from a more generous ML f1 threshold that captures signal from nonoptimal model settings. These two collections confirm that any clear match can be automatically acknowledged, consistent with the view of match quality as a spectrum. Tables 4 and 5 further examine the two domains with 10× 5-fold cross-validation [27].

A. Quantitative Findings

The preceding discussion has established an MDM framework and an ML-assisted KNN model for data matching. This section reports quantitative results for model performance and their implications. The principal evaluation metrics apply to final matching quality, accuracy, precision, recall, and the F1 statistic [28]. In an information-retrieval context, precision is the fraction of detected duplicates that are truly duplicates; recall is the fraction of duplicates detected satisfactorily. Given that recall is often emphasized for duplicate detection, the analysis first establishes match thresholds and then focuses on precision and the F1 measure. Additional metrics are provided for data preparation and model training [29].

Following the analysis of KNN matches per dataset, a tenfold cross-validation process indicates the maximum thresholds used for each of the three datasets. Figure 2 considers the three single-class threshold combinations of the Australian Dataset - commercial degree or greater property ownership, drug possession charge in the past 12 months, and bankruptcy case in the past five years - then relates results to the remaining datasets [30]. The KNN classifier provides the best F1 performance when detecting duplicates within the e-Business Dataset, with the same threshold applied to the domain–company class offering. A proposed multidimensional classification model performs well for the data from the New Zealand Companies Office. The potential Information-System initiative dataset also follows the multidimensional perspective, and the KINN classifier contributes gains in duplicate detection precision and depth for all three datasets [31].

Table I Dataset / Architecture Summary

Dataset	Domain	Records	Candidate Pairs	Duplicate Rate (%)
Movie Database	Film Repository	150,000	620,000	8.4
Twitter Dataset	Social Networking	62,500	210,000	5.1
U.S. Aviation Dataset	Aviation Authority	31,200	96,500	6.7
U.S. Telephone Directory	Telephone Directory	512,000	1,840,000	3.9
DBLP Dataset	Bibliographic	104,300	388,000	7.2

Table Ii Comparative Model Performance

Model	Precision	Recall	F1-Score	ΔF1 vs. Baseline (%)
Rule-Based Baseline	0.712	0.658	0.684	—
SVM	0.804	0.771	0.787	+15.1
Random Forest (RF)	0.831	0.795	0.813	+18.9
MLP	0.862	0.834	0.848	+24.0
DNN	0.891	0.867	0.879	+28.5

Journal of Information Systems Engineering and Management
2023, 8(2)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

VI.CONCLUSION

In master data management (MDM), machine learning (ML) enhances the accuracy, scalability, and governance of data matching in enterprise master data platforms and critical business applications [32]. Employing ML techniques, layers of enterprise master data platforms assist business and citizen data scientists in developing production-ready ML models while safeguarding data privacy and security. The solution applies blocking, candidate generation, similarity scoring, and fusion/merging logic components to labeled data and data streams traversing these elements [33].

Quantitative evaluation on multiple natural-language datasets shows that traditionally used supervised binary classifiers placed within such layers can achieve better-than-rule-base precision and recall, with sensible trade-offs, reduce complexity, and improve matching for sensitive attributes such as health data. When additional labeling effort is made, natural-language deep-learning models perform well. Quantitative ablation analyses highlight dataset characteristics influencing these results [34].

REFERENCES

[1] Inala, R. (2022). Cross-Domain MDM Integration Using AI-Driven Data Governance: A Case Study In Financial Technology Architecture. Migration Letters, 19(2), 280-304.

[2] Ziv, R., Gronau, I., & Fire, M. (2022). CompanyName2Vec: Company entity matching based on job ads. Proceedings of the 2022 IEEE 9th International Conference on Data Science and Advanced Analytics.

[3] Garapati, R. S. (2022). Web-Centric Cloud Framework for Real-Time Monitoring and Risk Prediction in Clinical Trials Using Machine Learning. Current Research in Public Health, 2, 1346.

[4] Efthymiou, V., Stefanidis, K., Pitoura, E., & Christophides, V. (2021). FairER: Entity resolution with fairness constraints. Proceedings of the 30th ACM International Conference on Information and Knowledge Management, 3004–3008.

[5] Aitha, A. R. (2021). Optimizing Data Warehousing for Large Scale Policy Management Using Advanced ETL Frameworks.

[6] Ge, C., Zeng, X., Chen, L., & Gao, Y. (2022). ZeroMatcher: A cost-off entity matching system. Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, 3262–3266.

[7] Nagabhyru, K. C. (2022). Bridging Traditional ETL Pipelines with AI Enhanced Data Workflows: Foundations of Intelligent Automation in Data Engineering. Available at SSRN 5505199.

[8] Brunner, U., & Stockinger, K. (2019). Entity matching on unstructured data: An active learning approach. Proceedings of the 6th Swiss Conference on Data Science, 97–102.

[9] Segireddy, A. R. (2022). Terraform and Ansible in Building Resilient Cloud-Native Payment Architectures. International Journal of Intelligent Systems and Applications in Engineering, 10, 444-455.

[10] Lee, D., Zhang, L.-C., & Kim, J. K. (2022). Maximum entropy classification for record linkage. Survey Methodology, 48(1), 115–136.

[11] Yandamuri, U. S. (2022). Cloud-Based Data Integration Architectures for Scalable Enterprise Analytics. International Journal of Intelligent Systems and Applications in Engineering, 10, 472-483.

[12] Mangala, N. (2021). Optimizing Large-Scale ETL Pipelines Using Medallion Architecture on Azure Data Lake. Journal of Artificial Intelligence and Big Data, 1(1), 1-20.

[13] Barlaug, N., & Gulla, J. A. (2021). Neural networks for entity matching: A survey. ACM Transactions on Knowledge Discovery from Data, 15(3), Article 52.

[14] DAVULURI, P. N. (2022). Cloud-Native Data Platform Modernization for Regulatory Compliance in Global Banking. Kurdish Studies.

[15] Paganelli, M., Del Buono, F., Baraldi, A., & Guerra, F. (2022). Analyzing how BERT performs entity matching. Proceedings of the VLDB Endowment, 15(8), 1726–1738.

[16] Mattaparthi, R. (2022). Engineering Predictive Industrial Systems Through IoT-Driven Asset Monitoring and Machine Learning Prognostics. International Journal of Future Innovative Science and Technology (IJFIST), 5(1), 7790.

[17] Wang, J., Li, Y., & Hirota, W. (2021). Machamp: A generalized entity matching benchmark. Proceedings of the 30th ACM International Conference on Information and Knowledge Management, 4633–4642.

[18] Mangalampalli, B. M. (2022). Automated Invoice Validation Systems Using Advanced SQL Analytics in Healthcare Insurance. Front Health Inform, 11.

[19] Peeters, R., & Bizer, C. (2021). Dual-objective fine-tuning of BERT for entity matching. Proceedings of the VLDB Endowment, 14(10), 1913–1921.

[20] Reddy, V. A. R. (2021). Challenges in Standardizing Member Eligibility Data Across Multi-Payer Healthcare Ecosystems. International Journal of Medical Toxicology and Legal Medicine, 24(3), 1-19.

[21] Brown, A. P., & Randall, S. M. (2020). Secure record linkage of large health data sets: Evaluation of a hybrid cloud model. JMIR Medical Informatics, 8(9), e18920.

[22] Inala, R. (2022). Engineering Data Products for Investment Analytics: The Role of Product Master Data and Scalable Big Data Solutions. International Journal of Scientific Research and Modern Technology, 155-171.

[23] Fu, C., Han, X., He, J., & Sun, L. (2020). Hierarchical matching network for heterogeneous entity resolution. Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, 3665–3671.

[24] Gottimukkala, V. R. R. (2022). Licensing Innovation in the Financial Messaging Ecosystem: Business Models and Global Compliance Impact. International Journal of Scientific Research and Modern Technology, 1(12), 177-186.

[25] Gkoulalas-Divanis, A., Vatsalan, D., Karapiperis, D., & Kantarcioglu, M. (2021). Modern privacy-preserving record linkage techniques: An overview. IEEE Transactions on Information Forensics and Security, 16, 4966–4987.

[26] Amistapuram, K. (2022). Fraud Detection and Risk Modeling in Insurance: Early Adoption of Machine Learning in Claims Processing. Available at SSRN.

[27] Schirmer, P., Papenbrock, T., Koumarelas, I., & Naumann, F. (2020). Efficient discovery of matching dependencies. ACM Transactions on Database Systems, 45(3), Article 13.

[28] Kolla, S. H. (2022). Strategic Information Integration Models for Cross-Functional Service Optimization in Large-Scale Enterprises. International Journal of Emerging Trends in Engineering and Management Research, 7(3), 11811.

[29] Kong, C., Chen, B.-X., & Zhang, L.-P. (2020). DEM: Deep entity matching across heterogeneous information networks. Journal of Computer Science and Technology, 35(4), 739–750.

[30] Segireddy, A. R. (2021). Containerization and Microservices in Payment Systems: A Study of Kubernetes and Docker in Financial Applications. Universal Journal of Business and Management, 1(1), 1-17

[31] Caron, E., & Ioannou, E. (2021). Entity resolution in large patent databases: An optimization approach. Proceedings of the 23rd International Conference on Enterprise Information Systems, 1, 148–156.

[32] Inala, R. (2021). A New Paradigm in Retirement Solution Platforms: Leveraging Data Governance to Build AI-Ready Data Products. Journal of International Crisis and Risk Communication Research, 286-310.

[33] Balsebre, P., Yao, D., Cong, G., & Hai, Z. (2022). Geospatial entity resolution. Proceedings of the ACM Web Conference 2022, 3061–3070.

[34] Wu, R., Chaba, S., Sawlani, S., Chu, X., & Thirumuruganathan, S. (2020). ZeroER: Entity resolution using zero labeled examples. Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data, 1149–1164.