A Multi-Model Framework for Autonomous Schema Discovery and Hybrid Natural Language Generation-Driven Augmented Business Intelligence

Randy Permana[1,*], Sarjon Defit[2], Gunadi Widi Nurcahyo[3]

[1,2,3]*Department of Information Technology, Faculty of Computer Science, Universitas Putra Indonesia YPTK Padang, Indonesia*

(Received: February 20, 2026; Revised: April 25, 2026; Accepted: July 20, 2026; Available online: August 9, 2026)

Abstract

Traditional Business Intelligence (BI) frameworks rely heavily on data professionals to manually map relational metadata into structured star schemas during the ETL process, creating a significant operational bottleneck in data preparation and downstream interpretation of insights. To address these limitations, this study introduces an augmented analytics framework for autonomous schema discovery and hybrid Natural Language Generation (NLG) driven Augmented BI. In the data representation layer, a weighted hybrid feature fusion mechanism combines structural database metadata with contextual text embeddings produced by a pre-trained Sentence Transformer (all-MiniLM-L6-v2). In the multi-model machine learning layer, a multi-paradigm execution engine combines unsupervised geometric clustering models (K-Means, K-Medoids, DBSCAN) and supervised classifiers (SVM, Random Forest), and performance is evaluated using Leave-One-Out Cross-Validation (LOOCV). The resulting schemas are then dynamically projected into an in-memory OLAP cube. At the downstream insight interpretation layer, a Hybrid NLG engine combines a context-aware, rule-based router with an autoregressive generative decoding mechanism to autonomously produce adaptive, actionable business commentaries triggered by user-driven OLAP exploration states. Experimental results demonstrate that applying linear semantic scaling optimization (α) substantially mitigates statistical semantic blindness and protects the framework from structural schema misclassification. On the E-Commerce dataset, the proposed K-Medoids+Semantic configuration demonstrated topological superiority, achieving a peak Silhouette Score of 0.611 and a compressed Davies-Bouldin Index of 0.514. Meanwhile, on the high-dimensional Superstore dataset, the pipeline maintained high functional flexibility, stabilizing overall classification accuracy up to 94.74%. Furthermore, the downstream Hybrid NLG engine (Template+Generative) demonstrated high factual integrity and linguistic flexibility, achieving a ROUGE-1 score of 0.85, a ROUGE-2 score of 0.82, and a BLEU score of 0.15. This research provides ABI frameworks that enable accelerated executive decision-making through seamless, data-to-insight automation.

Keywords: Business Intelligence, Multi-Model Framework, Autonomous Schema Discovery, Natural Language Generation (NLG), Augmented BI.

1. Introduction

In the modern corporate landscape, Business Intelligence (BI) and Data Warehousing (DW) systems serve as the backbone for data-driven strategic decision-making. BI systems serve as the core technological infrastructure for transforming raw transactional data into actionable strategic insights. The first generation of BI development focused on extracting data and generating reports, requiring specialized SQL Scripting and manual database administration. The second generation of BI, or Self-Service BI, was developed to enable end users and business users to interact with prebuilt data visualizations. The latest generation of BI involves Artificial Intelligence (AI) and Machine Learning (ML) to democratize access for non-technical users. This paradigm is called Augmented Analytics, in which the development focus is on automated data preparation and insight generation.

In this study, Augmented Business Intelligence (Augmented BI) refers to a BI framework that autonomously performs three core analytical tasks by automatic schema discovery during the ETL process, automatic multidimensional OLAP construction, and hybrid natural language generation for interpreting analytical results. Rather than evaluating user adoption or usability, this study operationalizes Augmented BI through the automation accuracy of schema discovery and the quality of generated business narratives.

Recent studies demonstrate that ML and AI enable autonomous and adaptive analytical solutions. For example, a heterogeneous ensemble feature selection strategy significantly improves learning accuracy in intelligent decision systems, showing that combining diverse computational methods enhances system robustness [1]. Similarly, machine

***Corresponding author: Randy Permana (randy_permana@upiyptk.ac.id)**
DOI: https://doi.org/10.47738/jads.v7i3.1479

learning is used to detect and classify DoS and DDoS attacks within IoT networks, illustrating how intelligent automation can strengthen complex data-driven environments[2]. Machine learning also automates complex pattern processing tasks, such as reconstructing textile motifs using combined affine segmentation techniques [3].

The evolution of modern BI has spurred extensive research to optimize the initial phases of the data value chain, with particular emphasis on Extract-Transform-Load (ETL) architectures and schema alignment. Traditional rule-based data mapping methods frequently encounter limitations due to structural schema drift and semantic inconsistencies present in heterogeneous corporate data streams. In response, recent studies have investigated the application of Large Language Models (LLMs) for advanced schema inference and automated generation of transformation logic. Frameworks incorporating LLM-enhanced data mapping within enterprise ETL pipelines have exhibited increased resilience and self-optimization by leveraging metadata-aware prompting, domain-specific exemplar retrieval, and iterative self-refinement strategies[4].

Unified and modular schema-matching frameworks have been developed to achieve finer granularity in multi-table enterprise environments by decomposing the integration process into distinct operational stages: schema preparation, table candidate selection, and column-level alignment. This modular structure frequently employs a two-stage optimization strategy, incorporating a Rollup module to consolidate semantically related columns into higher-order concepts and a Drilldown module to re-expand these concepts for precise mapping. In addition, the computational challenge of scaling text-to-SQL engines across large databases while avoiding context window limits has prompted the development of autonomous agent frameworks. The frameworks reformulate schema linking as an iterative, agent-driven process that dynamically explores and expands relevant schema subsets without processing the entire database architecture [5]. Meanwhile, recent developments in Natural Language Generation (NLG) facilitate the creation of narrative text from structured data, improving clarity in analytical systems. Template-driven NLG methods guarantee factual correctness but frequently result in inflexible or monotonous outputs, whereas neural methods generate more coherent language yet can struggle with factual reliability [6].

Hybrid approaches aim to integrate the strengths of both methodologies. However, research focusing on their implementation in BI via automated data-preparation processes remains limited. To address this gap, this study examines the structural and semantic fragmentation in the current automated BI literature by introducing an end-to-end, multi-model framework for autonomous schema discovery and hybrid Natural Language Generation (NLG)-driven augmented analytics. This study advances the implementation of Augmented Business Intelligence (BI) through several key contributions: First, at the data ingestion layer, we develop a hybrid feature fusion mechanism. This mechanism integrates structural metadata profiles with dense contextual text embeddings extracted via Sentence-BERT. By applying a linear semantic scaling optimization, parameterized by alpha, we stretch the computational metric space. This expansion increases semantic distances relative to raw statistical boundaries. As a result, we resolve the challenge of statistical semantic blindness and enable the precise, automated isolation of ambiguous high-cardinality nominal identifiers (such as postal codes) from true transactional metrics.

Second, the discovered schemas are dynamically mapped into automated Star Schema layouts. They are then projected onto an in-memory multidimensional OLAP (Online Analitycal Processing) cube. This setup facilitates sub-second exploratory querying without manual structural configurations. Third, at the downstream commentary interface, the framework introduces a Hybrid NLG engine. A hybrid approach combining these two models can mitigate their respective shortcomings: while template-based models can generate narratives that are highly effective and adhere strictly to narrative guidelines, they tend to feel rigid. Generative models, on the other hand, produce more flexible narratives but risk introducing hallucinations or shifts in meaning from the intended narrative. Combining these two models into a hybrid NLG system will deliver better narratives for end users, thereby representing the implementation of Augmented Analytics in Business Intelligence.

2. Literature Review

2.1. Business Intelligence

Business Intelligence (BI) refers to a technology-driven process for analyzing data and delivering actionable information that supports informed decision-making by executives, managers, and employees. BI enables organizations

 ISSN 2723-6471

to generate data-driven insights, facilitating real-time decision-making and allowing management to enhance operational efficiency, identify emerging opportunities, and conduct market segmentation within competitive environments [7]. Data processing and collection in BI typically use SQL, with outputs transformed into tables, dashboards, reports, and graphs to support decision-making. The initial stages of a BI system include gathering data from diverse corporate sources, followed by Extract, Transform, and Load (ETL) activities into a Data Warehouse. Data collection and analysis have long been central components of BI (figure 1).

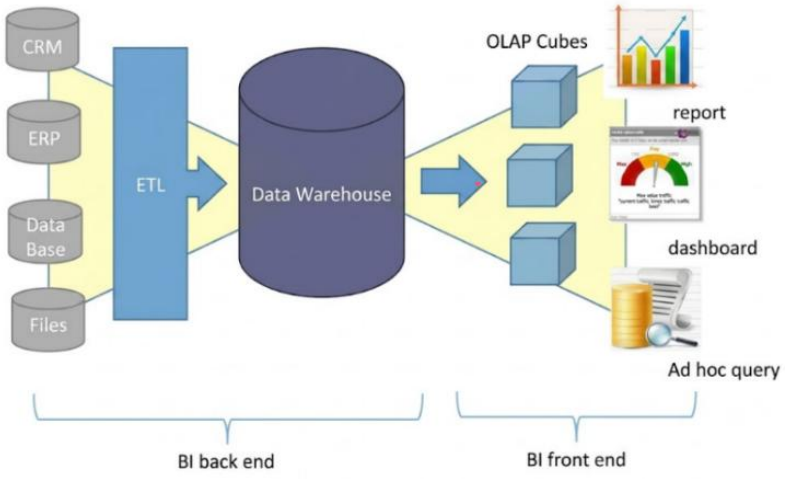

Figure 1. Architecture of Business Intelligence

Based on the BI architecture, ETL activities and the formation of a Data Warehouse determine the success of BI's ultimate goal. Data Cubes group data into a multidimensional matrix. Data Cubes allow data to be modeled and viewed in various dimensions. The multidimensional data model is organized by theme, such as sales and transactions. There is one or more fact tables representing the theme, each containing numerical measures and keys for each dimension.

2.2. Augmented Analytics

Augmented analytics represents an advanced analytical paradigm enhanced through the systematic application of Artificial Intelligence (AI), Machine Learning (ML), and Natural Language Processing (NLP) technologies [8]. Within this architecture, Machine Learning automates intricate analytical pipelines during the data preparation phase, while Natural Language Processing (NLP) serves as the core mechanism for presenting insights from processed data via user-driven conversational queries. Furthermore, Augmented Analytics leverages a symbiosis of linguistic and statistical methodologies to optimize data management performance, extending its utility across diverse enterprise domains including data analysis, data sharing, and Business Intelligence (BI) [9]. This approach significantly elevates analytical capabilities by utilizing conversational interfaces and exploiting human knowledge representations through digital assistants to extract latent insights [10] [11] . As a state-of-the-art methodology designed to navigate the complexities of Big Data, Augmented Analytics capitalizes on contemporary breakthroughs in artificial intelligence to fully automate the end-to-end analytical lifecycle [12].

2.3. Machine Learning

Machine Learning can be broadly defined as a computational method that utilizes experience gained from learning processes to improve performance and predictive accuracy[13]. Machine Learning offers numerous advantages over traditional statistical and experimental models, such as optimal accuracy, superior computational speed, responsiveness within complex environments, and enhanced cost-effectiveness [14]. The learning mechanism used in machine learning is inspired by human cognitive patterns, in which latent patterns are discovered from historical cases to address future scenarios. Consequently, Machine Learning primarily focuses on the application of data and algorithms to simulate and execute machine-driven learning, thereby enhancing overall system accuracy [15].

The workflow of Machine Learning can be categorized into several core sequential phases [16]: (1) Data Collection: The foundational step in machine learning involves gathering the necessary datasets required to train and evaluate the model. These data can encompass various formats, including imagery, textual corpora, or numerical values. (2) Data

Preprocessing: The ingested data is subsequently processed to eliminate anomalies, impute missing values, and standardize formats to ensure compatibility with machine learning models. (3) Model Selection: A critical phase that involves determining the specific algorithmic architecture appropriate for the target task, such as regression, decision trees, or neural networks. (4) Model Training: The model is trained by ingesting the training dataset. This training pipeline involves adjusting the model's internal parameters to capture latent patterns and structural relationships in the data. (5) Model Evaluation: Following the training phase, the model is benchmarked using a separate testing dataset withheld during training. Evaluation metrics are then used to quantify the model's ability to produce accurate predictions. (6) Prediction and Deployment: Once evaluated, the trained model can be deployed to execute predictions or drive automated decision-making processes on newly ingested data streams.

2.4. Natural Language Generation (NLG)

Natural Language Generation (NLG) is a specialized branch of AI programming that synthesizes written or spoken narratives from structured datasets, facilitating bidirectional interactions between humans and machines [17]. Remarkable advancements in generative modeling, propelled by neural networks and Large Language Models (LLMs), have driven the creation of intelligent conversational agents, enabling NLG to deliver highly natural language generation for human consumption [18]. Among the diverse computational models established within the NLG domain, architectures such as SLUG, seq2seq, and Long Short-Term Memory (LSTM) are frequently utilized as foundational frameworks[19].

Beyond these architectures, the core mechanism of Deep Learning-driven learning within NLG is governed by several integrated, sequential pipelines[20]. This operational lifecycle initiates with text or document structuring, which determines the optimal order and grouping of sentences within the generated text. This phase is followed by lexicalization, in which words, phrases, or morphological patterns are dynamically added to the lexicon to modify words in accordance with precise grammatical usage. Subsequently, the system performs referring expression generation to construct phrases that accurately represent nouns, culminating in linguistic realization, which compiles the actual surface text from the underlying syntactic representations.

The real-world application of NLG technology has demonstrated profound operational impacts across various strategic sectors, spanning media, customer service, public administration, and healthcare [21]. Emerging case studies in the industry underscore NLG's vast potential to optimize operational efficiency while maintaining the qualitative accuracy of data-driven information dissemination.

To avoid conceptual ambiguity, this study explicitly operationalizes the notion of Augmented Business Intelligence (Augmented BI). In the context of this research, Augmented BI is not treated as a generic AI-enhanced BI paradigm, but rather as an end-to-end analytical framework integrating four measurable capabilities: (1) autonomous schema discovery through machine learning-driven AutoETL, (2) dynamic multidimensional OLAP construction, (3) hybrid Natural Language Generation (NLG) for automated business narrative generation, and (4) semantic-aware optimization to mitigate statistical semantic blindness during schema discovery. Unlike previous studies that typically address only isolated components of the BI pipeline, the proposed framework unifies these capabilities within a single architecture. Table 1 summarizes the operational scope of existing approaches and highlights the distinctive capabilities of the proposed framework.

Table 1. Operational Comparison of Augmented BI Capabilities

Framework / Paradigm	Automated Schema Discovery	Dynamic OLAP	Automated Narrative Generation	Semantic-aware Optimization	Operational Scope
Aluso & Enyejo [4]	✓	-	LLM-based	-	ETL automation
Bozdemir & Bilgin [5]	✓	-	SQL generation	-	Schema retrieval
Sharma et al. [6]	-	-	Table-to-Text	-	Narrative generation

Valentine	✓	-	-	-	Dataset matching
Traditional BI	-	✓	Template	-	BI reporting
This Study	✓	✓	Hybrid NLG	✓	Augmented BI

3. Methodology

Figure 2 systematically illustrates the end-to-end framework for autonomous schema discovery and hybrid NLG-driven Augmented BI, providing a comprehensive operational view of the system. The architecture employs a modular, multi-tier pipeline to address manual transformation bottlenecks and overcome statistical semantic blindness.

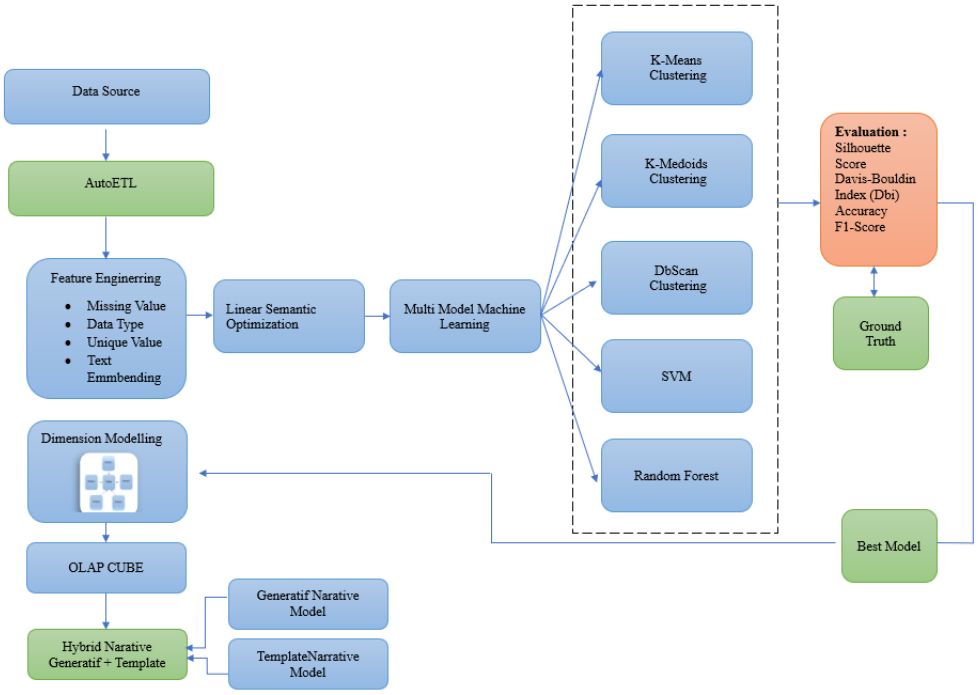

Figure 2. The proposed end-to-end multi-model framework for autonomous schema discovery (AutoETL) and hybrid NLG-driven Augmented BI.

To provide a comprehensive visual representation of how AutoETL, machine learning models, and hybrid text generation engines interact within a single pipeline, this end-to-end framework is divided into three main operational tiers.

3.1. Framework Scope and Functional Automation Boundary

The Augmented Analytics proposed in this framework emphasizes the application of AI and machine learning concepts to two layers within Business Intelligence: the Initialization Layer and the Interactive OLAP Exploration Phase. During the preliminary initialization phase, users perform three configuration steps: uploading the raw database schema instance in a standardized Comma-Separated Values (CSV) format; providing a one-time ground-truth binary annotation across the extracted metadata columns to serve as a baseline reference for supervised cross-validation tracking; and adjusting the programmatic semantic scaling slider to initialize the hyperparameter threshold α. Once these operational parameters have been set, the subsequent pipeline transitions into a hands-free computational cycle. The architecture sequentially orchestrates the extraction of physical statistical metadata, triggers the formulation of a dense embedding matrix via Sentence-BERT, and constructs a scale-invariant hybrid fused matrix to eliminate statistical semantic blindness. For multidimensional star-schema discovery, the system automatically deploys a multi-paradigm machine learning suite to determine dimensional separation rules, which are then dynamically projected into

an in-memory OLAP cube instance. On the final operational layer, the human user functions purely as an analytical explorer who interacts with the dynamic OLAP cube configurations.

3.2. Machine Learning-Driven AutoETL for Autonomous Schema Discovery

The System execution begins at Tier 1, the layer for data ingestion and semantic feature fusion via an automated processing pipeline (AutoETL). The workflow in the initial phase operates in parallel to extract internal dataset characteristics using two approaches: a data profiling module to extract structural statistical metrics, including primitive data types, the missing-value ratio, and the unique-value ratio, and to determine the cardinality of string values in raw data attributes. The phenomenon of statistical semantic blindness is formally defined within the context of heterogeneous feature engineering. This phenomenon refers to a structural anomaly that occurs when physical database profiling statistics (e.g., column cardinality ratios, null-value percentages, and table dimensions) are directly fused side-by-side with dense linguistic text embeddings derived from Sentence Transformer architectures [22]. Conceptually, text embeddings operate within a highly constrained, normalized coordinate hyperspace where dense vector values are bounded within a narrow range (typically between -1.0 and 1.0). In contrast, structural database metrics, specifically high-cardinality nominal identifiers such as primary keys or postal codes, frequently exhibit large empirical ranges and high statistical variance. When these two distinct data domains are combined without optimization, it will fail to reconcile heterogeneous attribute distributions. As a solution, the textual extraction pipeline uses a pre-trained transformer-based language model (Sentence-BERT with the all-MiniLM-L6-v2 architecture) to convert column names into dense embeddings. Before generating semantic embeddings, each column name in the dataset is subjected to lexical normalization to minimize syntactic variations and preserve the authenticity of the semantic results. The processing involves, first, replacing characters detected as underscores with spaces, and then converting all characters to lowercase. In the second step, the normalized attribute names are fed into a pre-prepared contextual prompt in the format "This database column represents <column_name>." The output of this predefined prompt is then encoded using the Sentence-BERT (all-MiniLM-L6-v2) model, which has been pre-trained to generate semantic embeddings. Once the semantic results are obtained, we have two features: the first is the statistical feature—which is the primary feature we extracted based on the previous process to obtain the statistical feature values—while the second feature is the result of the semantic analysis. These two feature representations of different dimensions are then combined into a hybrid fusion matrix, which is then fed into the StandardScaler module. The crucial point of this first stage lies in the application of weighted feature fusion through the injection of a linear semantic scaling parameter (α). The process of merging and scaling the combined feature space F_{fused} is mathematically defined by Equation (1):

$$F_{fused} = [\mathbb{S}(F_{stat}) \parallel \alpha \cdot \mathbb{S}(F_{sem})] \tag{1}$$

$F_{stat} \in \mathbb{R}^{N\times 3}$ represents the raw physical profiling matrix containing structural statistical metrics, $F_{sem} \in \mathbb{R}^{N\times 2}$ represents the contextual semantic feature matrix derived from the Sentence-BERT embeddings, and α is a semantic weighting factor controlling the relative influence of semantic information during feature fusion. The symbol $\parallel$ denotes the formal horizontal matrix concatenation operator, which directly aligns the standardized statistical space side-by-side with the scaled linguistic coordinates to construct a dimensionally stable $N \times 5$ joint matrix. To ensure scale-invariance across heterogeneous feature domains, the stabilization layer $\mathbb{S}(\cdot)$ applies an independent column-wise Z-score transformation, formulated as:

$$\mathbb{S}(F_{stat})_{i,j} = \frac{(F_{stat})_{i,j} - \mu_j^{stat}}{\sigma_j^{stat}} \tag{2}$$

$$\mathbb{S}(F_{sem})_{i,j} = \frac{(F_{sem})_{i,j} - \mu_j^{sem}}{\sigma_j^{sem}} \tag{3}$$

μ_j^{stat} and σ_j^{stat} represent the empirical mean and standard deviation computed over the j feature column of the statistical matrix, while μ_j^{sem} and σ_j^{sem} represent the empirical mean and standard deviation computed over the j feature column of the semantic matrix. This multi-dimensional mathematical transformation successfully shifts the metric coordinates of the joint feature space to balance linguistic similarity with the structural distribution. Consequently, nominal attributes with high cardinality, such as ZIP codes, are not misclassified as quantitative metrics.

Once the joint feature space F_{fused} has been fully optimized and standardized, the balanced data matrix is seamlessly fed into the multi-model execution phase. At this operational layer, the computing environment executes five machine learning algorithms to independently evaluate the performance of the data architecture and dimensional separation. This testing pipeline combines unsupervised clustering paradigms-represented by K-Means, K-Medoids, and DBSCAN-with supervised classification baselines, including Support Vector Machines (SVMs) and Random Forests. For supervised classification models, accuracy is evaluated using Leave-One-Out Cross-Validation (LOOCV). Because the number of labeled schema attributes is relatively limited, Leave-One-Out Cross-Validation (LOOCV) was employed for the supervised baseline models (Support Vector Machine and Random Forest). In each iteration, one schema attribute was used as the test instance, whereas the remaining labeled attributes were used for model training. This attribute-level validation strategy ensures that every schema attribute is evaluated exactly once while maximizing the amount of training data available during each iteration. Meanwhile, for clustering algorithms, alignment with the Ground Truth is measured in parallel using internal geometric metrics. Ground Truth was used to evaluate the proposed schema discovery framework, where a manually annotated reference dataset was constructed. The annotation process followed multidimensional data warehouse design principles, in which numerical measures representing aggregatable business metrics were labeled as facts, while categorical descriptors, identifiers, temporal attributes, and organizational entities were labeled as dimensions. The comprehensive details of the Ground Truth dataset are presented in table 2.

Table 2. Ground Truth Annotation of Database Attributes

No	Column Name	Data Type	Business Role	Ground Truth Label	Annotation Rationale
1	Customer_ID	Integer	Identifier	Dimension	Unique identifier used for grouping
2	Customer_Name	Text	Descriptor	Dimension	Categorical descriptive attribute
3	Product	Text	Descriptor	Dimension	Product category
4	Region	Text	Location	Dimension	Geographical grouping
5	Order_Date	Date	Time	Dimension	Temporal analysis
6	Quantity	Integer	Measure	Fact	Aggregatable numerical measure
7	Sales	Float	Measure	Fact	Continuous business metric
8	Profit	Float	Measure	Fact	Aggregatable business value

The annotated labels presented in table 2 serve as the reference ground truth for evaluating the proposed schema discovery framework. These labels were established based on multidimensional data warehouse design principles, where numerical attributes representing aggregatable business measures were classified as facts, while descriptive, categorical, temporal, and identifier attributes were classified as dimensions. Silhouette Score SS density index as in Equation (2) and the Davies-Bouldin Index DBI The inter-cluster distance dispersion index in Equation (3) is also used to evaluate the performance of clustering, whether semantic or non-semantic.

$$SS = \frac{b-a}{\max(a,b)} \tag{4}$$

$$DBI = \frac{1}{k}\sum_{i=1}^{k} \max_{j \neq i}\left(\frac{s_i+s_j}{d(c_i,c_j)}\right) \tag{5}$$

a denotes the average intra-cluster distance, b the average distance to the nearest inter-cluster neighbor, s the internal cluster dispersion, and $d(c\_i, c\_j)$ the geometric distance between cluster centers.

In addition, the accuracy of mapping the star schema elements to the ground truth was measured quantitatively; the system calculated external evaluation metrics in the form of Accuracy and F1-Score for all test models based on the calculation of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN), formulated using Equation (4) and Equation (5):

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{6}$$

$$\text{F1-Score} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \tag{7}$$

3.3. Multidimensional Modeling and Exploratory Data Representation via OLAP

Once the schema boundaries are identified by the machine learning engine, the framework automatically projects the classified metadata into a multidimensional structure, specifically a Star Schema. The discovered facts are mapped to central fact tables that contain numerical measures, while contextual attributes are clustered into surrounding dimension tables. This structured data warehouse serves as the foundation for building an interactive Online Analytical Processing (OLAP) cube. The framework leverages exploratory data representation through advanced OLAP operations-including roll-up, drill-down, slicing, and dicing-to navigate the multi-dimensional aggregate space. This allows the system to systematically mine hidden data trends, anomalies, and multi-layered business metrics from the underlying relational streams.

3.4. Hybrid Natural Language Generation for Narrative Explanation

The workflow in this proposed framework culminates in Tier 3, which serves as a layer for multidimensional cube projections and a hybrid business narrative interface. The Star Schema structural metadata, isolated in the previous stage, is loaded into memory (in-memory) to form an OLAP cube. This in-memory projection enables decision-makers to execute complex analytical queries through drill-down and slicing-dicing operations. As users navigate the data exploration interface, the aggregated results of pivot tables are fed into the natural language generation (Hybrid NLG) module. The intelligent execution splits text processing into two logical paths: the rule-based narrative model (Template Narrative Model) locks in transaction figures to ensure reporting boundaries that are 100% free of mathematical hallucinations. Then, the Long Short-Term Memory (LSTM) neural network generative model synthesizes syntactic phrase variations, flexible sentence transitions, and contextual action recommendations around these rigid metrics.

To evaluate the linguistic quality and textual similarity of the generated report text compared to the human-written reference text, the system calculates the BLEU score p_n based on modified n-gram alignment via Equation (8), as well as the ROUGE-L score R_{LCS} based on the combination of precision and recall of the longest common subsequences via Equation (7):

$$\text{BLEU} = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{8}$$

$$\text{ROUGE-L} = \frac{(1+\beta^2) \cdot R_{LCS} \cdot P_{LCS}}{R_{LCS} + \beta^2 \cdot P_{LCS}} \tag{9}$$

BP represents the Brevity Penalty to prevent bias in generating text that is too short, w_n is the geometric weight of the n-gram, p_n denotes the modified precision of the n-gram, while R_{LCS} and P_{LCS} define the recall and precision of the structural match of the summarized sentence elements, respectively. This narrative sequence, rich in business terminology, is ultimately displayed in real-time on an interactive dashboard.

4. Results and Discussion

This section presents the empirical evaluation and performance metrics of the proposed multi-model framework for autonomous schema discovery and hybrid NLG-driven augmented analytics. First, the accuracy and semantic discernment of the machine-learning-driven AutoETL engine are benchmarked against traditional statistical data profiling methods, with a specific focus on its ability to mitigate semantic blindness in high-cardinality-dimension columns. Second, the structural integrity and transactional processing agility of the automatically constructed OLAP cubes are assessed through multi-dimensional exploratory queries. Finally, the quality, narrative coherence, and factual precision of the generated business commentaries produced by the hybrid NLG engine.

To bridge the structural and semantic gap between database columns during the initial data ingestion phase, we developed a hybrid feature fusion mechanism [23]. Relying solely on data profiling often triggers statistical semantic blindness, causing continuous identifiers (e.g., postal codes or customer IDs) to be misclassified as facts due to their

high cardinality and numerical footprint. To resolve this, we leverage contextual text embeddings extracted via Sentence-BERT to capture the latent business meanings and semantic contexts of column headers. Furthermore, we apply a linear variable-weighting optimization inspired by foundational geometric clustering principles established in recent multi-table and schema-alignment architectures [24][25]. By adjusting the semantic scaling parameter (α), the metric space is stretched, amplifying semantic distances relative to statistical properties, thereby achieving a clean separation between facts and dimensions. Unlike conventional, isolated schema-matching frameworks that operate independently of downstream applications, this optimization directly produces a classified metadata layout tailored for immediate multidimensional warehousing.

4.1. Dataset

To rigorously evaluate the system's resilience in complex data environments, the framework was tested on a heterogeneous corporate benchmark dataset comprising transactional operational schemas. The data used in this research comes from a transactional marketplace and also a secondary dataset, Superstore, for benchmarking the results of this framework. The structural profiles of both datasets are presented in table 3.

Table 3. Dataset

No	Dataset Name	Number of Columns	Number of Rows
1	E-Commerce Transactional	12	4415
2	Superstore	20	9995

The E-Commerce Transactional dataset, despite its lower dimensionality, exhibits a high density of ambiguous data features. In this specific environment, the continuous numerical identifiers within the 12 columns tested the limits of the linear semantic scaling optimization. Because transactional streams are inherently prone to localized schema definitions, the feature fusion mechanism relied heavily on Sentence-BERT text embeddings to accurately anchor contextual definitions, ensuring that high-cardinality values were not falsely aggregated as core facts. In contrast, the Superstore dataset had a significantly broader dimensionality, with 20 columns and a larger historical footprint of 9,995 rows. This larger dimensional scale directly challenged the unsupervised clustering models.

4.2. Machine Learning-Driven AutoETL for Autonomous Schema Discovery

The workflow begins with the dual stages of raw data ingestion and feature engineering, during which statistical characteristics (data type, missing-value ratio, cardinality) are calculated, and contextual column-name features are extracted using Sentence-BERT text embeddings to mitigate the risk of structural blindness. Next, the feature matrix is linearly optimized using parameter weights (α) to adjust the metric space scale before being tested comparatively across datasets (E-Commerce and Superstore) using three learning models: K-Means, K-Medoids and DBSCAN. The semantic scaling factor (α) is operationally parameterized via an interactive adjustment slider constrained within the range [1.0, 10.0]. The default value of α=5.0 used in the main benchmark was determined through empirical experiments; this value represents the statistical inflection point at which the semantic clustering weights successfully balance the variation in database profiling metrics without causing structural distortion. The attribute clustering results are then rigorously evaluated using internal geometric metrics (Silhouette Score and Davies-Bouldin Index) and external alignment with expert assessments (Ground Truth) to identify the Best Model that effectively maps fact tables and dimensions. Meanwhile, we also used SVM and Random Forest models to classify dimensions and facts as a comparison for supervised learning algorithms.

Based on tests conducted in two different schema environments, sourced from the E-Commerce Dataset and the Superstore Dataset, integrating contextual textual features (Semantic) provides a significant advantage over a purely statistical approach (Statistical Only). The empirical results presented in table 4 and table 5 demonstrate that internal metrics such as the Silhouette Score (SH) and the Davies-Bouldin Index (DBI) are far more valid indicators for measuring the purity of the separation between fact tables and dimensions, compared to relying solely on external classification metrics like Accuracy, which are susceptible to manipulation of standard data distributions.

 ISSN 2723-6471

Table 4. Dimension Modeling by Machine Learning on E-Commerce Dataset

Machine Learning	Silhouette Score	Davies-Bouldin Index	Accuracy	F1-Score
K-Means	0.556	0.845	91.67%	90.91%
K-Medoids	0.556	0.845	91.67%	90.91%
DBSCAN	0.582	0.412	58.33%	66.67%
K-Means +Semantic	0.587	0.551	91.67%	90.91%
K-Medoids + Semantic	0.611	0.514	91.67%	90.91%
DBSCAN + Semantic	-0.294	1.152	50.00%	0.00%
SVM	N/A	N/A	83.33%	80.00%
Random Forest	N/A	N/A	91.67%	90.91%

In e-commerce, although variations of purely statistical models (K-Means and K-Medoids) achieved an Accuracy of 91.67% and an F1-Score of 90.91%, both produced suboptimal feature-space topologies, with a stagnant SH value of 0.556 and a relatively high cluster dispersion (DBI) of 0.845. When text embedding transformations using Sentence-BERT were integrated via the proposed linear semantic optimization module, the quality of the cluster's geometric structure improved significantly. The K-Means + Semantic configuration successfully boosted the SH value to 0.587 and reduced the DBI to 0.551. Peak performance was achieved by the K-Medoids + Semantic model, which recorded the highest SH score of 0.611 and the lowest DBI value of 0.514, while maintaining stable accuracy (91.67%) and F1-score (90.91%). The substantial decrease in the DBI index provides direct mathematical evidence that the proposed semantic weight parameter (α) effectively reduces intra-cluster distance and significantly increases inter-cluster separation.

Table 5. Dimension Modeling by Machine Learning on SuperStore Dataset

Machine Learning	Silhouette Score	Davies-Bouldin Index	Accuracy	F1-Score
K-Means	0.651	0.771	94.74%	90.91%
K-Medoids	0.334	1.428	63.16%	36.36%
DBSCAN	0.651	0.771	94.74%	90.91%
K-Means +Semantic	0.404	0.818	78.95%	71.43%
K-Medoids + Semantic	0.410	0.806	68.42%	62.50%
DBSCAN + Semantic	0.133	0.563	68.42%	0.00%
SVM	N/A	N/A	68.42%	0.00%
Random Forest	N/A	N/A	94.74%	88.89%

This semantic-based approach becomes even more robust when examining anomalies in the Superstore dataset. Without contextual guidance, pure clustering algorithms like K-Medoids (Statistical Only) suffer from clustering results, where the SH value drops to 0.334, and the DBI value skyrockets to 1.428, resulting in a drop in accuracy to just 63.16%. This phenomenon confirms the existence of "structural blindness," in which purely statistical models fail to recognize boundaries of business logic when faced with attributes that have similar numerical profiles but serve opposing BI architectural functions. With the proposed hybrid architecture, the K-Medoids + Semantic and K-Means + Semantic models served as a crucial step, restoring SH values to 0.410 and 0.404, respectively, and reducing the DBI damage from 1.428 to stable ranges of 0.806 and 0.818.

Although the semantic feature integration substantially improves clustering topology, as evidenced by higher Silhouette Scores and lower Davies–Bouldin Index values, the corresponding classification accuracy remains unchanged for several experiments, particularly on the E-Commerce dataset. This behavior indicates that the original statistical feature

 ISSN 2723-6471

representation was sufficient to correctly classify most schema attributes as either facts or dimensions. Consequently, the semantic features primarily improve the geometric organization of the feature space by producing more compact intra-cluster distributions and better inter-cluster separation rather than altering the final class assignments. These improvements are expected to enhance clustering robustness and stability, especially when applied to more heterogeneous schemas.

To complement the quantitative results presented in table 4 and 5, figures 3 and figure 4 provide a visual comparison of the evaluation metrics obtained from each schema discovery algorithm, using both basic statistical representations and semantically enriched feature representations.

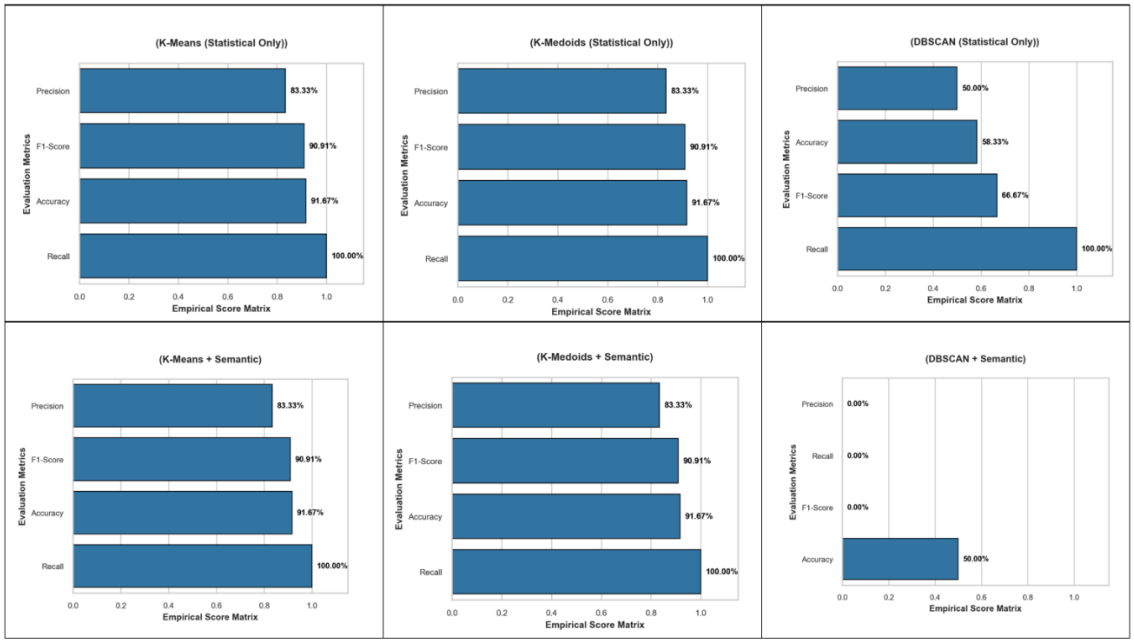

Figure 3. Performance Comparison of Schema Discovery Models (E-Commerce Dataset)

As illustrated in figure 3, the centroid-based algorithms, namely K-Means and K-Medoids, exhibit highly consistent performance under both statistical-only and semantic-enhanced feature representations. Both algorithms maintain an accuracy of 91.67%, a recall of 100%, and an F1-score of approximately 90.91%, indicating that the original statistical profiling features were already sufficient to correctly identify most fact and dimension attributes within the E-Commerce schema. Consequently, the integration of semantic features does not significantly alter the final classification labels for this relatively simple dataset.

Nevertheless, maintaining identical classification accuracy should not be interpreted as an absence of semantic contribution. As demonstrated by the clustering quality metrics discussed previously, the semantic feature integration improves the structural organization of the feature space by increasing cluster compactness and inter-cluster separation. Therefore, its primary contribution lies in enhancing representation quality and reducing semantic ambiguity rather than merely increasing predictive accuracy.

In contrast, DBSCAN shows significantly lower performance after semantic information is integrated. Specifically, DBSCAN differs from centroid-based clustering methods in that it relies on density estimates to determine cluster boundaries. Based on experiments conducted using a semantic weighting mechanism, this causes the density distribution in the combined feature space to change, resulting in a decrease in classification accuracy and F1 score.

ISSN 2723-6471

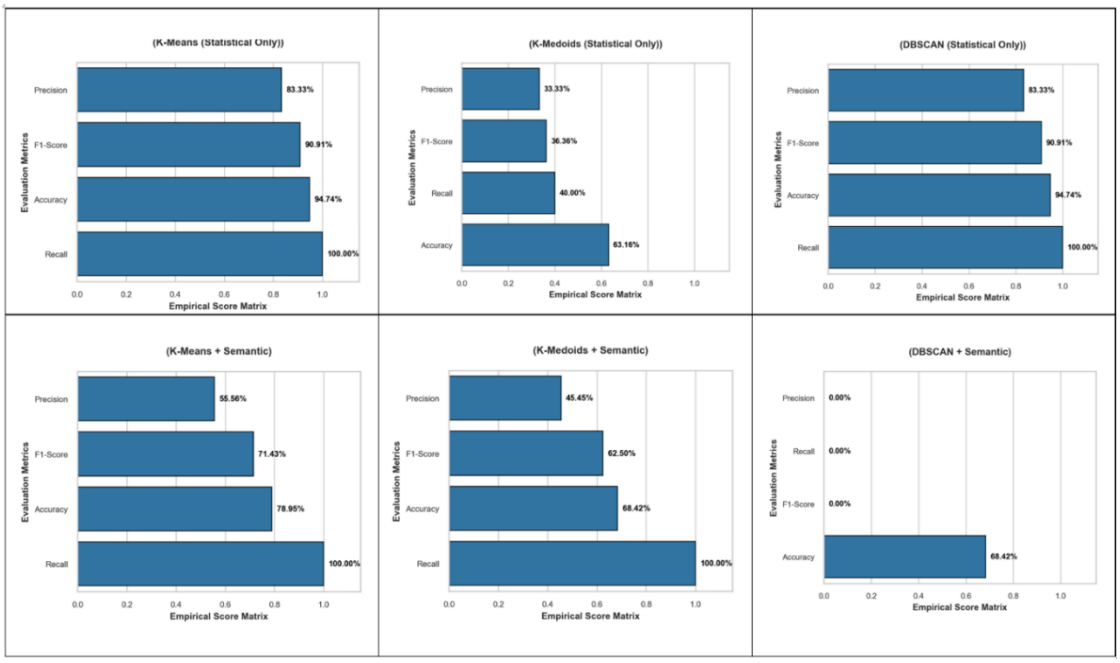

Figure 4. Performance Comparison of Schema Discovery Models (SuperStore Dataset)

Figure 4 presents a scenario using the SuperStore dataset, which has a more heterogeneous schema than the E-Commerce dataset. Under these conditions, the impact of semantic feature integration becomes more apparent. Although K-Means and K-Medoids continue to maintain high recall, accuracy, and precision change due to increased semantic diversity among the schema's attributes. The experimental results show that semantic information becomes increasingly influential as schema complexity increases. Rather than improving all evaluation metrics, semantic fusion primarily serves to organize heterogeneous attributes into more meaningful feature representations, thereby supporting more stable clustering behavior for centroid-based algorithms.

Measuring execution latency in the creation of multidimensional models, which are subsequently converted into OLAP cubes using this framework, is necessary to ensure good response performance in providing insights to end users. The results of the latency measurements are presented in table 6, broken down by the respective model-building algorithms and datasets used.

Table 6. OLAP Engine Latency

Dataset	Machine Learning	Latency Time
E-Commerce	K-Means	0.1751 s
	K-Medoids	0.1829 s
	DBSCAN	0.1863 s
	K-Means +Semantic	0.2940 s
	K-Medoids + Semantic	0.1781 s
	DBSCAN + Semantic	0.2231 s
	SVM	0.1908 s
	Random Forest	1.2440 s
SuperStore	K-Means	0.2788 s
	K-Medoids	0.2364 s
	DBSCAN	0.1981 s
	K-Means +Semantic	0.3031 s

K-Medoids + Semantic	0.2316 s
DBSCAN + Semantic	0.2145 s
SVM	0.2361 s
Random Forest	2.0132 s

Most algorithms work well at creating multidimensional models for OLAP cubes. For the E-Commerce dataset, most clustering-based approaches complete the preprocessing phase within 0.17 to 0.29 seconds, indicating that the proposed framework can prepare the analytical environment in well below one second. Similar behavior is observed on the larger SuperStore dataset, where preprocessing remains below 0.31 seconds for all unsupervised approaches despite the increased schema complexity. In contrast, the supervised Random Forest baseline requires substantially longer execution times. This indicates that lightweight clustering-based schema discovery provides a more efficient preprocessing strategy for interactive Business Intelligence environments.

4.3. Hybrid Natural Language Generation

Therefore, to mitigate this issue, the narrative accompanying the table will serve as a valuable tool in harmonizing perceptions regarding the results of data interpretation. The conceptualized hybrid model is depicted as illustrated in figure 5.

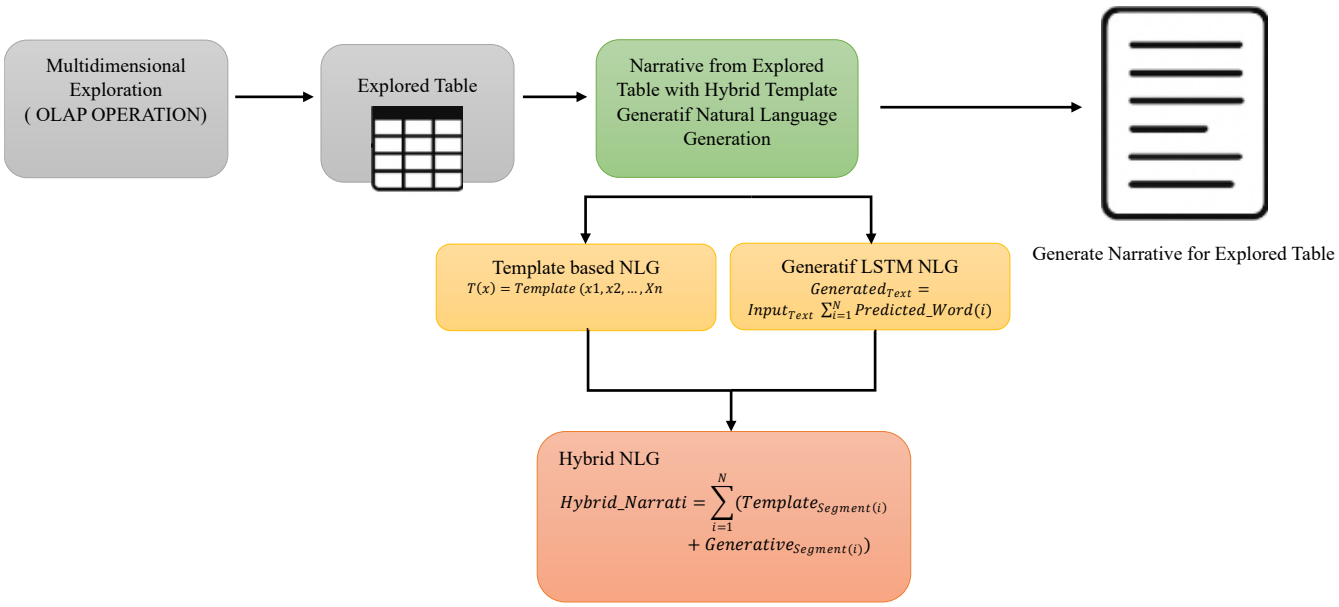

Figure 5. hybrid Natural language Generation Model

For each operation that emerges from data exploration using pivots, the narrative is constructed using template-based NLG. The system will autonomously discern statistical operations, such as the identification of maximum or minimum values from the pivot table results, which are to be utilized as key insights, i.e., growth or increase, reflecting the highest aggregate value and combinations of particular dimensions. To enhance the expressiveness of the narrative, a hybrid approach to NLG is employed, wherein template-based narratives serve as the foundational framework for neural-based token resistance models predicated on sequence learning. The objective of the LSTM component is not to develop a general-purpose language model. Instead, it serves as a domain-specific narrative refinement module that extends structured analytical templates into more natural executive-level narratives. Consequently, the training corpus is intentionally composed of curated Business Intelligence narrative templates rather than a large open-domain text corpus. The specific architectural and hyperparameter configurations of this LSTM model within the Hybrid NLG framework are outlined in table 7.

Table 7. LSTM Configuration in Hybrid NLG

Component	Configuration
Training Corpus	Curated Business Intelligence narrative templates
Corpus Size	100 domain-specific narrative templates
Tokenization	Keras Tokenizer (word-level)
Embedding Layer	100-dimensional embedding vectors
LSTM Layer 1	150 hidden units (return_sequences=True)
LSTM Layer 2	100 hidden units
Output Layer	Dense Softmax
Loss Function	Categorical Cross-Entropy
Optimizer	Adam
Training Epochs	150 Epoch

Since the narrative generator operates only after structured business facts have been extracted by the template engine, the LSTM is responsible for improving fluency and contextual coherence instead of generating unrestricted natural language. Consequently, the model requires substantially less training data than open-domain neural language generation systems. To substantiate the results, evaluations were conducted using BLEU scores, as well as precision, recall, and F1, ROUGE-1, and ROUGE-L metrics against the resulting narrative, as delineated in table 8.

Table 8. Hybrid NLG Evaluation Score

Model	BLEU	ROUGE-1 (P)	ROUGE-1 (R)	ROUGE-1 (F1)	ROUGE-L (P)	ROUGE-L (R)	ROUGE-L (F1)
Template-based	0.14	0.25	0.65	0.36	0.23	0.59	0.33
LSTM Generative	0.07	0.32	0.11	0.32	0.41	0.41	0.41
Hybrid NLG	**0.15**	**0.23**	**0.85**	**0.36**	**0.22**	**0.82**	**0.35**

Table 8 shows the evaluation results of the narrative generation models. BLEU measures the n-gram overlap between the generated narrative and the reference narrative, whereas ROUGE-1 and ROUGE-L are reported using Precision (P), Recall (R), and F1-score (F1) to provide a comprehensive assessment of lexical overlap and sequence similarity. Hi gher values indicate better agreement with the reference narrative. We can see that Hybrid NLG is better than Templa te-based NLG in terms of BLEU scores. Therefore, the Hybrid approach can be considered superior to the Template-b ased NLG model, which is known for its strong ability to preserve the narrative context of the generated text.

5. Conclusion

This study has successfully developed and validated an end-to-end, integrated multi-model framework for Augmented Business Intelligence (BI). By combining machine learning-driven AutoETL, in-memory multidimensional OLAP cubes, and a context-aware hybrid Natural Language Generation (NLG) engine, the proposed architecture delivers a resilient pipeline that transforms raw, heterogeneous enterprise data streams into executive-ready insights without requiring manual intervention. The primary contribution of this proposed framework lies in its ability to bridge the gap between back-end data structural profiling and front-end semantic narration. By implementing a hybrid feature-fusion loop that combines Sentence-BERT embeddings with linear semantic-scaling optimization, the framework provides a robust, dataset-agnostic solution to the persistent challenge of statistical semantic blindness, bypassing the resource-intensive bottlenecks traditionally associated with manual schema matching. The empirical scalability and architectural stability of the unified pipeline were rigorously validated across varying data complexities using the E-Commerce and high-dimensional Superstore datasets. At the multi-model machine learning layer, the integration of contextual features successfully insulated schema modeling from structural schema misclassification. On the E-Commerce dataset, the

proposed K-Medoids+Semantic configuration demonstrated topological superiority by achieving a peak Silhouette Score of 0.611 and a compressed Davies-Bouldin Index of 0.514. Meanwhile, on the higher-dimensional Superstore dataset, the pipeline maintained high functional flexibility, achieving optimal performance with overall classification accuracy up to 94.74%. Crucially, in the downstream insight interpretation stage, the proposed Hybrid NLG module integrates template-based analysis narratives and LSTM-based narrative refinement to generate better business narratives. Experimental results show that the hybrid model achieves the highest BLEU score of 0.15 compared to either the template-based or LSTM-based model. Although the improvement is still relatively small (an increase of 0.1), this hybrid architecture effectively balances factual consistency and linguistic fluency while preserving the analytical information generated from OLAP exploration. The proposed framework demonstrates consistent performance on two heterogeneous business datasets, its experimental evaluation is restricted to relatively simple dataset schemas. Consequently, the framework's generalization capability in large-scale enterprise environments with hundreds of attributes, multiple relational tables, highly heterogeneous schemas, and noisy metadata remains to be thoroughly examined. Future research should assess this framework using larger industrial datasets and a broader range of business domains to evaluate its scalability and robustness in real-world scenarios.

6. Declarations

6.1. Author Contributions

Conceptualization: R.P.., S.D., and G.W.N.; Methodology: R.P..; Software: R.P..; Validation: R.P.., S.D., and G.W.N.; Formal Analysis: R.P.., S.D., and G.W.N.; Investigation: R.P..; Resources: R.P..; Data Curation: R.P..; Writing Original Draft Preparation: R.P.., S.D., and G.W.N.; Writing Review and Editing: R.P.., S.D., and G.W.N.; Visualization: R.P..; All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

This work was fully funded and supported by Universitas Putra Indonesia YPTK Padang, which provided both financial sponsorship and the institutional infrastructure necessary to conduct this research.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] B. M. Olukoya, G. O. Ogunleye, P. O. Olabisi, and A. S. Adegoke, "Heterogeneous Ensemble Feature Selection: An Enhancement Approach to Machine Learning for Phishing Detection," *Int. J. Softw. Eng. Comput. Syst*., vol. 10, no. 1, pp. 60–74, 2024, doi: 10.15282/ijsecs.10.1.2024.6.0124.

[2] A. I. Gide and A. A. Mu'azu, "A Novel Approach for Addressing IoT Networks Vulnerabilities in Detection and Classification of DoS/DDoS Attacks," *Int. J. Softw. Eng. Comput. Syst*., vol. 10, no. 1, pp. 50–59, 2024, doi: 10.15282/ijsecs.10.1.2024.5.0123.

[3] A. Ramadhanu, J. Na'am, G. W. Nurcahyo, and Y. Yuhandri, "Development of affine transformation method in the reconstruction of songket motif," *Int. J. Adv. Sci. Eng. Inf. Technol*., vol. 12, no. 2, pp. 600–606, 2022, doi: 10.18517/ijaseit.12.2.14069.

[4] L. Aluso and J. O. Enyejo, "Integrating ETL workflows with LLM-augmented data mapping for automated business intelligence systems," *Int. J. Sci. Res. Mod. Technol*., vol. 2, no. 11, pp. 76–89, 2023, doi: 10.38124/ijsrmt.v2i11.1078.

[5] M. Bozdemir and M. Bilgin, "Schema Retrieval with Embeddings and Vector Stores Using Retrieval-Augmented Generation and LLM-Based SQL Query Generation," *Appl. Sci*., vol. 16, no. 2, pp. 1–32, 2026, doi: 10.3390/app16020586.

[6] S. Sharma, Z. Li, and Q. Zhang, "Table-to-Text Generation: A Survey on Challenges, Methods, and Directions," *Inf. Fusion*, vol. 86, pp. 1–15, 2022, doi: 10.1016/j.inffus.2022.07.003.

[7] A. Tripathi, T. Bagga, and R. K. Aggarwal, "Strategic impact of business intelligence: A review of literature," *Prabandhan Indian J. Manag*., vol. 13, no. 3, pp. 35–48, 2020, doi: 10.17010/pijom/2020/v13i3/151175.

[8] N. Prat, "Augmented Analytics," *Bus. Inf. Syst. Eng*., vol. 61, no. 3, pp. 375–380, 2019, doi: 10.1007/s12599-019-00589-0.

[9] Z. Ahmad and M. S. M. Sanjudharan, "Augmented Analytics: The Future of Business Intelligence," *Recent Trends Comput. Sci. Softw. Technol*., vol. 5, no. 1, pp. 7–13, 2020, doi: 10.5281/zenodo.3757837.

[10] K. Lepenioti, A. Bousdekis, D. Apostolou, and G. Mentzas, "Human-augmented prescriptive analytics with interactive multi-objective reinforcement learning," *IEEE Access*, vol. 9, no. 2, pp. 100677–100693, 2021, doi: 10.1109/ACCESS.2021.3096662.

[11] R. Odogwu, J. C. Ogeawuchi, A. A. Abayomi, O. A. Agboola, and S. Owoade, "Bridging the gap between data science and decision makers: a review of augmented analytics in business intelligence," *Int. J. Manag. Organ. Res.*, vol. 2, no. 3, pp. 61–69, 2023, doi: 10.54660/IJMOR.2023.2.3.61-69.

[12] T. D. Oesterreich, E. Anton, and F. Xu, "Augmenting the Future: An Exploratory Analysis of the Main Resources, Use Cases and Implications of Augmented Analytics," *ECIS 2021 Res. Pap*., vol. 2021, no. 19, p. 19, 2021.

[13] M. Mohri, A. Rostamizadeh, and A. Talwalkar, Foundations of machine learning. London: MIT press, 2018.

[14] M. Mohtasham Moein et al., "Predictive models for concrete properties using machine learning and deep learning approaches: A review," *J. Build. Eng*., vol. 63, no. 1, p. 105444, 2023, doi: 10.1016/j.jobe.2022.105444.

[15] M. Soori, B. Arezoo, and R. Dastres, "Machine learning and artificial intelligence in CNC machine tools, A review," *Sustain. Manuf. Serv. Econ*., vol. 2, no. 4, p. 100009, 2023, doi: 10.1016/j.smse.2023.100009.

[16] Y. Ma, P. Xu, M. Li, X. Ji, W. Zhao, and W. Lu, "The mastery of details in the workflow of materials machine learning," *npj Comput. Mater*., vol. 10, no. 1, p. 141, 2024, doi: 10.1038/s41524-024-01331-5.

[17] A. K. Pandey and S. S. Roy, "Natural Language Generation Using Sequential Models: A Survey," *Neural Process. Lett.*, vol. 55, no. 6, pp. 7709–7742, 2023, doi: 10.1007/s11063-023-11281-6.

[18] V. Srinivasan, S. Santhanam, and S. Shaikh, "Using reinforcement learning with external rewards for open-domain natural language generation," *J. Intell. Inf. Syst.*, vol. 56, no. 1, pp. 189–206, 2021, doi: 10.1007/s10844-020-00626-5.

[19] T. Liu, W. Wei, and W. Y. Wang, "Table-to-Text Natural Language Generation with Unseen Schemas," arXiv:1911.03601v1, vol. 2019, no. 11, p. 10, 2019.

[20] Y. Guo et al., "Urania: Visualizing Data Analysis Pipelines for Natural Language-Based Data Exploration," *ACM Trans. Interact. Intell. Syst*., vol. 16, no. 1, pp. 1–26, Jan. 2026, doi: 10.1145/3770071.

[21] M. Lyu et al., "Natural language generation in healthcare: A review of methods and applications," *J. Biomed. Inform*., vol. 176, no. 4, p. 104997, 2026, doi: 10.1016/j.jbi.2026.104997.

[22] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Nov. 2019, pp. 3982–3992. doi: 10.18653/v1/D19-1410.

[23] C. Koutras et al., "Valentine: Evaluating Matching Techniques for Dataset Discovery," in *2021 IEEE 37th International Conference on Data Engineering (ICDE)*, 2021, pp. 468–479. doi: 10.1109/ICDE51399.2021.00047.

[24] J. Z. Huang, M. K. Ng, H. Rong, and Z. Li, "Automated variable weighting in k-means type clustering," *IEEE Trans. Pattern Anal. Mach. Intell*., vol. 27, no. 5, pp. 657–668, 2005, doi: 10.1109/TPAMI.2005.95.

[25] R. Cordeiro de Amorim and B. Mirkin, "Minkowski metric, feature weighting and anomalous cluster initializing in K-Means clustering," *Pattern Recognit*., vol. 45, no. 3, pp. 1061–1075, 2012, doi: 10.1016/j.patcog.2011.08.012.