systems

MDPI

Article

BERT-Based Schema Matching for Integrating Heterogeneous Flood Data: A Case Study in Korea

Taeyoung Choe [1], Mincheol Shin [1], Kwangyoung Kim [2], Myungseok Yang [2], Ka Lok Man [3] and Mucheol Kim [1,*]

[1] Department of Computer Science and Engineering, Chung-Ang University, Seoul 06974, Republic of Korea; tychoe9601@gmail.com (T.C.); mcsin1648@gmail.com (M.S.)
[2] Department of Data Headquarters Strategy Team, Korea Institute of Science and Technology Information (KISTI), Daejeon 34141, Republic of Korea; glorykim@kisti.re.kr (K.K.); msyang@kisti.re.kr (M.Y.)
[3] School of Advanced Technology, Xi'an Jiaotong-Liverpool University, Suzhou 215123, China; ka.man@xjtlu.edu.cn
* Correspondence: kimm@cau.ac.kr

Abstract

Integrating flood-response datasets across municipalities is often hindered by heterogeneous and non-standard variable names, a challenge amplified in Korean by local naming conventions and linguistic variation. This study addresses scalable schema alignment to standardize municipal flood datasets with reduced manual effort while maintaining semantic consistency for downstream modeling. We propose a BERT-based schema matching framework that augments standardized attribute names with paraphrases generated by a generative language model and filtered to reduce semantic drift. Both standardized and target variable names are encoded using a flood-domain-adapted Korean BERT model, and candidate correspondences are retrieved via cosine-similarity ranking to produce top-k match suggestions for automated or human-in-the-loop alignment. Experiments on real flood-related tables from Busan and Incheon, evaluated jointly to diversify variable expressions, show that augmentation substantially improves top-k retrieval accuracy. In the combined evaluation, Hit@5 improves from 0.71 to 0.95, supporting more reliable schema harmonization for simulation-ready inputs.

Keywords: disaster management; schema matching; LLMs

Check for updates

Academic Editor: Zaoli Yang

Received: 19 December 2025
Revised: 9 February 2026
Accepted: 15 February 2026
Published: 2 March 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Urban flooding represents a severe hazard. It arises when intense rainfall exceeds the capacity of drainage and stormwater systems in densely built areas [1]. In recent years, Korea and many other regions worldwide experience repeated large-scale flood damage during heavy rainfall. Climate change intensifies short-duration precipitation extremes. It amplifies flood risk in urban environments [2]. Such events damage residential areas, industrial facilities, and transport networks, with long-term disruptions imposed on overall urban functionality [3]. Systematic data collection and analysis are critical to risk assessment, providing the essential foundation for effective disaster management in urban areas [4].

In Korea, multiple municipalities and public agencies manage extensive flood-related datasets, but their schemas are heterogeneous in both format and naming conventions. Beyond surface name differences, attributes may be produced by different agencies and vendors, recorded with abbreviations or professional terminology, and stored under different

unit conventions or measurement definitions (e.g., instantaneous vs. cumulative values). These practical inconsistencies are amplified in Korean, where short column names often rely on Sino-Korean compounds and local shorthand; thus, character-level similarity can be unreliable even when two attributes refer to the same concept. For instance, as shown in Table 1, the attributes 'Gaebyeol Beonho (개별번호)' and 'Goyu Sikbyeolja (고유식별자)' both denote a unique identifier. Similarly, 'Gwangeo Hyeongtae (관거형태)' and 'Gujomul Hyeongtae (구조물형태)' both indicate sewer structure type. Such inconsistencies hinder integration and delay analytical workflows that support rapid disaster response.

Table 1. Variable Names for Sewer Network Data by Municipality.

English Description	Korean Variable Name (Busan City)	Korean Variable Name (Incheon City)
Unique identifier	Gaebyeol Beonho (개별번호)	Goyu Sikbyeolja (고유식별자)
Installation date	Seolchi Ilja (설치일자)	Seolchi Ilja (설치일자)
Sewer structure type	Gwangeo Hyeongtae (관거형태)	Gujomul Hyeongtae (구조물형태)
Manhole diameter	Ttukkeong Gugyeong (뚜껑구경)	Ttukkeong Gugyeong (뚜껑구경)
Manhole material	Ttukkeong Jaejil (뚜껑재질)	Ttukkeong Jaejil (뚜껑재질)
Administrative district	Haengjeong-dong (행정동)	Haengjeong-dong (행정동)

This study therefore asks whether a domain-adapted semantic matching approach can provide schema alignment that is accurate, robust, and practical for real municipal integration pipelines, where schema harmonization is often the gating step for cross-jurisdiction flood analytics. We formulate the task as top- k retrieval: for each non-standard municipal attribute name, the system returns a short candidate list of standardized attributes that can be accepted automatically under high confidence or verified in a human-in-the-loop workflow.

Schema matching refers to the process of identifying and aligning semantically equivalent attributes across heterogeneous data sources [5]. Prior work classifies schema matching approaches into explicit knowledge-driven methods and data-driven learning methods [6]. Rule-based methods rely on predefined conditions or regular expressions to detect patterns in column names. This approach ensures simplicity, which in turn enables fast deployment. However, it requires continuous rule updates, which limits adaptability when previously unseen expressions appear. Dictionary-based approaches utilize glossaries or thesauri to map variable names to standardized expressions. They support the initial stages of integration, with high maintenance costs in domains where terminology evolves rapidly and new concepts emerge frequently. Similarity-based approaches calculate string distances or statistical similarity scores to identify candidate matches. In morphologically complex languages such as Korean, character-level similarity makes it hard to capture semantic equivalence. Ontology-based approaches define hierarchical domain concepts and relationships to support systematic alignment. Yet, they require intensive expert knowledge for construction and maintenance. This dependency makes it challenging to keep pace with rapidly changing disaster-related terminology [7].

Recent advances in deep learning and natural language processing (NLP) drive the adoption of semantic similarity models for schema matching [8,9]. Embedding-based techniques represent contextual meaning [10]. Pre-trained large language models (LLMs) further extend this capability to support automated mapping between heterogeneous variable names [11]. For instance, embedding models identify attributes that describe highly similar concepts even when surface forms differ [12]. Recent work has also explored schema alignment with prompted large language models (LLMs) [11,13]. In contrast, we use an LLM only for offline lexicon expansion of standardized attribute names and perform the

matching stage through embedding-based retrieval using a flood-domain-adapted Korean BERT encoder. This hybrid design improves lexical coverage while keeping online matching deterministic and auditable, producing an explicit similarity-ranked candidate list that can be inspected and logged in operational pipelines.

Despite these advantages, existing approaches reveal important limitations. Many approaches depend on specific data formats. This dependency restricts generalizability across domains [14]. Some methods require extensive domain expertise for reliable performance [7]. Furthermore, several studies emphasize semantic similarity alone. They only partially capture contextual variations in terminology [13]. These challenges intensify in morphologically complex languages such as Korean, where schema alignment is hindered by Sino-Korean compounds and frequent neologisms.

To address these challenges, this study proposes a BERT-based automatic schema matching model for flood data integration. The model fine-tunes a Korean BERT with domain-specific corpora, including flood-related public documents and news articles, to capture terminology and contextual nuances in disaster datasets. It computes similarity scores on embeddings of variable names to determine matches between non-standard and standardized attributes. The approach introduces a text augmentation technique based on generative language models. This technique exposes the system to diverse paraphrases, which in turn enhances generalization. The contributions of this study are as follows:

1. Proposes an automated schema matching method using BERT that reduces data preparation costs for flood data integration.
2. Develops a model capable of standardizing heterogeneous variable names without requiring explicit domain expertise.
3. Introduces generative text augmentation that enriches the model's ability to learn diverse expressions and leads to higher accuracy in matching.
4. Demonstrates that the proposed method achieves notable accuracy improvements compared to existing embedding-based approaches.
5. Validates the practical utility of the model through a case study that applies the method to real Korean flood datasets for simulation-based disaster response.

2. Related Work

The literature commonly classifies schema matching approaches into explicit knowledge-driven methods and data-driven learning methods [5,15,16]. Explicit approaches rely on manually defined rules, dictionaries, and ontologies, while data-driven approaches employ statistical similarity, embeddings, and pre-trained models. This study discusses prior research accordingly, following a similar categorization.

2.1. Explicit Knowledge-Driven Approaches

2.1.1. Rule-Based Methods

Rule-based approaches represent one of the earliest directions in schema matching. Do & Rahm [17] introduce the COMA system, which flexibly combines rule-based heuristics with other strategies. Elmagarmid et al. [18] discuss the use of rule-based detection in data cleaning and integration tasks, while Kedad & Xue [19] extend the idea to XML data. Chen et al. [20] present BigGorilla, an open-source ecosystem that integrates multiple rule-based and heuristic components for end-to-end data integration, demonstrating the continued relevance of rule-based heuristics in practical systems. These methods provide simplicity and fast deployment, yet they require frequent updates when new terms appear. Consequently, in large-scale disaster datasets where regional variations in terms are common, rule-based strategies become costly to maintain.

2.1.2. Dictionary-Based Methods

Dictionary-based approaches map variable names to standardized expressions using glossaries or thesauri. Early studies employ resources such as WordNet to support general-purpose schema alignment [16]. Rashid et al. [21] propose a Semantic Data Dictionary that formalizes column descriptions using standard controlled vocabularies and ontological concepts to improve metadata consistency across heterogeneous biomedical datasets. Asif-Ur-Rahman et al. [22] develop a semi-automated hybrid dictionary framework for vegetation data, showing that controlled vocabularies can bridge heterogeneous ecological datasets. Despite these advances, dictionary-based methods still face high maintenance costs when neologisms or domain-specific terms emerge frequently.

2.1.3. Ontology-Based Methods

Ontology-based approaches define domain concepts and their relationships hierarchically to support semantic integration. Z. Wu et al. [7] propose an ontology framework for urban flood data integration, which improved consistency across heterogeneous sources. K. Wu et al. [23] incorporate semantic constraints into ontology-based schema matching to enhance accuracy. Although ontologies provide structured and interpretable mappings, they demand intensive expert involvement and frequent updates, which reduce their applicability in fast-evolving domains such as disaster management.

2.2. Data-Driven Learning Approaches

2.2.1. Embedding-Based Models

The development of word embeddings has shifted schema matching from surface-level similarity to semantic representation. Pan et al. [24] apply semantic similarity metrics for integrating building energy datasets, showing effectiveness when standardized terminology exists. Oh et al. [12] employ word embeddings for schema alignment in lifecycle management datasets, showing improvements over rule-based baselines. Kired et al. [13] extend embedding-based matching to disparate data sources by demonstrating that embeddings capture semantic proximity across heterogeneous attributes. Koutras et al. [8] adopt graph embeddings in REMA, where embeddings supported relational schema alignment. Mukherjee et al. [25] propose learning knowledge graphs for schema matching, highlighting that embedding representations can be enhanced with structural and semantic constraints. These studies collectively show that embedding-based approaches enhance generalization beyond string similarity. However, their performance often depends on the training corpus, and they may fail when confronted with domain-specific terminology that does not present in the embedding space.

2.2.2. Pre-Trained Language Models

The introduction of pre-trained language models has further advanced schema matching research. Hättasch et al. [9] propose AI Match, a two-step approach that uses embeddings derived from pre-trained models to improve matching accuracy. Ayala et al. [10] introduce LEAPME, a learning-based property matching system leveraging embeddings, which achieved high performance across multiple datasets. These works highlight that pre-trained models allow for more accurate alignment of attributes with surface differences when they capture contextual meaning. Nevertheless, their reliance on large training corpora and limited domain adaptation restrict their effectiveness in specialized areas such as disaster management.

2.2.3. Generative Models and Data Augmentation

Recent work explores generative models to improve schema matching. Narayan et al. [26] question the extent to which foundation models can handle data wrangling and integration tasks, emphasizing both the promise and the risks of relying on large pretrained models. Sheetrit et al. [14] introduce ReMatch, which enhances schema alignment through retrieval-augmented generation, demonstrating the utility of large language models (LLMs) in schema tasks. Parciak et al. [11] show through their study that schema matching with LLMs depends significantly on prompt design and context provision. These studies indicate that LLMs can act not only as standalone matchers but also as components that enhance schema matching workflows. Building on this perspective, we argue that data augmentation strategies, for example by exposing models to diverse paraphrases, can further improve generalization. However, such strategies also raise concerns about the quality and reliability of synthetic data, which requires careful validation.

This paper proposes an automatic schema matching model using BERT to address the need for heterogeneous data integration in disaster management. The approach addresses the limitations of ontology-based methods by transforming flood data attributes into high-dimensional embedding vectors, which support matching between standardized and non-standardized variable names. The proposed model effectively handles the diversity and heterogeneity of data while delivering higher performance compared to existing methods.

3. Schema Matching on Flood Data with LLMs

3.1. Preliminary

BERT represents a key advancement in natural language processing. It introduces bidirectional pre-training by capturing contextual meaning from both the left and the right of a target token [27]. The model employs a multi-layer Transformer encoder that applies self-attention across tokens to represent each word in relation to its surrounding context [28]. The pre-training strategy of BERT includes Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). MLM randomly masks tokens in the input, which forces the model to predict the missing elements and thereby achieve bidirectional dependency learning. NSP trains the model to determine whether two sentences appear consecutively, which enables understanding of inter-sentence relationships. These tasks produce generalizable language representations that can be applied to a wide range of downstream tasks.

Variants of BERT extend this framework to multilingual and domain-specific contexts. Multilingual BERT (mBERT) encodes multiple languages with a shared vocabulary, demonstrating cross-lingual transfer in morphologically complex settings [29]. BioBERT enhances the ability to capture specialized terminology by fine-tuning BERT on biomedical corpora [30]. KR-BERT provides a pre-trained foundation that reflects linguistic features specific to Korean [31]. KoBERT is another widely used Korean language model released through an official repository by SKT Brain [32].

This study fine-tunes a Korean BERT model with flood-related public documents and news articles. Fine-tuning enables the encoder to recognize contextual nuances of domain-specific terminology and the generation of embeddings of attribute names. These embeddings establish the foundation for the schema matching framework described in the following sections.

3.2. Overview

This study presents a BERT-based schema matching model for the integration of flood disaster data. The model focuses on unifying heterogeneous column names in flood datasets while addressing the complexity of Korean linguistic expression and the non-

standard naming that arises from diverse sources. It consists of three modules: a text augmentation module, an embedding generation module, and an automatic matching module. Each module contributes to broadening data diversity and reflecting semantic similarity with precision.

Figure 1 illustrates the overall schema matching process. The framework processes multi-source flood data through the three modules in sequence, producing unified column names for schema matching. The subsequent section describes each module in detail.

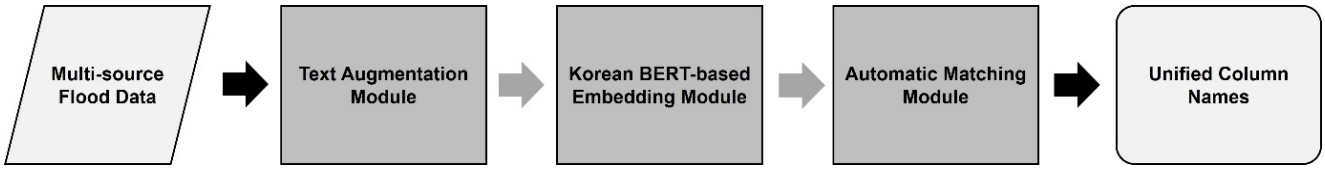


Figure 1. Overview of the proposed schema matching model.

3.3. Text Augmentation Module

The text augmentation module employs a generative language model to enrich the representation of variable names in flood data (Figure 2). Given an input variable name, the model generates a small set of alternative expressions intended to preserve the original semantics while reflecting linguistic variation in Korean (e.g., paraphrases, synonymous terms, and common variants). For example, when the standard variable name 'shape of pipe (관의 모양)' is provided, the generative model produces expressions such as 'sewer pipe form (하수관 형태)' and 'pipeline shape (배관 모양)'. To mitigate semantic drift, we apply similarity-based filtering: for each generated candidate, we compute cosine similarity between its embedding and the embedding of the source variable name, and discard candidates that fall below a threshold τ. The refined alternatives are then used as additional training signals for the embedding module, improving robustness to non-standard and region-specific naming.

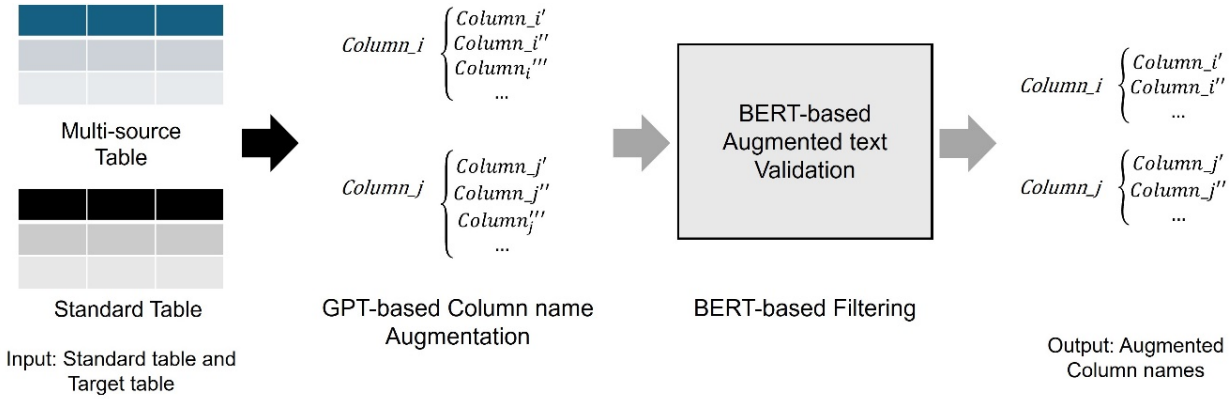


Figure 2. Description of the text augmentation module.

Through this procedure, the model acquires robustness and generalization capacity, enabling it to align non-standard column names with standard variables even when they employ different expressions. This approach proves particularly effective in Korean data environments where Sino-Korean compounds and complex phrasing frequently occur.

3.4. BERT-Based Embedding Module

The Korean BERT-based embedding module converts the variable names derived in the previous stage into high-dimensional vectors. The Korean BERT model learns from a corpus collected from diverse news articles, with additional fine-tuning on news and

public data to specialize in disaster management. This process equips the model with the capacity to interpret flood data. It also provides the ability to process specialized terminology and varied linguistic forms. To adapt the encoder to the flood domain, we compiled a Korean corpus from publicly available flood-related documents (e.g., guidelines, reports, and technical materials) and flood-related news articles. We applied basic text cleaning (removing markup and non-linguistic artifacts, normalizing whitespace, and deduplicating near-identical lines) before fine-tuning the encoder for sentence-level semantic similarity. This domain adaptation is intended to better capture specialized terminology frequently observed in municipal flood datasets.

During data preprocessing, the module normalizes column names by removing unnecessary special characters and spaces, which enhances consistency. It also standardizes variable names that appear in multiple formats, including English words, English acronyms, and Korean terms, according to naming conventions such as camel case. It then transforms the standardized names into both English and Korean representations. These steps enable the alignment of attributes that share the same concept even when they differ in surface form.

In the embedding stage, the module receives normalized variable names as input, generating high-dimensional vectors that reflect semantic similarity. These vectors serve as input for the subsequent automatic matching module. Figure 3 illustrates the preprocessing and embedding procedure. Through this process, the framework evaluates semantic similarity between standard and non-standard variable names. It then prepares them for matching in the next stage.

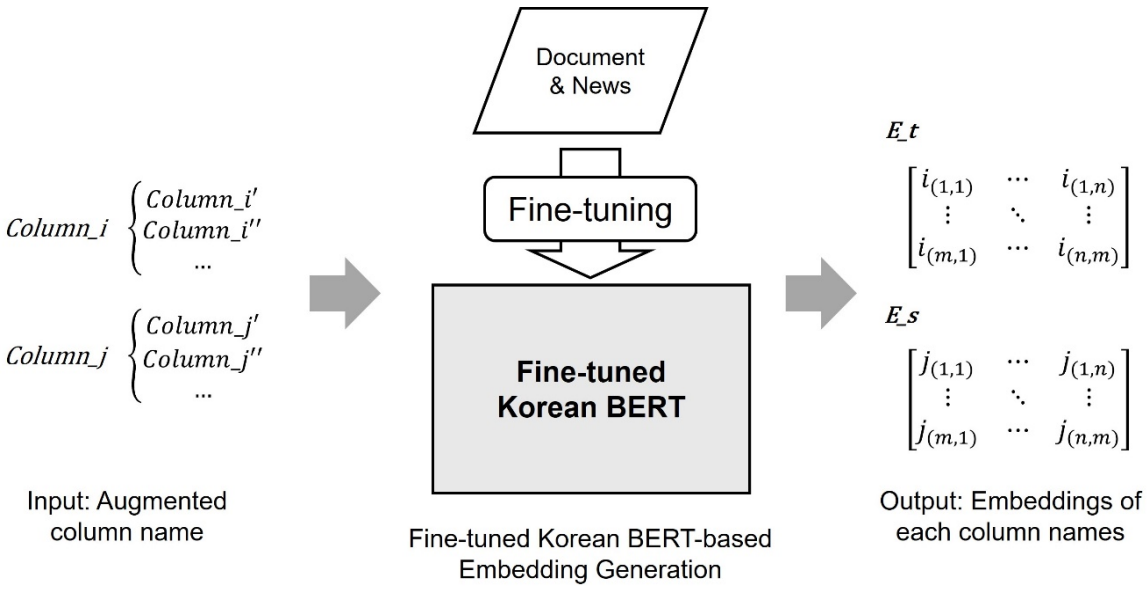


Figure 3. Description of the BERT-based embedding module.

3.5. Automatic Matching Module

The third module performs the task of aligning non-standard column names with standard ones, thereby automating data integration. The embedding vectors produced in the previous stage undergo comparison using cosine similarity, with values above a predefined threshold indicating a match. Equation (1) shows the computation of cosine similarity.

$$Cosine\ Similarity(E_t,\ E_s) = \frac{E_t \cdot E_s}{\|E_t\|\|E_s\|} \quad (1)$$

where E_t denotes the embedding vector of a target column name and E_s denotes the embedding vector of a standard column name. The numerator represents the inner product of the

two vectors, while the denominator represents the product of their magnitudes. Through this matching process based on Equation (1), the module transforms non-standard column names into corresponding standard variables. The method goes beyond surface-level string comparison by reflecting contextual meaning and relationships across data, which produces refined and reliable results. This approach addresses variation in expression and heterogeneity across sources, thereby enhancing both reliability and accuracy of data integration.

Figure 4 illustrates the matching process, including how non-standard column names align with standard column names based on cosine similarity. The documents unified with standard column names through the proposed model serve as inputs for subsequent simulations that support decision-making in flood response strategies.

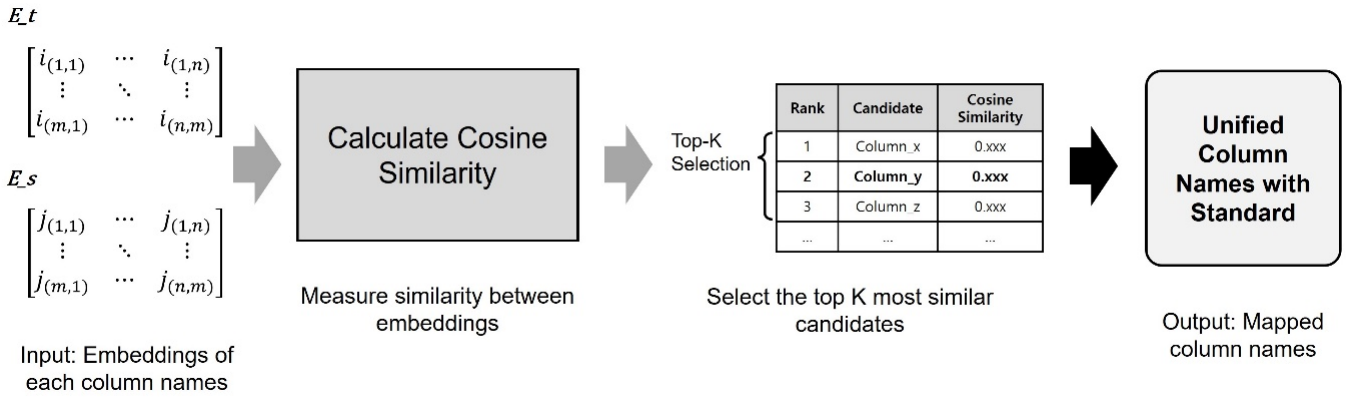


Figure 4. Description of the automatic matching module.

Algorithm 1 presents the three main stages of the proposed model—text augmentation, embedding generation, and automatic matching—from an input–output perspective. The inputs consist of a set of standard column names S, a set of target column names T, a pre-trained generative language model G, and a BERT model B fine-tuned on disaster-domain data. The final output is the mapping result M, which aligns target column names with standard column names. Stage 1 corresponds to text augmentation. For each standard name s in the set S, the generative model G produces a variety of synonyms A_s. The framework adds them to the augmented standard column set $S_{augmented}$. Stage 2 converts both the augmented standard column names and the target column name into embedding vectors. The fine-tuned BERT model B computes embeddings E_s and E_t for each set. Stages 3 and 4 perform the actual mapping. For each target column name t, the framework calculates cosine similarity with all standard names, assigning t to the standard name s that yields the highest similarity score and thus producing the final mapping result M. Through this procedure, the model improves the accuracy of disaster data integration by applying a Top-K matching strategy to derive optimal results.

Computational complexity. Let $|S_{aug}|$ be the number of standardized name strings after augmentation and $|T|$ the number of target attribute names. After embedding generation, the retrieval stage computes cosine similarity between each target vector and all standard vectors, which is $O(|T||S_{aug}|d)$ for d-dimensional embeddings. Because embeddings for S_{aug} can be precomputed offline, integrating a new municipal table requires encoding $|T|$ strings and a similarity search over the cached standard vectors.

Algorithm 1: LLM-based Schema Matching for Flood Data

Input:
- S: Set of standard column names.
- T: Set of target column names.
- G: Generative language model.
- B: Fine-tuned BERT model.

Output:
- M: Mapping of target column names to standard column names.

Steps:
1. Text Augmentation:
   For each $s \in S$:
      (a) Generate augmented variations A_s using G.
      (b) Add A_s to $S_{augmented}$.
2. Embedding Generation:
   (a) Compute embeddings E_s for each $s \in S_{augmented}$ using B.
   (b) Compute embeddings E_t for each $t \in T$ using B.
3. Similarity Calculation:
   For each $t \in T$:
      (a) Calculate cosine similarity between E_t and E_s.
      (b) Identify the top-k similar s based on similarity scores.
4. Schema Mapping:
   For each $t \in T$:
      Assign t to s with the highest similarity score.
      Update M with the mapping $t \rightarrow s$.

Return: M.

4. Empirical Study

4.1. Experimental Setting

To evaluate the performance of the proposed model, we conduct the task of column name matching between transformation tables and a standard table. The dataset consists of real flood-related data provided by the Korean municipalities of Busan and Incheon together with a table of standard variable names. To increase the diversity of variable expressions, this study combines the Busan and Incheon tables into a single dataset for evaluation. Table 2 presents the English and Korean representations of the standard variables used as the basis for integration. In this study, the standard schema consists of a fixed set of standardized attributes used for downstream analysis, while the target datasets contain municipality-specific (non-standard) attribute names collected from Busan and Incheon. The target names often include abbreviations, mixed Sino-Korean compounds, and local naming conventions, as illustrated by the examples in Table 1. We evaluate schema alignment by ranking standardized candidates for each target attribute and reporting whether the correct standardized attribute is retrieved within the top-k list.

Table 2. Examples of Standard Variable Names for Flood Data.

English Name	Korean Name/Explanation
Name	Conduit Name (관거 이름)
FromNode	Start Node of Conduit (관거 시작노드)
ToNode	End Node of Conduit (관거 종점노드)
Shape	Conduit Shape (관거 모양)
Thickness	Conduit Thickness (관거 두께)
Diameter	Conduit Diameter (관거 지름)
SedimentHeight	Sediment Height (퇴적물의 높이)
Roughness	Roughness Coefficient (거칠기 계수)
Length	Conduit Length (관거 길이)
InitialFlow	Initial Flow (초기 유량)
MaximumFlow	Maximum Flow (최대 유량)
SeepageLossRate	Seepage Loss Rate (누수 속도)
FlapGate	Presence of Flap Gate (플랩 게이트 존재 여부)

The evaluation employs multiple metrics, including Hit@k, Precision, Recall, and Mean Reciprocal Rank (MRR). Hit@k (equivalently Acc@k) measures whether the correct standard variable appears within the top-k predictions. Precision and Recall capture the trade-off between accuracy and coverage of matches. MRR reflects the average rank quality of the correct variable in the candidate list. These metrics collectively assess not only the accuracy of individual matches but also the robustness of the framework in handling diverse non-standard variable names. Equations (2)–(5) describe the metrics.

$$Hit@k = \frac{1}{N}\sum_{i=1}^{N} 1\left\{ y_i \in \ Top_{k(\hat{y}_i)} \right\} \tag{2}$$

where N is the total number of target variables, y_i is the ground-truth standard variable for instance i, $\hat{y}_i$ is the predicted ranking list, and $Top_{k(\hat{y}_i)}$ denotes the top-k predictions.

$$Precision = \frac{TP}{(TP + FP)} \tag{3}$$

$$Recall = \frac{TP}{(TP + FN)} \tag{4}$$

where TP is the number of true positives, FP is the number of false positives and FN is the number of false negatives.

$$MRR = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{rank_i}\right) \tag{5}$$

where $rank_i$ is the rank position of the correct standard variable in the prediction list for instance i.

The comparison involves Word2Vec and fine-tuned BERT models. To verify the effectiveness of the proposed text augmentation technique, the evaluation examines performance differences with and without the augmentation module. The following describes the baseline models.

Word2Vec [33]: Word2Vec learns embeddings through Skip-gram and CBOW, taking surrounding context into account. It captures local context by reflecting context-dependent word meaning, but its performance remains limited compared with recent models such as BERT.

LaBSE [34]: LaBSE trains on 109 languages by combining masked language modeling, translation language modeling, and translation ranking. It demonstrates strong performance in multilingual sentence retrieval and semantic similarity tasks.

BERT_finetuned: BERT [27] learns bidirectional context to produce embeddings that reflect semantic meaning. We use this fine-tuned encoder as the fine-tuning-only baseline (BERT_fin) with the same cosine-similarity retrieval setup, and compare it with the proposed model, which adds the text-augmentation module on top of the same adapted encoder; thus, the comparison isolates the incremental contribution of augmentation over fine-tuning.

We used a Korean BERT encoder as the backbone of the embedding module and performed domain adaptation via fine-tuning on flood-related text. For the encoder fine-tuning, we set the maximum sequence length to 300 and used cosine similarity for retrieval. We report Hit@k with $k \in \{1, 3, 5\}$ and additionally report Precision, Recall, and MRR to reflect candidate quality and coverage in downstream integration.

Implementation details. In our experiments, the generative model G is GPT-4o. Decoding parameters were left at their default values. For each standardized attribute, we generate three short alternatives in a column-name style using a fixed prompt template that provides the standardized name and requests a comma-separated list while avoiding the introduction of additional concepts. We remove duplicates and long candidates by simple post-processing. To control semantic drift, each candidate is embedded with the same flood-domain-adapted Korean BERT encoder used for matching (Section 3.4), and candidates with cosine similarity below $\tau = 0.70$ with the source name are discarded. The remaining candidates are added to $S_{augmented}$ and used only as additional anchors during top-k retrieval; the downstream standardized schema remains unchanged.

Standard schema and datasets. The standardized schema S contains 17 sewer-network attributes required to build SWMM-based flood simulation inputs (examples are shown in Table 2). The target datasets contain municipality-specific column names extracted from sewer-network tables provided by Busan (11 columns) and Incheon (44 columns). Domain experts annotated which target columns correspond to the standardized schema; columns marked as unrelated were excluded from Hit@k evaluation. To increase linguistic diversity of non-standard expressions, the Busan and Incheon target sets are merged for the main Hit@k evaluation in Section 4.2, while the Busan dataset is used for the end-to-end case study in Section 4.3.

Encoder fine-tuning. The embedding backbone is adapted to the flood domain using unsupervised SimCSE training on a Korean flood-related corpus assembled from public technical documents and flood-related news articles. The corpus was prepared as a plain-text file with one sentence per line after basic cleaning (character normalization, whitespace normalization, and deduplication). Fine-tuning was run for 3005 steps (one epoch) with per-device batch size 16, learning rate 2×10^{-5}, maximum sequence length 300, and SimCSE temperature 0.07, using FP16. We use CLS pooling and train the SimCSE MLP only.

4.2. Experimental Evaluation

Each model computes embeddings for the input data, with an evaluation of the similarity between non-standard and standard variable names using cosine similarity. The evaluation metric is Hit@k, which measures the proportion of cases where a standard variable appears within the top k ∈ {1, 3, 5} candidates. Tables 3 and 4 present examples of variable names from the Incheon dataset, showing the results of applying the proposed model compared with the baseline BERT. In these results, the proposed approach successfully aligns variables such as 'Pipe Diameter (관경)' and 'Length (연장)', which fine-tuned BERT does not correctly transform.

https://doi.org/10.3390/systems14030267

Table 3. Schema matching with fine-tuned BERT.

Target Variable Name	Transformed Name (Top-3 Similarity)	Standard Variable Name	Correct Matching
Unique Identifier (고유 식별자)	Conduit Name (관의 이름) End Node of Conduit (관의 종점노드) Roughness Coefficient (거칠기 계수)	Conduit Name (관의 이름)	Y
Structure Type (구조물 형태)	Conduit Shape (관의 모양) Sediment Height (퇴적물의 높이) Roughness Coefficient (거칠기 계수)	Conduit Shape (관의 모양)	Y
Pipe Diameter (관경)	Roughness Coefficient (거칠기 계수) Conduit Length (관의 길이) Conduit Shape (관의 모양)	Conduit Diameter (관거 지름)	N
Material texture (재질)	Roughness Coefficient (거칠기 계수) Conduit Thickness (관의 두께) Conduit Shape (관의 모양)	Roughness Coefficient (거칠기 계수)	Y
Length (연장)	Roughness Coefficient (거칠기 계수) Conduit Thickness (관의 두께) Conduit Shape (관의 모양)	Conduit Length (관의 길이)	N
Dry-weather Flow Velocity (청천시 유속)	Initial Flow (초기 유량) Maximum Flow (최대 유량) Sediment Height (퇴적물의 높이)	Initial Flow (초기 유량)	Y

Table 4. Schema matching with proposed model.

Target Variable Name	Transformed Name (Top-3 Similarity)	Standard Variable Name	Correct Matching
Unique Identifier (고유 식별자)	Conduit Name (관의 이름) Conduit Shape (관의 모양) Conduit Thickness (관의 두께)	Conduit Name (관의 이름)	Y
Structure Type (구조물 형태)	Conduit Shape (관의 모양) Sediment Height (퇴적물의 높이) Conduit Diameter (관거 지름)	Conduit Shape (관의 모양)	Y

Table 4. *Cont.*

Target Variable Name	Transformed Name (Top-3 Similarity)	Standard Variable Name	Correct Matching
Pipe Diameter (관경)	Roughness Coefficient (거칠기 계수) Conduit Diameter (관거 지름) Conduit Shape (관의 모양)	Conduit Diameter (관거 지름)	Y
Material texture (재질)	Roughness Coefficient (거칠기 계수) Conduit Thickness (관의 두께) Initial Flow (초기 유량)	Roughness Coefficient (거칠기 계수)	Y
Length (연장)	Roughness Coefficient (거칠기 계수) Conduit Length (관의 길이) Conduit Shape (관의 모양)	Conduit Length (관의 길이)	Y
Dry-weather Flow Velocity (청천시 유속)	Initial Flow (초기 유량) Maximum Flow (최대 유량) Sediment Height (퇴적물의 높이)	Initial Flow (초기 유량)	Y

Figure 5 compares matching accuracy across models. In the combined Busan–Incheon evaluation, the fine-tuned BERT baseline achieves Hit@1/3/5 of 0.43/0.57/0.71, whereas the proposed model with augmentation achieves 0.33/0.86/0.95. These results indicate that augmentation substantially improves retrieval accuracy at higher k. An interpretation of the top-1 vs. top-k trade-off and representative failure cases is provided in Section 4.4.

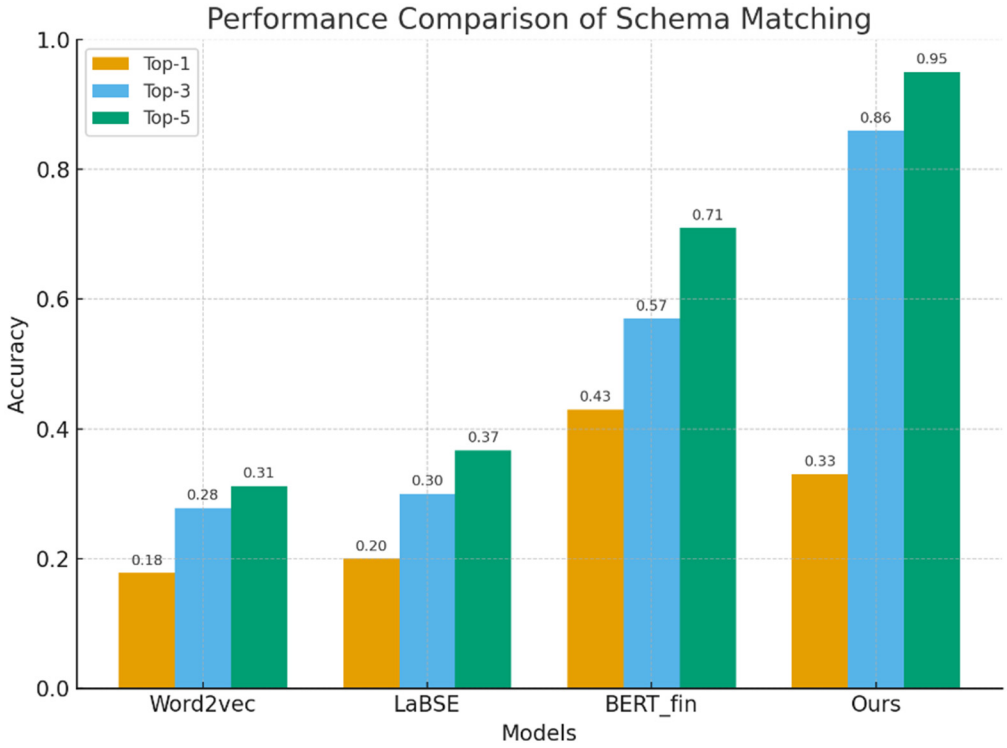


Figure 5. Performance comparison of the existing word embedding models and our model.

Despite these limitations, text augmentation substantially improves accuracy at top-3 and top-5. The model exhibits greater flexibility and robustness when handling non-standard variable names with diverse expressions and complex contexts. The results in Tables 3 and 4 and Figure 5 indicate that the augmentation strategy enables the language model to learn semantic similarity with standard variables across a broader range. Without complex modifications, the proposed model demonstrates superior performance over existing language models in standardizing non-standard variable names.

4.3. Case Study

This case study applies the proposed schema matching model to flood data provided by Busan Metropolitan City to demonstrate its applicability in real disaster response scenarios. The simulation assumes inland flooding caused by heavy rainfall at intensities of 50, 105, and 150 mm per hour. The simulation requires essential data attributes listed in Table 2. The proposed model transforms raw data from the municipality into the standardized schema. As shown in Table 4, our framework integrates the transformed dataset, which supports scenario-based disaster response.

Figure 6 illustrates the simulation outcome for the case of 105 mm per hour rainfall in Busan. Blue markers indicate CCTV locations, green markers denote evacuation shelters, and red areas represent inundated regions. The prediction of flood-affected zones provides critical information for identifying potential risks and supporting early response planning. Figure 7 further demonstrates how inundation estimates contribute to the identification of evacuation routes. Yellow, light green, and red regions indicate flooded areas, where brighter colors represent shallow inundation and darker shades indicate deeper water. Orange markers designate starting points, and blue markers denote destinations. These outputs illustrate that, under severe rainfall conditions, individuals located in heavily flooded zones can be guided toward accessible shelters through optimized evacuation paths, thereby reducing potential casualties.

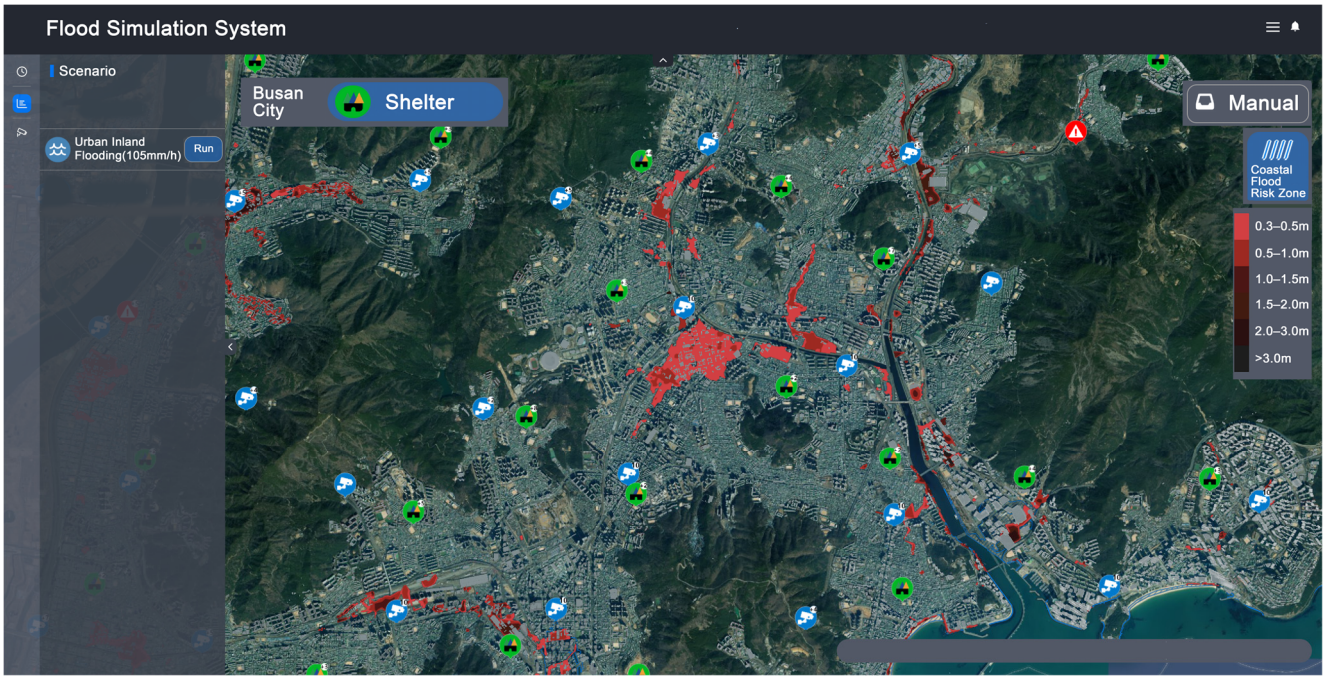


Figure 6. Example of flood simulation using automatically matched data.



Figure 7. Example of evacuation planning based on flood simulation results.

While Section 4.2 focuses on controlled experimental evaluation with Hit@k, the case study requires metrics that better capture operational reliability. In disaster simulations, missing critical attributes may distort outcomes, highlighting the importance of Recall. Conversely, incorrect mappings may propagate errors throughout the pipeline, making Precision equally critical. Table 5 summarizes schema matching performance on the Busan dataset using Precision, Recall, and MRR.

Table 5. Summary of column matching performance (Busan city).

Model	Hit@1	Hit@3	Precision	Recall	MRR	Note
Rule-based Method	0.044	0.133	0.066	0.255	0.107	Dictionary rules (synonyms)
KoSBERT	0.144	0.267	0.101	0.255	0.200	General KoSBERT model
BERT_fin	0.450	0.650	0.217	0.590	0.533	Fine-tuned on flood corpus
Ours (fin+aug)	0.526	0.737	0.246	0.700	0.614	Data augmentation (3 types)

The results in Table 5 demonstrate that the proposed model (finetuned BERT with data augmentation) achieves the highest performance across most evaluation metrics. In particular, it records a Recall of 0.700, which surpasses the fine-tuned BERT baseline (0.590) and indicates stronger robustness in capturing diverse expressions of non-standard variable names. The model also improves Hit@1 and Hit@3 (0.526 and 0.737, respectively) compared to the baseline (0.450 and 0.650), confirming that the augmentation strategy contributes to more accurate candidate selection. While Precision increases moderately (0.246 versus 0.217), the model maintains a balanced trade-off between accuracy and reliability, ensuring that matched variables remain useful for downstream flood simulations. Furthermore, the higher MRR (0.614 compared to 0.533) highlights its ability to rank correct matches more consistently at the top. These findings validate that the proposed schema matching framework not only aligns attributes more effectively than conventional baselines but also enhances the reliability of integrated data for flood simulation and evacuation planning in real-world contexts.

4.4. Discussion

Section 4.2 shows that domain adaptation and lexicon expansion play complementary roles in schema matching under heterogeneous Korean municipal naming. As reported in Figure 5, adding augmentation substantially improves top-k retrieval (Hit@3/Hit@5: 0.57/0.71 → 0.86/0.95), even though Hit@1 decreases (0.43 → 0.33). This behavior is practically meaningful because real integration pipelines often follow a shortlist-and-verify workflow, where the system proposes a small candidate set and a validator confirms the final mapping.

The top-1 vs. top-k trade-off is consistent with augmentation increasing near-neighbor anchors and occasionally introducing noise or overly general variants, which can reshuffle the highest-ranked candidate for short or abbreviated attribute names. This is visible in Table 3, where the baseline ranks "Roughness Coefficient" near "Pipe Diameter" and fails to return the correct mapping within the top-3, highlighting ambiguity when only the name string is available. When additional schema context exists, incorporating it would likely reduce such ambiguous mismatches. In this workflow, the main sensitivity knob is the augmentation filtering threshold τ, which controls the precision–coverage balance of generated variants. A lower τ retains more diverse expressions and tends to increase Hit@k at higher k, while a higher τ is more conservative and can stabilize the top-1 candidate. We therefore interpret the reported results as supporting top-k shortlist retrieval for practical schema integration, where validation resolves residual ambiguity.

From a systems perspective, the workflow is implementable as a repeatable pipeline: offline augmentation/filtering of the standardized schema, embedding precomputation for the augmented anchors, and online top-k retrieval by cosine similarity. This design keeps matching auditable as an explicit ranked candidate list and avoids invoking a fully generative LLM at inference time for every attribute, while still benefiting from LLM-assisted lexical coverage in the offline stage. In this setting, the main operational knobs control the precision–coverage balance and the observed top-1 vs. top-k behavior.

In the case study, the proposed model improves operationally relevant reliability metrics on the Busan dataset, supporting the motivation that higher coverage and better ranking reduce manual harmonization burden and help prevent missing critical inputs for downstream simulation pipelines. More broadly, the approach is transferable by replacing the target standard schema and adapting the encoder to the target domain/language; however, our evaluation currently covers two municipalities, and broader multi-city benchmarks would further test generalizability.

5. Conclusions

This study presents a novel approach that applies a BERT-based entity matching model to integrate heterogeneous flood data automatically. The framework addresses variability in data representation and semantic inconsistency through text augmentation with generative language models and the domain-specific fine-tuning of Korean BERT. It demonstrates substantial improvements in matching accuracy, even in environments characterized by complex and non-standard variable names. The proposed model not only simplifies the data integration process required for flood simulations but also provides reliable data that enhance decision-making in disaster response planning. A case study using data from the Busan municipality illustrates the practical applicability of the framework. Future research will extend this model to other disaster scenarios or incorporate additional data sources to reinforce adaptability and robustness. By resolving the complexity of heterogeneous data integration, this study contributes to automated data unification and effective strategy development in disaster management.

https://doi.org/10.3390/systems14030267

Author Contributions: Conceptualization, T.C. and M.K.; methodology, T.C. and M.S.; software, T.C.; validation, T.C., M.S. and M.K.; formal analysis, T.C.; investigation, M.S., K.K. and M.Y.; resources, M.K., K.K. and M.Y.; data curation, M.K., K.K. and M.Y.; writing—original draft preparation, T.C.; writing—review and editing, M.K. and K.L.M.; visualization, M.S., K.K. and M.Y.; supervision, M.K. and K.L.M.; project administration, M.K. and K.L.M. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2025-02217071) and also by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (No. RS-2025-02305436; Development of Digital Innovative Element Technologies for Rapid Prediction of Potential Complex Disasters and Continuous Disaster Prevention).

Data Availability Statement: Data available on request due to restrictions.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Jiang, Y.; Zevenbergen, C.; Ma, Y. Urban Pluvial Flooding and Stormwater Management: A Contemporary Review of China's Challenges and "Sponge Cities" Strategy. *Env. Sci. Policy* **2018**, *80*, 132–143.
2. Westra, S.; Fowler, H.J.; Evans, J.P.; Alexander, L.V.; Berg, P.; Johnson, F.; Kendon, E.J.; Lenderink, G.; Roberts, N. Future Changes to the Intensity and Frequency of Short-duration Extreme Rainfall. *Rev. Geophys.* **2014**, *52*, 522–555. [CrossRef]
3. Jongman, B.; Ward, P.J.; Aerts, J.C.J.H. Global Exposure to River and Coastal Flooding: Long Term Trends and Changes. *Glob. Environ. Chang.* **2012**, *22*, 823–835. [CrossRef]
4. Li, C.; Sun, N.; Lu, Y.; Guo, B.; Wang, Y.; Sun, X.; Yao, Y. Review on Urban Flood Risk Assessment. *Sustainability* **2022**, *15*, 765. [CrossRef]
5. Rahm, E.; Bernstein, P.A. A Survey of Approaches to Automatic Schema Matching. *VLDB J.* **2001**, *10*, 334–350. [CrossRef]
6. Shvaiko, P.; Euzenat, J. Ontology Matching: State of the Art and Future Challenges. *IEEE Trans. Knowl. Data Eng.* **2011**, *25*, 158–176. [CrossRef]
7. Wu, Z.; Shen, Y.; Wang, H.; Wu, M. An Ontology-Based Framework for Heterogeneous Data Management and Its Application for Urban Flood Disasters. *Earth Sci. Inf.* **2020**, *13*, 377–390.
8. Koutras, C.; Fragkoulis, M.; Katsifodimos, A.; Lofi, C. REMA: Graph Embeddings-Based Relational Schema Matching. In Proceedings of the EDBT/ICDT Workshops, Copenhagen, Denmark, 30 March 2020; p. 17.
9. Hättasch, B.; Truong-Ngoc, M.; Schmidt, A.; Binnig, C. It's AI Match: A Two-Step Approach for Schema Matching Using Embeddings. *arXiv* **2022**, arXiv:2203.04366.
10. Ayala, D.; Hernández, I.; Ruiz, D.; Rahm, E. Leapme: Learning-Based Property Matching with Embeddings. *Data Knowl. Eng.* **2022**, *137*, 101943. [CrossRef]
11. Parciak, M.; Vandevoort, B.; Neven, F.; Peeters, L.M.; Vansummeren, S. Schema Matching with Large Language Models: An Experimental Study. *arXiv* **2024**, arXiv:2407.11852. [CrossRef]
12. Oh, H.; Jones, A.; Finin, T. Employing Word-Embedding for Schema Matching in Standard Lifecycle Management. *J. Ind. Inf. Integr.* **2024**, *38*, 100547.
13. Sheetrit, E.; Brief, M.; Mishaeli, M.; Elisha, O. Rematch: Retrieval Enhanced Schema Matching with LLMs. *arXiv* **2024**, arXiv:2403.01567. [CrossRef]
14. Kired, N.E.; Ravat, F.; Song, J.; Teste, O. Embedding-Based Data Matching for Disparate Data Sources. In *Proceedings of the International Conference on Big Data Analytics and Knowledge Discovery*; Springer: Berlin/Heidelberg, Germany, 2024; pp. 66–71.
15. Doan, A.; Halevy, A.Y. Semantic Integration Research in the Database Community: A Brief Survey. *AI Mag.* **2005**, *26*, 83.
16. Bellahsene, Z.; Bonifati, A.; Rahm, E. *Schema Matching and Mapping*, 1st ed.; Bellahsene, Z., Bonifati, A., Rahm, E., Eds.; Springer: Berlin/Heidelberg, Germany, 2011; ISBN 978-3-642-16517-7.
17. Do, H.-H.; Rahm, E. COMA—A System for Flexible Combination of Schema Matching Approaches. In *Proceedings of the VLDB'02: Proceedings of the 28th International Conference on Very Large Databases*; Elsevier: Amsterdam, The Netherlands, 2002; pp. 610–621.
18. Elmagarmid, A.K.; Ipeirotis, P.G.; Verykios, V.S. Duplicate Record Detection: A Survey. *IEEE Trans. Knowl. Data Eng.* **2006**, *19*, 1–16. [CrossRef]
19. Kedad, Z.; Xue, X. Mapping Discovery for XML Data Integration. In *Proceedings of the OTM Confederated International Conferences "On the Move to Meaningful Internet Systems"*; Springer: Berlin/Heidelberg, Germany, 2005; pp. 166–182.
20. Chen, C.; Golshan, B.; Halevy, A.Y.; Tan, W.-C.; Doan, A. BigGorilla: An Open-Source Ecosystem for Data Preparation and Integration. *IEEE Data Eng. Bull.* **2018**, *41*, 10–22.

21. Rashid, S.M.; McCusker, J.P.; Pinheiro, P.; Bax, M.P.; Santos, H.O.; Stingone, J.A.; Das, A.K.; McGuinness, D.L. The Semantic Data Dictionary–an Approach for Describing and Annotating Data. *Data Intell.* **2020**, *2*, 443–486. [CrossRef] [PubMed]
22. Asif-Ur-Rahman, M.; Hossain, B.A.; Bewong, M.; Islam, M.Z.; Zhao, Y.; Groves, J.; Judith, R. A Semi-Automated Hybrid Schema Matching Framework for Vegetation Data Integration. *Expert. Syst. Appl.* **2023**, *229*, 120405.
23. Wu, K.; Zhang, J.; Ho, J.C. CONSchema: Schema Matching with Semantics and Constraints. In *Proceedings of the European Conference on Advances in Databases and Information Systems*; Springer: Berlin/Heidelberg, Germany, 2023; pp. 231–241.
24. Pan, Z.; Pan, G.; Monti, A. Semantic-Similarity-Based Schema Matching for Management of Building Energy Data. *Energies* **2022**, *15*, 8894. [CrossRef]
25. Mukherjee, D.; Bandyopadhyay, A.; Chowdhury, R.; Bhattacharya, I. Learning Knowledge Graph for Target-Driven Schema Matching. In *Proceedings of the 3rd ACM India Joint International Conference on Data Science & Management of Data (8th ACM IKDD CODS & 26th COMAD)*; ACM: New York, NY, USA, 2021; pp. 65–73.
26. Narayan, A.; Chami, I.; Orr, L.; Arora, S.; Ré, C. Can Foundation Models Wrangle Your Data? *arXiv* **2022**, arXiv:2205.09911. [CrossRef]
27. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*; Association for Computational Linguistics: Minneapolis, MN, USA, 2019; pp. 4171–4186.
28. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. *Adv. Neural Inf. Process Syst.* **2017**, *30*.
29. Pires, T.; Schlinger, E.; Garrette, D. How Multilingual Is Multilingual BERT? *arXiv* **2019**, arXiv:1906.01502. [CrossRef]
30. Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C.H.; Kang, J. BioBERT: A Pre-Trained Biomedical Language Representation Model for Biomedical Text Mining. *Bioinformatics* **2020**, *36*, 1234–1240. [CrossRef]
31. Lee, S.; Jang, H.; Baik, Y.; Park, S.; Shin, H. Kr-Bert: A Small-Scale Korean-Specific Language Model. *arXiv* **2020**, arXiv:2008.03979.
32. Brain, S.K.T. KoBERT: Korean BERT Pre-Trained Cased 2019. GitHub repository. Available online: https://github.com/SKTBrain/KoBERT (accessed on 10 February 2026).
33. Church, K.W. Word2Vec. *Nat. Lang. Eng.* **2017**, *23*, 155–162.
34. Feng, F.; Yang, Y.; Cer, D.; Arivazhagan, N.; Wang, W. Language-Agnostic BERT Sentence Embedding. In *Proceedings of the Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*; Association for Computational Linguistics: Dublin, Ireland, 2022; pp. 878–891.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.