Systems and Soft Computing 7 (2025) 200373

Contents lists available at ScienceDirect

Systems and Soft Computing

journal homepage: www.elsevier.com/locate/soft-computing-letters

Enhancing consistency in piping and instrumentation diagrams using DistilBERT and smart PID systems

F.S. Gómez-Vega [a], O. Acuña [a], Andrea C. Camargo [a], Jeison D. Jimenez [a,c], Sara M. Galeano [a], Isabella E. Franco [a], Laura L. Lozano [a], Jenifer Vásquez [b], Edwin Puertas [c,*]

[a] Department of Innovation, SERINGTEC S.A.S, Km. 5 Vía Mamonal - Complejo Logístico del Caribe (CLC), Local 11, Cartagena, 130013, Bolivar, Colombia
[b] School of Engineering, Architecture & Design, Universidad Tecnológica de Bolívar, Km 1 via a Turbaco, Turbaco, 130001, Bolivar, Colombia
[c] School of Digital Transformation, Universidad Tecnológica de Bolívar, Km 1 via a Turbaco, Turbaco, 130001, Bolivar, Colombia

ARTICLE INFO

Keywords:
Databases
P&ID
Consistency
DistilBERT
Machine learning
Design software

ABSTRACT

This study presents a novel approach utilizing DistilBERT, a lightweight variant of BERT, to identify inconsistencies in piping and instrumentation diagrams (P&IDs) within SmartPID systems. A structured dataset was constructed by extracting engineering design data from a SQL-based SmartPID database, monitoring all modifications and updates made throughout the design phase. The DistilBERT model was fine-tuned on this dataset to recognize inconsistencies in real-time, achieving an impressive F1 score of 99% and a loss of 0.04%. The model's performance was validated by domain experts, who confirmed the detected inconsistencies as highly accurate. Our approach significantly reduces the manual effort required for P&ID review and improves design consistency, demonstrating the potential for enhanced safety and efficiency in complex industrial projects. Future work will focus on refining the model's parameters and expanding its application across different industries.

1. Introduction

In the engineering industry, the accuracy and consistency of piping and instrumentation diagrams (P&IDs) are of paramount importance for the safety and success of complex industrial projects. P&IDs are instrumental in visualizing the design of process systems; however, their manual review is both time-consuming and susceptible to error, frequently resulting in design inconsistencies that may lead to costly delays and safety risks. The conventional approach to reviewing P&IDs places considerable reliance on the expertise of technical personnel, which can lead to the introduction of human error and inefficiency into the process.

Recent advancements in machine learning (ML) and deep learning (DL) offer significant potential for the automation and enhancement of this review process. In particular, the advent of transformer-based models, such as BERT [1], has transformed the domain of natural language processing (NLP), facilitating machines' comprehension of contextual and semantic nuances in textual data. In this context, DistilBERT [2]—a variant of BERT that is both more lightweight and faster—has emerged as a promising tool for detecting textual inconsistencies while maintaining much of the performance of the original BERT model.

In recent years, there has been a notable increase in the adoption of artificial intelligence (AI) in various industries, including engineering, which has been actively supported by government initiatives. This trend has been reinforced by government programs at the departmental level in Colombia and the national level through the Ministry of Information and Communications Technology (MINTIC). These initiatives encourage the use of AI technologies to optimize operational activities and improve efficiency across sectors. However, to successfully implement AI in the industrial field, it is crucial to collate, interpret, and utilize the vast amounts of data generated in this sector to identify the specific requirements of each company and create innovative, bespoke solutions.

In this study, we employ DistilBERT to identify inconsistencies in P&ID designs stored in a SmartPID system. SmartPID, which forms part of the Hexagon PPM suite, is a widely used software tool in the field of industrial engineering, employed for the design and management of P&IDs. The proposed approach automates the identification of design inconsistencies by integrating a structured dataset extracted from SmartPID's SQL-based database, which records all changes made during the design process. DistilBERT was fine-tuned on this dataset to detect potential errors in real time, thereby markedly enhancing the speed and accuracy of the review process.

* Corresponding author.
E-mail address: epuerta@utb.edu.co (E. Puertas).

https://doi.org/10.1016/j.sasc.2025.200373
Received 16 October 2024; Received in revised form 25 June 2025; Accepted 30 July 2025
Available online 9 August 2025

While machine learning has been successfully applied to object detection and feature recognition tasks in P&IDs, the majority of these approaches concentrate on the visual aspects of the diagrams. To date, there have been few studies that have addressed the challenge of detecting textual inconsistencies within P&ID databases. Our work addresses this gap by combining natural language processing with industrial design, thereby providing a robust solution to enhance P&ID consistency checks.

The following is a description of the structure of this paper: Section 2 presents a review of related work in the application of artificial intelligence to process and instrument diagram (P&ID) databases. Section 3 outlines the methodology and details the construction of the dataset. Section 4 provides an account of the experimental results and the performance of the model. Sections 5 and 6 discuss the implications of the findings and conclude with an outline of future research directions.

2. Related work

A research opportunity has been identified in the scientific literature on the topic of consistent detection in databases for P&ID applications. Following a comprehensive review of over 100 articles employing the ProKnow-C methodology [3,4], it became apparent that the majority of research efforts have been concentrated in the domain of computer vision, particularly in the areas of instrumentation and piping diagrams. However, research has been largely centered on object detection and feature recognition within P&IDs, with other critical aspects, such as the detection of textual inconsistencies, receiving comparatively little attention.

Deep learning models such as YOLO (You Only Look Once) have been effectively utilized to automatically detect symbols and components within P&ID layouts, as evidenced by the literature [5–7]. Although these models demonstrate proficiency in identifying visual elements, they exhibit limitations in discerning the intricate variations observed across different industries, particularly in the context of complex layouts and symbol variations. This gap in the literature indicates a need for more sophisticated approaches that not only focus on visual elements but also address the textual and relational aspects of P&IDs [8–13].

Despite the increasing prevalence of natural language processing (NLP) in a range of industries, its deployment for the identification of inconsistencies in technical documents, such as those found in P&ID databases, has remained relatively limited. NLP models such as BERT have been demonstrated to be effective in tasks such as text classification and error detection in fields including finance and healthcare. These models have exhibited robust capabilities in comprehending the context and semantics of technical texts. Nevertheless, there has been a paucity of research investigating the utilization of such models in the domain of industrial engineering, particularly in the context of identifying inconsistencies within databases. Our work exploits the strengths of transformer-based models, adapting them to the particular requirements of P&ID systems, where text and data must be aligned to guarantee design consistency [14].

2.1. Adaptation to the oil and gas industry

Beyond P&ID-related applications, AI techniques have also been explored in other areas more closely linked to natural language processing, such as address validation and material property data management. These approaches, although not originally developed for P&ID systems, can be adapted to benefit industries like oil and gas, specifically for our project. In the context of address validation, for instance, the use of RoBERTa has significantly improved the ability to handle polysemy and contextual ambiguities. The study ‘‘A RoBERTa-Based Approach for Address Validation’’ demonstrated advancements over traditional methods but highlighted ongoing challenges in parsing complex, multilingual address structures [15].

Similarly, artificial neural networks (ANNs) have been used for material property data validation and imputation, marking a significant improvement in database quality. However, the simplicity of the ANN models, often limited to single-layer architectures, has constrained their ability to capture complex, nonlinear relationships in large datasets, affecting the accuracy of imputed values. Despite these advancements, current methods face limitations when applied to diverse and complex inputs, such as intricate P&ID layouts or non-standardized address formats [16].

While (AI) techniques such as object detection, feature recognition, and address validation have proven useful in certain contexts, they often lack the necessary scalability and robustness for broader industrial applications. Furthermore, many of these solutions entail a compromise between model complexity and performance. In many cases, simpler models are unable to adequately capture the nuances of real-world data. It is therefore evident that more sophisticated approaches are required to meet the increasing demands of industries such as oil and gas [17–19].

2.2. DistilBERT as a solution

DistilBERT has emerged as a promising candidate for addressing these challenges. It provides a streamlined alternative to BERT, reducing the computational cost and memory consumption while largely maintaining BERT’s performance. In the process of knowledge distillation, DistilBERT learns to emulate the behavior of the larger BERT model by utilizing BERT’s outputs as soft targets during training. This approach permits the retention of BERT-level performance in a more efficient model that is well-suited to resource-constrained environments, such as those encountered in industrial settings [20,21].

In order to optimize the use of DistilBERT, a variety of transformer architectures have been evaluated based on both performance and resource requirements. The positive outcomes of these evaluations can be applied to enhance the consistency detection capabilities of our model. The objective of this project is to introduce a more sophisticated model that is capable of handling the inherent complexities of large, real-world datasets, thereby advancing AI-driven solutions for P&ID consistency detection [22].

2.3. Existing methods for P&ID analysis and validation

Existing approaches to P&ID analysis have predominantly emphasized digitization and visual recognition, often overlooking the detection of textual and relational inconsistencies within engineering databases. Recent studies employing Graph Convolutional Networks (GCNs) have achieved significant advancements, such as the use of GraphSAGE, which attains precision rates of 98.48% for binary text/non-text classification and 90.82% for three-class classification in mechanical engineering drawings [23]. Similarly, the OSSR-PID methodology utilizes Dynamic Graph Convolutional Neural Networks tailored explicitly for P&IDs, reporting an F1-score of 85.98% through one-shot learning techniques [24]. Nevertheless, these GCN-based methodologies inherently focus on graphical elements and neglect semantic relationships inherent in textual data, while also depending on intricate vectorial preprocessing, thus introducing potential inaccuracies.

Approaches to automatic digitization incorporating computer vision techniques have demonstrated varied success rates. For instance, a zero-shot method using Faster R-CNN and Siamese networks achieves detection accuracy of 75.4% and classification precision of 74.6% [25]. The Digitize-PID pipeline, integrating CRAFT for text detection, morphological operations, and deep learning models like FCN and TBMSL-Net, achieves symbol recognition F1-scores of 85.98% and line detection precision of 99.34% [26]. Advanced implementations improving

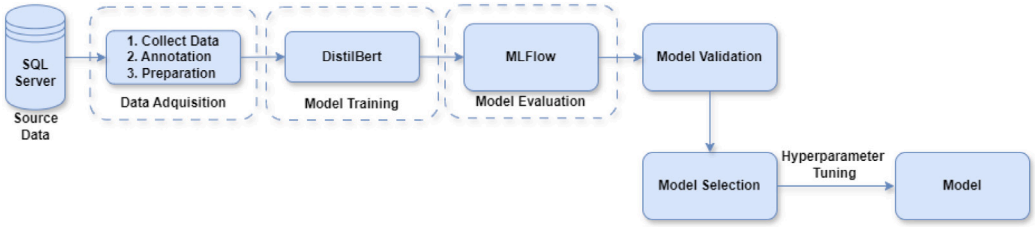

Fig. 1. Pipeline MLOps defined for the project.

upon Faster R-CNN have achieved up to 96.3% mAP by substituting VGG16 with ResNet101 for recognizing industrial control diagram elements [27]. Despite these advancements, their effectiveness remains heavily contingent on image quality, and they cannot assess logical consistency between interconnected components—essential for comprehensive P&ID validation.

Commercial solutions like Smart P&ID depend heavily on predefined rules for consistency verification, demonstrating significant rigidity and limited scalability to novel configurations and contextual variations.

Therefore, three key limitations emerge from existing methodologies: exclusive reliance on visual processing, inability to validate relational consistency comprehensively, and extensive dependency on manual intervention or large annotated datasets. Utilizing DistilBERT directly addresses these issues by performing semantic analyses on structured textual data from Smart P&ID systems. This reduces preprocessing errors and enhances the detection of complex semantic inconsistencies, such as discrepancies in properties among related components, which are otherwise undetectable through visual-only methodologies. Consequently, integrating DistilBERT and similar NLP techniques provides a robust and scalable alternative, addressing critical gaps identified in prior methods and significantly enhancing the semantic understanding necessary for effective P&ID validation.

3. Methodology

This section describes the environment, data acquisition, and methodology developed for the construction of the dataset and training of the AI model using DistilBERT for P&ID inconsistency detection. The methodology follows the MLOps pipeline illustrated in Fig. 1, which outlines the stages involved in transforming the model into a usable product for operational purposes within the company.

3.1. Research process

This work follows an action—research approach, involving multidisciplinary team members who actively participate in the development of the described solution. The team comprised nine professionals: an IT Project Manager, a Senior AI Consultant, a Junior Process Engineer, a Junior Mechatronics Engineer, a Junior Electronic Engineer, two Junior Civil Engineers, an Industrial Engineer, and a Junior Electrical Engineer. This diverse group ensured comprehensive coverage of both the technical aspects of model development and the evaluation criteria related to engineering drawings.

The validation of the deep learning (DL) model employed a mixed strategy. Initially, developers verified internal prediction consistency via unit testing on labeled datasets. Subsequently, senior engineering experts participated as validators, assessing model outputs in complex or critical cases. The validation targeted specific types of inconsistencies, including:

1. Labeling inconsistencies related to the standardization of symbols within P&ID diagrams.
2. Errors in accurately extracting and interpreting functional relationships between components.
3. Semantic ambiguities or inaccuracies impacting the engineering interpretation of the results.

The primary objective of the validation was to confirm label consistency with standardized symbols, followed by assessing accuracy in functional relationship extraction and ensuring semantic clarity from an engineering perspective—this expert intervention aimed to optimize resource utilization (man-hours) while maintaining high-quality standards.

The project timeline spanned January to August 2024 and included the implementation of a user interface (not discussed herein), enabling continuous interaction and validation by engineering experts. These iterative improvements align closely with MLOps methodologies, as illustrated in Fig. 1.

3.2. Dataset

The dataset was collected from a Smart P&ID system, which stores all the engineering design information in a SQL database. Smart P&ID is part of the Hexagon PPM suite, an advanced engineering tool that follows a data-centric approach. To build the dataset, we tracked all changes made in the design process, capturing data from every INSERT, UPDATE, and DELETE operation in the database. This Change Data Capture (CDC) approach allowed for the collection of over 100,000 rows of engineering data, focusing on key components such as equipment, instruments, labels, and piping runs. After gathering this data, preprocessing was applied to clean and prepare it for the model, ensuring that only relevant columns were included.

3.2.1. Data acquisition

For the collect data stage, the dataset created for the project consisted of the capture of engineering design information in the database, as shown in Fig. 2. This involved the identification and documentation of each element that composed the database, (created, modified, or update) [28,29]. This review was conducted methodically and rigorously, ensuring the accuracy and precision of the information captured.

For this purpose, the designer started using Smart P&ID and we tracked every change. Smart PID is an intelligent engineering design tool that is integrated into the Hexagon PPM application suite. It is an advanced software tool that is based on a data-centric model and engineering rules. This tool facilitates the creation and management of PIDs and also deploys a set of functionalities that optimize the efficiency and quality of engineering products. The first step was to identify all the components in the design shown in Fig. 3. This included equipment, instruments, notes, labels, piping runs and instrumentation signals. The aim was to identify the tables that corresponded to each of the components.

All detected changes were saved in a structured format as shown in Table 1, ensuring that each modification was properly documented where the first column called "table" represents the existing tables in the database, "column" is the specific column in which information update occurs within the tables (creation, modification, update), "logic" was defined as the column referring to the role of the table within the database, whether it is a domain table or a tool configuration table

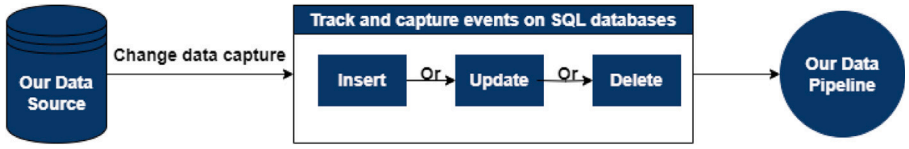

Fig. 2. Change Data Capture (CDC) extracts change events such as INSERTs, UPDATEs or DELETEs from data.

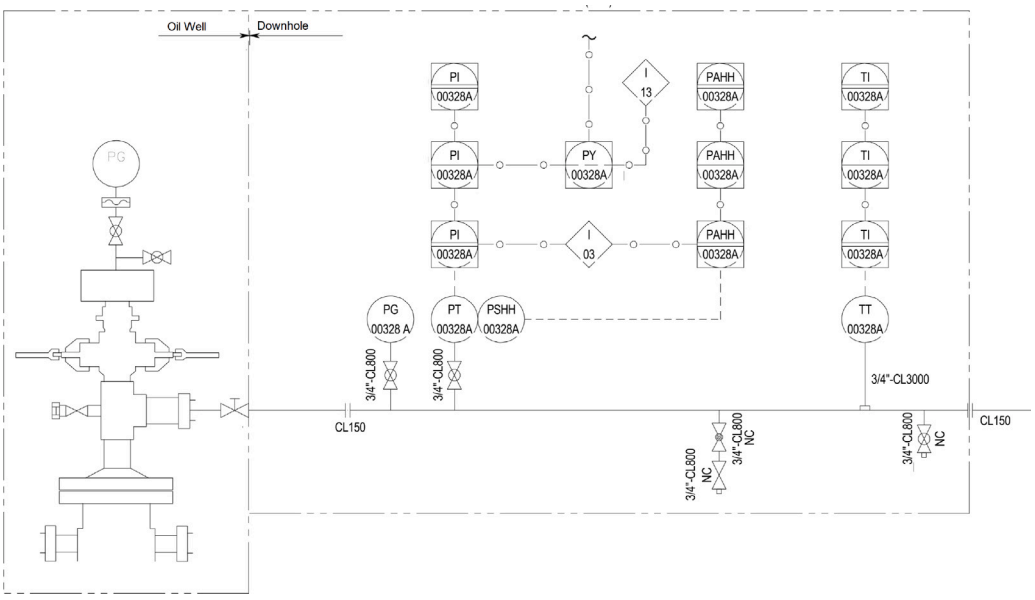

Fig. 3. P&ID designed for the crude oil extraction process for a client, here each of the most common components of a P&ID, equipment, instruments, notes, circuit labels (loop tags), pipe runs, signal lines.

Table 1
Extracted from the dictionary created to understand and build the views defined to acquire the dataset.

Table	Column	Data type	Max length	Logic	Description
T_Inconsistency	Name	nvarchar	160	Dominio	Records the inconsistencies generated in the drawing.
T_Inconsistency	Severity	nvarchar	4	Dominio	
T_Inconsistency	SP_ID	nvarchar	64	Dominio	
T_Inconsistency	SP_RelationshipID	nvarchar	64	Dominio	
T_Relationship	GraphicOID	nvarchar	4	Configuración	
T_Relationship	SP_DrawingID	nvarchar	64	Configuración	Records the types of relationships that are allowed in the project.
T_Relationship	SP_ID	nvarchar	64	Configuración	
T_Relationship	SP_Item1ID	nvarchar	64	Configuración	
T_Relationship	SP_Item2ID	nvarchar	64	Configuración	
T_Representation	GraphicOID	int	4	Configuración	
T_Representation	InStockpile	int	4	Configuración	Records the symbols added in the drawing and where each of these symbols are taken off from.
T_Representation	SP_DrawingID	nvarchar	64	Configuración	
T_Representation	SP_FileLastModifiedTime	datetime	8	Configuración	
T_Representation	SP_ID	nvarchar	64	Configuración	
T_Representation	SP_ModelItemID	nvarchar	64	Configuración	

and finally a general description of what each table corresponds to understand its role within the design process.

This exercise resulted in a data dictionary to understand the structure and meaning of the data stored in the SmartPID database. The data dictionary provides clear guidance on the relevant tables and their attributes, helping to understand the application logic and select the appropriate tables to create a view that will serve as a basis for training an artificial intelligence model.

Creating views is fundamental to focusing, simplifying, and personalizing the perception of data stored in the database. Views allow the definition of subsets of data that specifically reflect elements of interest for analysis or modeling. In this case, two main views have been defined: one focused on unresolved inconsistencies and another on the history of changes in the database. These views are constructed through SQL queries that combine several related tables to obtain the necessary information for model training.

The first view, shown in Fig. 4, focuses on the relationships between engineering elements and the graphical representations associated with these elements. The SQL query uses the T-History, T-Relationship, and T-Representation tables to obtain information about the relationships' location, type, and other attributes and graphical representations.

The second view, Fig. 5, focuses on unresolved inconsistencies between engineering elements. The SQL query combines the T-Relationship, T-Representation, and T-Inconsistency tables to obtain information about the inconsistencies, including their description, status, and severity. This view allows identifying the inconsistencies that need to be addressed and provides labeled data that can be used to train the

Fig. 4. View of T_History_Representation_Relationship.

Fig. 5. View of T_Inconsistency_Relationship_Representation.

artificial intelligence model in detecting and resolving problems in engineering design. Both views are fundamental to the training process of the artificial intelligence model, as they provide structured and relevant data that allow the model to learn patterns and relationships between engineering elements, as well as identify and resolve inconsistencies in design.

The dataset was constructed according to the previously selected views. This process resulted in a query that can be secured for future testing, aiming to make the process more robust. After verifying the functionalities in a test instance, this query was executed in a production environment. This allowed for the verification of both consistency and inconsistency on a large scale, using a greater amount of data. For this exercise, two projects that had been successfully completed by the client and one ongoing project were used as references.

3.2.2. Annotations

For the annotations, we made them take into account different inconsistencies that could show the tool as, "Inconsistent Property Value", which refers to a situation in which the value assigned to a property does not match the expected or required value according to the rules or standards of the system. On the other hand, the second type of inconsistency, called "Required Connection Point Unbound", indicates that a connection point, required to establish a link to another component, remains unbound, creating a design inconsistency.

The third type of inconsistency, called "Inconsistent flow direction", occurs when the direction of flow, whether in a pipe, circuit or any other matter or energy transport system, does not match the intended or designated direction according to the rules of the system. The fourth type of inconsistency, "No applicable standards", suggests that there are no relevant or applicable standards that can be used to assess the consistency of the design element in question. The fifth type of inconsistency, known as "Connector not attached", occurs when a connector that should be attached to other components is left unconnected.

The model can detect any discrepancies in the process lines, which are categorized as primary, secondary, or utility. Primary lines are defined as pipes that transport the main process fluid, thus constituting the primary flow route within the plant. Secondary lines are defined as those that branch off or connect to a primary line. In contrast, utility lines do not directly contribute to the primary process; rather, they furnish vital services such as water, steam, electricity, or gas to other equipment or processes. The model is principally designed to assess whether the information related to these different types of lines remains consistent across various components and representations.

To illustrate this point, consider a scenario in which a model identifies discrepancies in the property values of two pipe runs that appear to be related. These inconsistencies, categorized as 'Inconsistent Property Value', are identified when properties such as Nominal Diameter, Insulation Type, or Heat Trace Requirements deviate between the two items, despite the expectation of congruence. The comparison process systematically checks attributes such as insulation density, fluid code, or coating requirements and highlights any mismatches that violate predefined engineering rules. The model has been developed to address these inconsistencies by comparing corresponding properties and visually indicating any deviations that require engineering review. This capacity is imperative to preserve the integrity and consistency of data throughout the design and documentation of industrial systems.

Table 2
Training parameters of the DistilBERT model on Smart P&ID intelligent design databases.

Model/ Parameters	distil-pid-v1	distil-pid-v2
Epoch	1	1
Learning rate	5.00E−05	1.00E−05
Training Batch Size	–	16
Eval Batch Size	–	16
Optimizer	Adam	Adam
Eval steps	–	500
Step per Epoch	–	–
Validation Steps	–	500

3.3. Model training

Consequently, an information flow of over 100,000 rows was obtained, providing a substantial and reliable dataset essential for training the model. The data was subjected to preprocessing before its loading for training. This entailed the cleansing of the dataset to ensure its exclusive focus on the target columns, thereby precluding any potential interference with the model's performance. However, before the execution of the experiments, the model underwent training to ascertain its computational capabilities and to verify its ability to process more information. This was done to enhance the robustness of the dataset, as discussed at the outset of this text while avoiding the pitfall of overfitting the model, as evidenced in Table 2.

In the experiments, we use DistilBert to build a model that can detect inconsistencies in smart P&ID databases according to the given dataset. The initial learning rate was 1e-5, the epoch was 1 according to large dataset, for the training batch size we chose 16, and the evaluation batch size was 4. The experiments were conducted on a single machine with Intel i7-10700 CPU, 32G memory.

3.4. Model evaluation

This section presents our evaluation using MLflow, an open-source visualization tool designed for experiment tracking in machine learning. The library provides a streamlined approach to logging performance metrics, hyperparameters, and statistical characteristics of the data. Its importance within the machine learning community is reflected in its widespread adoption. MLflow supports the management of the machine learning lifecycle [30,31], offering access to experiment logs through either a web-based user interface or a programmatic API [32,33].

Table 3
Summary of DistilBERT model results on Smart P&ID intelligent design databases train by our team.

Model	distil-pid-v1	distil-pid-v2
Duration	1.1d	3.0d
epoch	1	1
eval_accuracy	0	0.99997
eval_f1	0	0.99995
eval_loss	0	0.00017
eval_precision	0	0.99994
eval_recall	0	0.99997
eval_runtime	–	9815.2047
eval_samples_per_second	–	3.445
eval_steps_per_second	–	0.215
grad_norm	1.8139E−05	0.00055
loss	0.0044	0.0004
total_flos	3.5209E+16	1.7916E+16
train_loss	0.00798821	0.00503
train_runtime	92 931.9346	262 155.162
train_samples_per_second	2.86	0.516
train_steps_per_second	0.357	0.032

In this way, the performance can be evaluated using the F1 score metric. The F1 score is calculated as each class's harmonic mean of precision and recall. The macro-F1 score represents the average across all classes. The loss function tracks the loss value, which the model aims to minimize during training. Lower loss values indicate superior model performance. Gradient norms are a crucial metric in this context, as they represent the direction and rate of change of the loss function concerning the model's parameters. Therefore, these metrics are paramount in gauging the model's overall performance.

4. Experiment

This section presents and analyzes the results obtained during the training process of the DistilBERT model on a specific database. The results obtained from the distilbase-v1 dataset were deemed an experimental run, the objective of which was twofold: firstly, to ensure that the code configuration was appropriate, and secondly, to assess the computational capacity of the training machine. Additionally, the experiment facilitated the expansion of the database from two to three projects, thereby enhancing the dataset. Although the augmented data volume likely resulted in prolonged training periods, this was imperative to minimize the requisite number of epochs, as illustrated in Table 3. A principal factor for comparison is the duration of the training process, which exhibited a discrepancy between the two experiments. This discrepancy can be attributed to the incorporation of the evaluation dataset into Experiment v2. Despite the augmented complexity, the training times exhibited a proportionality, reflecting the augmented data volume.

Experiment distil-pid-v2 was particularly notable for the high precision values and promising model behavior observed under the established parameters. The F1 score, as illustrated in Fig. 6, showed a rapid increase during the initial training steps before stabilizing, indicating that the model quickly reached and maintained a high level of performance. This suggests that the model effectively captures the essential features of the data early in the training process. The precision metric demonstrated a consistent and linear increase, reaching a value of 0.9181 by the end of the process, indicating continuous improvement in the model. Significant peaks in the gradient norms shown in Fig. 6 were observed throughout the training, particularly around steps 1k, 2k, and 5k, reflecting moments of considerable parameter adjustment.

While these results are promising, it is crucial to interpret their broader implications. The high precision and low loss metrics suggest that the DistilBERT model is well-suited for the specific task of database consistency detection in intelligent P&ID design. However, the scalability of these results to other domains remains to be tested. Compared to previous work, the model's performance exceeds initial expectations, particularly in its ability to generalize across more complex datasets. This suggests a significant advancement in the field, especially considering the model's adaptability to increased data volumes. Future research should further validate these results across different datasets and explore the model's applicability to other industrial contexts (see Fig. 7).

The loss metric as shown in Fig. 8 exhibited a sharp decline within the first few thousand steps, after which it stabilized at a very low value. This behavior reflects the model's efficiency in minimizing errors and suggests a high level of convergence, implying that further training is unlikely to yield significant improvements. However, while the low loss is indicative of a strong model, it is essential to compare these results with industry benchmarks to assess their real-world applicability. The results obtained indicate an effective training process for the DistilBERT model, but the broader implications of these findings warrant further discussion. The following sections will delve into these implications, comparing them to previous work and identifying areas for future research based on the observed outcomes.

5. Conclusion

The training of the DistilBERT model on an expanded dataset produced promising outcomes, highlighting its efficiency and adaptability in processing larger-scale data within the context of intelligent P&ID design. Experiment 2 demonstrated the model' s capacity to rapidly attain a high level of performance, as evidenced by the F1 score' s swift stabilization and the precision metric' s unwavering enhancement, reaching 0.9181. These outcomes reflect a significant advancement in the field of database consistency detection. The model effectively captured essential features at an early stage of training, thereby minimizing error rates as indicated by the low and stable loss metrics.

This study makes a significant contribution to the field by demonstrating the robustness and scalability of DistilBERT within the specialized context of database consistency detection. The model's capacity to generalize across more complex and larger datasets represents a significant advancement over previous approaches, particularly in terms of achieving high precision and rapid convergence. Furthermore, the incorporation of data augmentation techniques, when feasible, could serve to augment the model's generalization capabilities by diversifying the training dataset. The application of cross-validation and continuous monitoring of learning curves ensured that the model's performance remained robust and free from overfitting, which is essential for the rigorous development and tuning of AI models.

Notwithstanding these achievements, the study is constrained by its focus on a specific industrial application and dataset, which raises questions about the generalizability of these findings to other domains or data types. The constraints imposed by limited computational resources and project execution times, particularly within the three-month timeframe stipulated by the innovation strengthening call in Colombia, precluded the comprehensive implementation of certain hyperparameter optimization strategies. Furthermore, the augmented training times resulting from dataset expansion underscore the trade-offs between data volume and computational efficiency, which may restrict the practical applicability of this approach in resource-constrained environments.

6. Future directions

Extending this research to diverse datasets and industrial contexts is crucial to validate the model's scalability and adaptability. Future work should also investigate the integration of DistilBERT with other machine learning techniques or models, such as those in computer vision, to develop a more comprehensive solution for P&ID design validation. Benchmarking these results against industry standards will be essential for evaluating the model's practical applicability and identifying potential avenues for further optimization. Moreover, addressing the computational constraints that constrained this study will be vital for achieving more extensive and refined hyperparameter tuning in future iterations.

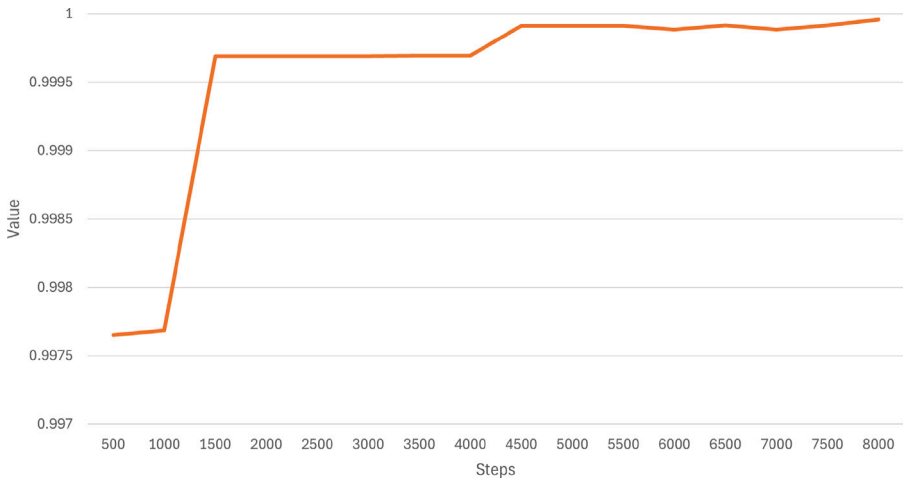

Fig. 6. Evolution of F1 score in DistilBERT training.

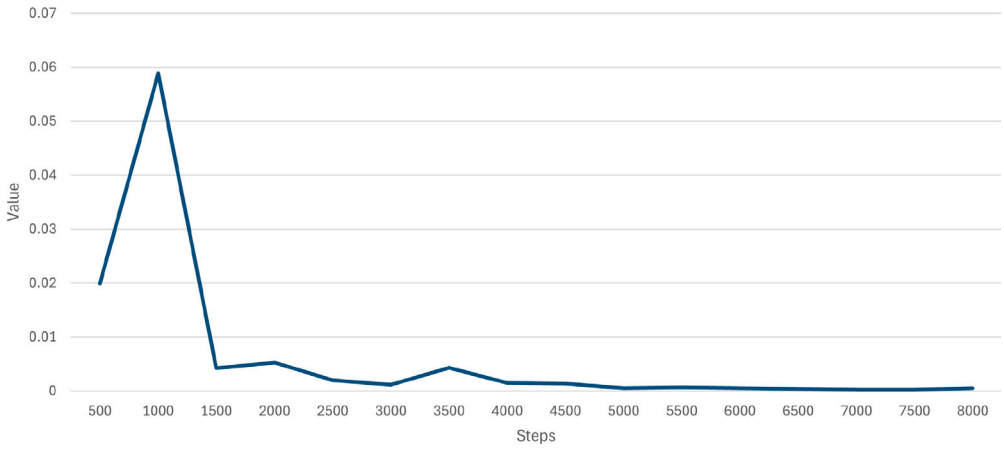

Fig. 7. Gradient Norms during DistilBert training.

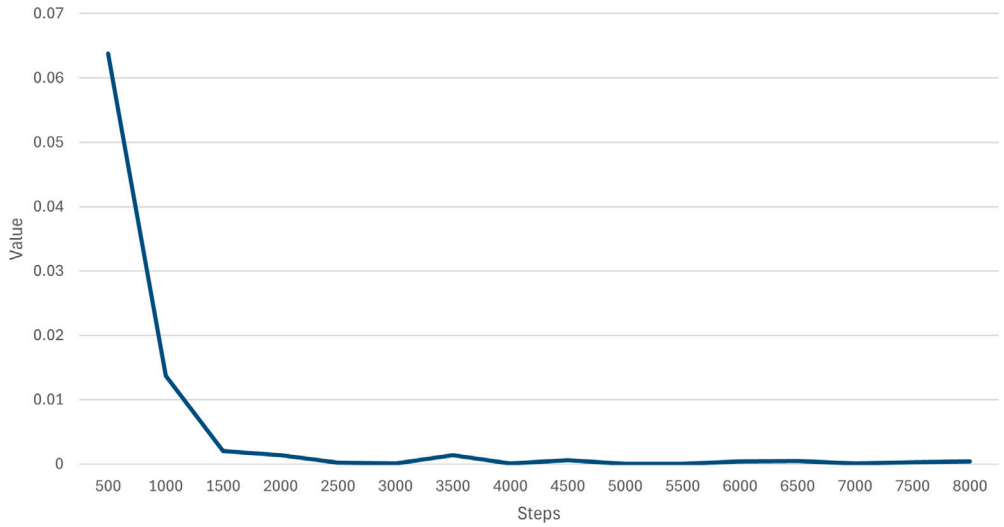

Fig. 8. Training Loss for distilbert-pid-v2.

CRediT authorship contribution statement

F.S. Gómez-Vega: Writing – original draft, Visualization, Validation, Investigation, Data curation. **O. Acuña:** Supervision, Investigation, Funding acquisition, Formal analysis. **Andrea C. Camargo:** Visualization, Validation, Resources, Data curation, Conceptualization. **Jeison D. Jimenez:** Writing – original draft, Visualization, Validation, Supervision, Resources, Methodology, Data curation, Conceptualization. **Sara M. Galeano:** Visualization, Validation, Resources, Formal analysis, Data curation. **Isabella E. Franco:** Validation, Supervision, Resources, Investigation, Formal analysis, Data curation. **Laura L. Lozano:** Visualization, Validation, Resources, Investigation, Formal analysis, Data curation. **Jenifer Vásquez:** Writing – review & editing, Supervision, Resources, Methodology, Formal analysis, Conceptualization. **Edwin Puertas:** Writing – review & editing, Writing – original draft, Supervision, Resources, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Edwin Puertas reports article publishing charges was provided by Tecnologica University of Bolivar. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

The authors would like to acknowledge the support provided by the master's degree scholarship program in engineering at the Universidad Tecnologica de Bolivar (UTB) in Cartagena, Colombia.

Data availability

The authors do not have permission to share data.

References

[1] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, 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, Association for Computational Linguistics, 2019, pp. 4171–4186, http://dx.doi.org/10.18653/v1/N19-1423.

[2] V. Sanh, L. Debut, J. Chaumond, T. Wolf, DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, in: NeurIPS Workshop on Energy Efficient Machine Learning, 2019, URL https://arxiv.org/abs/1910.01108.

[3] E.L. Vieira, S.E.G. da Costa, E.P. de Lima, C.C. Ferreira, Application of the Proknow-C methodology in the search of literature on performance indicators for energy management in manufacturing and industry 4.0, Procedia Manuf. 39 (2019) 1259–1269, http://dx.doi.org/10.1016/j.promfg.2020.01.343.

[4] S. Krüger, M. Borsato, Developing knowledge on digital manufacturing to digital twin: a bibliometric and systemic analysis, Procedia Manuf. 38 (2019) 1174–1180, http://dx.doi.org/10.1016/j.promfg.2020.01.207.

[5] S.M. Gajbhiye, S.R. Bhamre, L.N.T. Tadepalli, M.R. Pillai, D. Uplaonkar, Advancing P&ID digitization with YOLOv5, in: International Conference on Integrated Intelligence and Communication Systems, ICIICS, 2023, http://dx.doi.org/10.1109/ICIICS59993.2023.10421368.

[6] R. Dzhusupova, R. Banotra, J. Bosch, H.H. Olsson, Using artificial intelligence to find design errors in engineering drawings, J. Softw.: Evol. Process. 35 (12) (2023) http://dx.doi.org/10.1002/smr.2543.

[7] G. Su, S. Zhao, T. Li, S. Liu, Y. Li, G. Zhao, Z. Li, Image format pipeline and instrument diagram recognition method based on deep learning, Biomim. Intell. Robot. 4 (1) (2024) 100142, http://dx.doi.org/10.1016/j.birob.2023.100142.

[8] Y.M. Ibrahim, T.C. Lukins, X. Zhang, E. Trucco, A.P. Kaka, Towards automated progress assessment of workpackage components in construction projects using computer vision, Adv. Eng. Informat. 23 (1) (2009) 93–103, http://dx.doi.org/10.1016/j.aei.2008.07.002.

[9] E.S. Yu, J.M. Cha, T. Lee, J. Kim, D. Mun, Features recognition from piping and instrumentation diagrams in image format using a deep learning network, Energies 12 (23) (2019) 4425, http://dx.doi.org/10.3390/en12234425.

[10] M.K. Gellaboina, V.G. Venkoparao, Graphic symbol recognition using auto associative neural network model, in: Proceedings of the 7th International Conference on Advances in Pattern Recognition, ICAPR, 2009, pp. 297–301, http://dx.doi.org/10.1109/ICAPR.2009.45.

[11] R. Rahul, S. Paliwal, M. Sharma, L. Vig, Automatic information extraction from piping and instrumentation diagrams, in: International Conference on Pattern Recognition Applications and Methods, 2019, pp. 163–172, http://dx.doi.org/10.5220/0007376401630172.

[12] J. Oeing, W. Welscher, N. Krink, L. Jansen, F. Henke, N. Kockmann, Using artificial intelligence to support the drawing of piping and instrumentation diagrams using DEXPI standard, Digit. Chem. Eng. 4 (2022) 100038, http://dx.doi.org/10.1016/j.dche.2022.100038.

[13] H. Kim, W. Lee, M. Kim, Y. Moon, T. Lee, M. Cho, D. Mun, Deep-learning-based recognition of symbols and texts at an industrially applicable level from images of high-density piping and instrumentation diagrams, Expert Syst. Appl. 183 (2021) 115337, http://dx.doi.org/10.1016/j.eswa.2021.115337.

[14] V.K. Bhanuse, J. Kulkarni, S. Patankar, A.R. Bauskar, A.M. Bhilare, S.N. Bhosale, Enhancing piping and instrumentation diagram recognition: A machine learning approach, in: IEEE International Conference for Convergence in Technology, I2CT, 2024, http://dx.doi.org/10.1109/I2CT61223.2024.10543300.

[15] Y. Guermazi, S. Sellami, O. Boucelma, A RoBERTa based approach for address validation, in: Communications in Computer and Information Science, vol. 1652, 2022, pp. 157–166, http://dx.doi.org/10.1007/978-3-031-15743-1_15.

[16] P.C. Verpoort, P. MacDonald, G.J. Conduit, Materials data validation and imputation with an artificial neural network, Comput. Mater. Sci. 147 (2018) 176–185, http://dx.doi.org/10.1016/j.commatsci.2018.02.002.

[17] S. Ghosh, A. Zaboli, J. Hong, J. Kwon, Object-focused risk evaluation of AI-driven perception systems in autonomous vehicles, in: IEEE Transportation Electrification Conference and Expo, ITEC, 2024, http://dx.doi.org/10.1109/ITEC60657.2024.10599086.

[18] K. Kubola, B. Kongon, P. Boonmee, P. Jitngernmadan, Optimal AI model for industrial AR application based on object detection, in: International Conference on Information Technology (InCIT), 2023, pp. 38–42, http://dx.doi.org/10.1109/INCIT60207.2023.10413004.

[19] A.L. Wozniak, N.Q. Duong, I. Benderitter, S. Leroy, S. Segura, R. Mazo, Robustness testing of an industrial road object detection system, in: IEEE International Conference on Artificial Intelligence Testing (AITest), 2023, pp. 82–89, http://dx.doi.org/10.1109/AITEST58265.2023.00022.

[20] R.N.S. Singla, Comparative analysis of transformer based pre-trained NLP models, Int. J. Comput. Appl. 8 (11) (2020) 40–44.

[21] R. Qasim, W.H. Bangyal, M.A. Alqarni, A.A. Almazroi, A fine-tuned BERT-based transfer learning approach for text classification, J. Heal. Eng. 2022 (2022) http://dx.doi.org/10.1155/2022/3498123.

[22] A. Nagarajan, Optimizing Transformers with Approximate Computing for Faster, Smaller and more Accurate NLP Models, Tech. rep, 2020, arXiv, URL https://arxiv.org/abs/2010.03688.

[23] W. Zhang, J. Joseph, Y. Yin, L. Xie, T. Furuhata, S. Yamakawa, K. Shimada, L.B. Kara, Component segmentation of engineering drawings using graph convolutional networks, Comput. Ind. 147 (2023) 103885, http://dx.doi.org/10.1016/j.compind.2023.103885.

[24] S. Paliwal, M. Sharma, L. Vig, OSSR-PID: One-shot symbol recognition in P&ID sheets using path sampling and GCN, in: International Joint Conference on Neural Networks, IJCNN, 2021, http://dx.doi.org/10.1109/IJCNN52387.2021.9534122, URL https://arxiv.org/abs/2109.03849.

[25] S. Sarkar, P. Pandey, S. Kar, Automatic Detection and Classification of Symbols in Engineering Drawings, Tech. rep, 2022, http://dx.doi.org/10.48550/arXiv.2204.13277, arXiv, URL https://arxiv.org/abs/2204.13277.

[26] S. Paliwal, A. Jain, M. Sharma, L. Vig, Digitize-PID: Automatic digitization of piping and instrumentation diagrams, in: Lecture Notes in Computer Science, vol. 12705, 2021, pp. 168–180, http://dx.doi.org/10.1007/978-3-030-75015-2_17.

[27] S. Wu, Y. Wang, H. Yang, P. Wang, Improved faster R-CNN for the detection method of industrial control logic graph recognition, Front. Bioeng. Biotechnol. 10 (2022) http://dx.doi.org/10.3389/fbioe.2022.944944.

[28] A. Aspin, Change tracking and change data capture, in: SQL Server 2012 Data Integration Recipes, 2012, pp. 681–729, http://dx.doi.org/10.1007/978-1-4302-4792-0_12.

[29] L. Hao, T. Jiang, Y. Lin, Y. Lu, Methods for solving the change data capture problem, in: Lecture Notes on Data Engineering and Communications Technologies, vol. 153, 2023, pp. 781–788, http://dx.doi.org/10.1007/978-3-031-20738-9_87.

[30] A.O. Salau, M.M. Beyene, Software defined networking based network traffic classification using machine learning techniques, Sci. Rep. 14 (1) (2024) http://dx.doi.org/10.1038/s41598-024-70983-6.

[31] P. Singh, Systematic review of data-centric approaches in artificial intelligence and machine learning, Data Sci. Manag. 6 (3) (2023) 144–157, http://dx.doi.org/10.1016/j.dsm.2023.06.001.

[32] P.L. Foalem, F. Khomh, H. Li, Studying logging practice in machine learning-based applications, Inf. Softw. Technol. 170 (2024) 107450, http://dx.doi.org/10.1016/j.infsof.2024.107450.

[33] A. Chen, A. Chow, A. Davidson, A. Dcunha, A. Ghodsi, S.A. Hong, A. Konwinski, C. Mewald, S. Murching, T. Nykodym, P. Ogilvie, M. Parkhe, A. Singh, F. Xie, M. Zaharia, R. Zang, J. Zheng, C. Zumar, Developments in MLflow: A system to accelerate the machine learning lifecycle, in: Workshop on Data Management for End-To-End Machine Learning, 2020, http://dx.doi.org/10.1145/3399579.3399867.