Detecting Semantic Data Smells with BERT: A Transformer-Based Approach to Data Quality

Gilberto Recupito[1], Giammaria Giordano[1], Dario Di Nucci[1] and Fabio Palomba[1]

[1]SeSa Lab - University of Salerno, Italy

Abstract

In recent years, the integrity of data used in machine learning pipelines has become increasingly critical, as even state-of-the-art models are constrained by the quality of their input. Among the various threats to data reliability, data smells—subtle and often semantic anomalies—pose unique challenges due to their nuanced and context-sensitive nature. This paper presents an innovative approach to detect semantic data smells using BERT, a transformer-based language model originally designed to understand natural language. We focus specifically on two underexplored categories: *Split Value Smells* and *Multiple Value Smells*, which respectively reflect improper data fragmentation and conflation. To facilitate model training, we constructed a large and heterogeneous corpus comprising synthetic and real datasets, utilizing data augmentation techniques via Faker and SDV. We trained two binary classifiers using fine-tuned BERT models, achieving high performance (F1 scores of 0.88 and 0.98) in detecting these smells. Our findings demonstrate that transformer-based models are not only effective in capturing structural patterns in tabular data but also capable of generalizing across diverse semantic anomalies. This work establishes a foundation for the broader application of language models in data quality assurance, opening new avenues for semantic-level data cleaning automation.

Keywords

Data Smells, Data Quality, Software Engineering for AI, MLOps, Empirical Software Engineering

1. Introduction

As machine learning (ML)-enabled systems evolve from experimental prototypes to deployed applications, ensuring their reliability requires attention to all stages of the ML lifecycle, from data preparation to model deployment and monitoring [1]. To support this shift, MLOps has emerged as a collection of engineering practices that extend the DevOps principles to ML systems [2]. A key aspect of MLOps is continuous quality assurance, which involves maintaining the performance, integrity, and stability of ML systems throughout their operation. Although current MLOps efforts often emphasize monitoring model performance and retraining strategies, data quality remains a critical and challenging factor [3]. Low-quality data can silently introduce errors, degrade model performance, and increase the risk of biased or misleading predictions. Among the most subtle and impactful threats to data quality are data smells, data value-based indications of latent data quality issues caused by poor practices that may lead to future failures [4]. Although certain types of data smell—particularly those related to syntax or formatting—can be addressed using existing validation tools, current approaches fall short when it comes to semantic data smells. These involve deeper contextual and structural inconsistencies that are difficult to detect with rule-based techniques [4, 5].

In this preliminary work, our goal is to investigate initial opportunities for modeling and identifying subtle semantic anomalies in tabular data. To start exploring these capabilities, we focused on two simple data quality issues related to the distribution of information among the columns of a dataset: *Split Value Smells* and *Multiple Value Smells*. By transforming tabular data into textual representations,

MLOps25: Workshop on Machine Learning Operations. October 25, 2025, Bologna, Italy

*Corresponding author.

†These authors contributed equally.

grecupito@unisa.it (G. Recupito); giagiordano@unisa.it (G. Giordano); ddinucci@unisa.it (D. D. Nucci); fpalomba@unisa.it (F. Palomba)

0000-0001-8088-1001 (G. Recupito); 0000-0003-2567-440X (G. Giordano); 0000-0002-3861-1902 (D. D. Nucci); 0000-0001-9337-5116 (F. Palomba)

we leveraged the ability of BERT to capture semantic patterns and irregularities. Our approach supports automated semantic data quality checks that can be integrated into MLOps pipelines, reducing manual overhead, and improving the robustness of ML-enabled systems over time.

To train and evaluate our models, we created a synthetic dataset using a combination of the `Faker` library [6] and the `Synthetic Data Vault (SDV)` [7]. We fine-tuned two separate BERT-based classifiers and evaluated their performance using standard classification metrics. The model trained for *Split Value Smells* achieved an accuracy of 88.99% with perfect precision and an F1 score of 0.8763. The model trained for *Multiple Value Smells* achieved an even higher accuracy of 98% and an F1 score of 0.9795. We provide the following contributions:

- We introduce a novel BERT-based approach to detect semantic data smells in tabular datasets.
- We define and formalize two complex smell types—*Split Value Smells* and *Multiple Value Smells*—and propose an automated way to inject them into synthetic data.
- We build two meta-datasets to explore further research on semantic data smells.

2. Background and Related Work

In this section, we outline key data quality issues relevant to ML-enabled systems, establishing the broader context in which these issues arise. We then review related work on the detection of such issues, with a particular focus on recent advances in data smell identification.

2.1. Data Quality Assurance in MLOps Environments

In MLOps pipelines, data quality plays a central role in ensuring long-term performance, reliability, and maintainability of ML-enabled systems. Unlike traditional software systems, where source code is the primary artifact, ML-enabled systems are highly dependent on data, an inherently dynamic and evolving asset. Renggli et al. [8] introduced a data quality-driven perspective of MLOps, arguing that many MLOps challenges are fundamentally rooted in data management. They emphasized how key data quality dimensions are distributed across different stages of the ML pipeline. Degraded data quality directly impacts model performance. Mohammed et al. [9] investigated this relationship, demonstrating that data completeness and consistency issues have a significant impact on both classification and regression tasks. This finding supports earlier work by the data management community, which focused on automating data quality assurance practices and tools [10]. Therefore, current data validation techniques are typically integrated into CI/CD pipelines and focus on syntactic validations, such as type checking, missing values, or schema conformity.

In addition to classical data quality issues, researchers have introduced the concept of data smells, recurring patterns in data that suggest latent quality problems. Foidl et al. [4] proposed a taxonomy of these smells, distinguishing between syntactic and semantic issues. Building on this work, Recupito et al. [5] expanded the taxonomy by introducing two new types of semantic data smells relevant to tabular datasets: *Split Value Smells* and *Multiple Value Smells*.

The former refers to cases where a single piece of information is unnecessarily distributed across multiple columns, for example, splitting a date of birth into separate month, day and year components, as illustrated in Table 1. Such fragmentation can hinder downstream processing and reduce data usability. The latter involves instances where multiple distinct pieces of information are inappropriately merged into a single column, for example, by storing both an email address and a phone number in the same field, as shown in Table 2. This practice complicates data interpretation and parsing. Both smells represent semantic anomalies that go beyond syntactic validation and require contextual understanding to detect and resolve effectively. These semantic issues are particularly problematic in production-grade ML systems, where they can silently propagate through automated workflows, resulting in weak models and unreliable predictions. However, such techniques often fail to capture more complex violations of the data's semantic intent, such as improperly merged or split values. Overcoming these limitations requires

more adaptive and context-aware strategies that go beyond rule-based checks and can understand the structure and meaning of the data within its application context.

Table 1
Example of Split Value Smell.

Date of Birth	→	Month	Day	Year
12-08-1995	→	12	08	1995
01-12-1993	→	01	12	1993
05-12-2000	→	05	12	2000
12-01-2000	→	12	01	2000
...	→	...	...	...

Table 2
Example of Multiple Value Smell.

Email	Phone Number	→	Contact
alice@example.com	+391234567890	→	alice@example.com, +391234567890
bob@example.com	+391234567891	→	bob@example.com, +391234567891
john@example.com	+391234567892	→	john@example.com, +391234567892
mark@example.com	+391234567893	→	mark@example.com, +391234567893
...	...	→	...

2.2. Related Work

Recent studies have increasingly focused on formalizing and detecting data smells, recurring patterns in data that may indicate latent quality issues. Foidl et al. [4] introduced a taxonomy of data smells and proposed DSD, a tool to detect a subset of syntactic data smells. They also developed a machine learning–based approach for identifying format inconsistency smells. However, most existing detection methods remain limited to easily measurable features, such as null values or outliers, and overlook semantic-level anomalies that span multiple fields or require contextual interpretation. Li et al. [11] presented CleanML, a tool that automatically detects and cleans data quality issues, including missing, extreme, and duplicate values. Their empirical study demonstrated that applying data cleaning techniques can significantly improve the performance of downstream ML models. However, these approaches largely focus on surface-level issues and fall short of addressing complex data semantics, particularly in structured data used in ML pipelines.

In parallel, recent advances in transformer-based models—particularly BERT and its derivatives—have shown strong potential for processing tabular and hybrid (text-tabular) data. These models have been successfully adapted to capture the structure and semantics of tables in diverse tasks. For example, TaBERT [12] and TAPAS [13] jointly model natural language and tabular data for Table Question Answering, incorporating table-specific attention mechanisms and positional encodings. For semantic table understanding, models such as TURL [14] and TUTA [15] leverage structure-aware transformers to perform entity linking, relation extraction, and hierarchical parsing. In self-supervised settings, TABBIE [16] and TABNER [17] pretrain BERT-based architectures for tasks like corrupted cell detection and named entity recognition in spreadsheets, demonstrating the models' ability to learn meaningful structural patterns. These works illustrate BERT's strong contextual embedding capabilities, making it a compelling foundation for detecting semantic anomalies in structured data. Building on this foundation, our work bridges the gap between data quality and research communities by applying BERT to semantic data validation. Specifically, we reinterpret rows in tabular datasets as natural language sequences and fine-tune BERT to detect data smells that manifest semantic and contextual irregularities. Our approach

contributes to data quality assurance by offering a flexible, learning-based solution to identify complex data quality issues.

Contribution to the State of the Art.

While existing work on data quality predominantly addresses structural and syntactic issues, this work explores the capabilities of transformer-based models of detecting semantic data smells that require semantic understanding.

3. Research Method

The *goal* of this study is to explore whether BERT can support automated detection of semantic smells in tabular data sets. We focus our investigation on two key categories: *Split Value Smells*, which occur when a single logical entity (e.g., a person's full name or an address) is improperly fragmented across multiple columns; and *Multiple Value Smells*, which result from the incorrect merging of distinct values, such as a phone number and an email, into a single cell.

This study is carried out from a dual *perspective*. On the one hand, our aim is to contribute to the body of knowledge by understanding the applicability of language models in semantic data quality tasks, thereby addressing the interests of researchers in empirical data-centric ML. On the other hand, we respond to the practical needs of developers and data engineers, who routinely face the challenge of detecting and correcting these types of anomaly at scale. In terms of reporting, we follow the guidelines of Wohlin et al. [18]. We formulated two research questions to guide our investigation:

RQ_1. On the Use of Bert in Detecting Split Value Smells.

Can a BERT-based model detect Split Value Smells in tabular datasets?

This question aims to assess the extent to which a BERT-based model can recognize the presence of *Split Value Smells* in tabular data.

To answer this question, we trained a fine-tuned BERT model using pairs or groups of columns as inputs, treating them as serialized sequences with embedded positional markers. We then evaluated the model's predictions using standard classification metrics, including precision, recall, and F1 score, to measure its sensitivity to split patterns across different data layouts.

RQ_2. On the Use of Bert in Detecting Multiple Value Smells.

Can a BERT-based model detect Multiple Value Smells in tabular datasets?

With this question, we explore ability of BERT to identify cases in which multiple logically independent values are improperly merged into a single cell (i.e., Multiple Value Smells). For this task, we fine-tuned BERT to process individual cells as input, leveraging its token-level attention to learn whether a given entry contains signs of conflation—e.g., delimiters, unexpected semantic shifts, or irregular formatting. As with RQ1, we used a combination of accuracy, precision, recall, and F1 score to assess performance and complemented this with confusion matrix analysis to uncover boundary cases and failure modes.

To conduct our study, we constructed a synthetic dataset of labeled tabular entries. These were designed to represent realistic examples of both clean and anomalous data. Where possible, we introduced smells based on known real-world issues observed in open datasets, supplemented by manually engineered cases to ensure balanced coverage across scenarios.

All records were serialized in a column-based format suitable for the input structure of BERT, and each instance was annotated with a binary label corresponding to the presence of a smell, allowing us to train and evaluate the models under consistent conditions.

Figure 1 overviews the main steps we performed to conduct our experiments, which will be explained in more detail in the following sections.

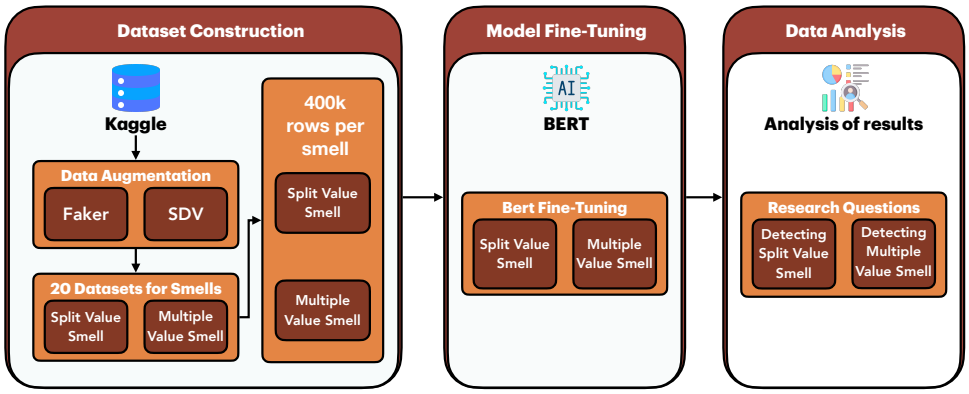

Figure 1: Overview of the Research Method.

3.1. Dataset Construction

Due to the limited availability of well-defined semantic data smells datasets, we constructed two comprehensive meta-datasets[1] for detecting *Split Value Smells* and *Multiple Value Smells*, respectively [4].

Source and Synthetic Data Generation. We collected real-world datasets from public repositories such as Kaggle.[2] However, since real datasets rarely include explicit annotations of data smells, we employed synthetic data generation to create controlled and diverse examples. Specifically, we used the Python Faker library [6] to generate human-readable tabular data that can be affected by these data smells, including names, addresses, and dates of birth, enabling us to simulate plausible examples of well-formed and smelly data structures.

To enhance diversity and quantity, we expanded the dataset using Synthetic Data Vault (SDV) [7], which enables the generation of synthetic tabular data while preserving the statistical properties of the original data. We trained SDV synthesizers in our initial datasets to generate additional records, thus balancing and scaling the data. The validity of the generated data was confirmed using the built-in diagnostic tools of SDV[3]. All data sets achieved 100% synthetic data validation scores, indicating high fidelity in capturing the original data distributions and relational structures.

Data Smell Injection. We injected two data smells: *Split Value Smells* and *Multiple Value Smells*. On the one hand, we programmatically fragmented single logical values (e.g., full addresses or dates) into multiple columns. On the other hand, we concatenated logically distinct values into single fields, such as merging "email" and "phone number" into a single string-separated column.

Tables 1 and 2 show examples of these transformations. In total, we generated 40 base datasets (20 per smell type), each consisting of 5,000 rows. These were then aggregated into two meta-datasets with approximately 400,000 labeled records. Due to the injection strategy, the class distribution in each raw meta-dataset was naturally imbalanced. For example, in the case of *Split Value Smells*, the smelly examples were overrepresented because each split introduced multiple rows. Conversely, *Multiple Value Smells* tended to be underrepresented due to the merging of values into fewer fields. In addition, row deduplication is performed across all data sets to ensure diversity. Finally, an undersampling strategy was applied to mitigate learning bias. The final version used for training and evaluation contained a balanced 50/50 distribution of positive and negative labels. Specifically, we applied stratified sampling to preserve variation between different base datasets and structural patterns. After this step, two meta-datasets, each containing 5,000 rows, are created.

[1] A meta-dataset refers to a collection of datasets, where each entry captures general information about an individual dataset.
[2] https://www.kaggle.com/
[3] SDV diagnostic tool: https://docs.sdv.dev/sdv/single-table-data/evaluation/diagnostic

Preprocessing and Labeling. Each row in the meta-dataset was labeled as "smelly" or "not smelly" based on whether it contained a known transformation. Additionally, to prepare the data for model training with BERT, we transformed each row into a structured text format (e.g., `Feature: Value`), allowing the model to treat tabular records as natural language sequences, as recommended in the recent literature [12]. After preprocessing, we serialized each data row into a text sequence and analyzed the sequence length of each row to conform to BERT input requirements. The 94% of the sequences were less than 50 tokens, with a maximum sequence length of 110 tokens.

3.2. Model Fine-Tuning

To effectively detect semantic data smells within tabular data, we adopted a fine-tuning approach on a pretrained BERT model. This decision leverages the strong contextualization capabilities of BERT, which, although initially designed for natural language processing tasks, has recently demonstrated high effectiveness in structured data domains [12].

Preprocessing and Input Representation. The fine-tuning process began transforming tabular records into text sequences. Each row was serialized into a single textual string by concatenating feature names with their corresponding values (e.g., `Feature: Value`). This format ensures that the BERT tokenizer can effectively segment and encode the data while preserving the structure of the original features. All inputs were tokenized using the BERT base uncased tokenizer and padded or truncated to a maximum sequence length of 128 tokens. Given the binary nature of our task (smelly vs. not smelly), each data sample was associated with a binary label. The input sequences were then divided into training (80%) and validation (20%) sets using stratified sampling to maintain label balance.

Model Architecture and Training Setup. We used the `bert-base-uncased` model from the Hugging Face Transformers library.[4] A classification head—consisting of a dropout layer followed by a dense layer—was added on top of the final [CLS] token representation. This setup enables binary classification by mapping the contextualized embedding to a single probability score. The training was carried out using AdamW optimizer [19], with a learning rate of 2×10^{-5}, the Binary Cross-Entropy loss function, and a batch size of eight, three epochs, and a maximum sequence length of 128 tokens. Each model was trained independently to avoid confusion caused by overlapping structural patterns between smell types. This phase was performed on a NVIDIA Tesla T4 through Google Colab.

3.3. Public Data Availability

To support replicability and facilitate future research, we have made all materials, including scripts and datasets, publicly available in a permanent online appendix [20].

4. Results of our Work

This section reports on the findings of the study. To improve clarity, we discuss each **RQ** separately.

4.1. On the Use of BERT in Detecting Split Value Smell

As shown in Table 3a, the model achieved an accuracy of 88.99%, with perfect precision (1.000) and an F1 score of 0.8763. Although these results confirm the ability of the model to detect *Split Value Smells* with high reliability, the recall was 0.7799, indicating that some true positives were missed.

Table 3b shows the confusion matrix for the *Split Value Smells* model. In particular, the model did not produce false positives, resulting in perfect precision. However, it did not detect 11% of smelly instances, which were mistakenly classified as clean (false negatives), highlighting a common trade-off in binary classification: the model favors precision over recall. Such behavior is particularly desirable

[4]Available at: https://huggingface.co/google-bert/bert-base-uncased

Metric	Value
Accuracy	0.8899
Precision	1.0000
Recall	0.7799
F1 Score	0.8763

(a) Performance Metrics

	Predicted	
Actual	Negative	Positive
Negative	50%	0%
Positive	11%	39%

(b) Confusion Matrix

Table 3
BERT's performance and confusion matrix for Split Value Smells.

in contexts where overflagging clean data is more costly than missing subtle anomalies, but it can be detrimental if undetected anomalies significantly impact downstream tasks.

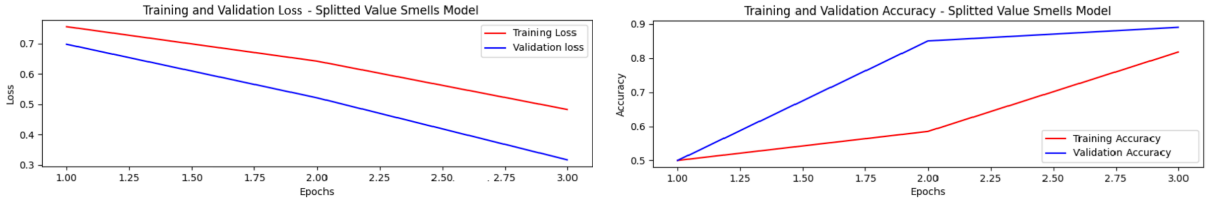

Figure 2: Training and validation curves for split Value Smell.

Figure 2 illustrates the evolution of loss and accuracy throughout training and validation. The convergence of the validation and training loss curves, along with the parallel increase in accuracy, suggests that the model generalizes well and does not suffer from overfitting. This stability reinforces the idea that BERT is effective in capturing semantic patterns associated with *Split Value Smells*, even when trained on relatively limited data.

One notable challenge lies in the *semantic variability* of *Split Value Smells*. Unlike syntactic anomalies, these smells are often context-dependent. For example, splitting an address into "Street", "City", and "Zip Code" might be acceptable in some domains but smelly in others if those components are only meaningful when considered as a whole. This ambiguity limits the model's ability to generalize across domains and may explain the relatively lower recall. Moreover, different instances of splitting (e.g., full name into first, middle, and last names) lack a consistent pattern, which increases intra-class variability and training noise. The zero false-positive rate of the model demonstrates that it serves as a *reliable filter* for high-confidence cases. In practice, this makes it well suited for human-in-the-loop data quality workflows, where automatic detection aids but does not replace manual review. A flagged instance can be trusted to be problematic, while borderline or context-sensitive cases may still require user judgment.

In summary, *BERT can effectively detect Split Value Smell, offering high precision and generalizability despite the challenges posed by semantic ambiguity.* Future improvements could involve integrating auxiliary metadata to capture implicit semantic groupings better.

RQ_1 — Summary of the Results.

BERT effectively detects *Split Value Smells* with high precision and generalizability. Although the recall is modest due to the semantic variability of smells, the model serves as a strong, high-confidence filter in human-in-the-loop data cleaning workflows.

4.2. On the Use of BERT in Detecting Multiple Value Smells

As shown in Table 4a, the BERT-based model for detecting *Multiple Value Smells* achieved exceptional results. The model reported an overall accuracy of 98.00%, with perfect precision (1.000) and a remarkably high recall of 0.9599, which led to an F1 score of 0.9795. These metrics indicate that the model not only correctly classifies true positives but also maintains a near-zero false negative rate, substantially outperforming the counterpart model developed for *Split Value Smells*.

Metric	Value
Accuracy	0.9800
Precision	1.000
Recall	0.9599
F1 Score	0.9795

(a) Performance Metrics for BERT.

	Predicted	
Actual	Negative	Positive
Negative	50%	0%
Positive	2%	48%

(b) Confusion Matrix for BERT.

Table 4
BERT's performance and confusion matrix for detecting Multiple Value Smells.

The confusion matrix in Table 4b confirms this performance: the model correctly identifies nearly all instances with a smell, with only a minimal number of false negatives. As in the previous model, no false positives were recorded, preserving perfect precision. This combination of high recall and perfect precision makes the model both sensitive and specific, a highly desirable trait in anomaly detection.

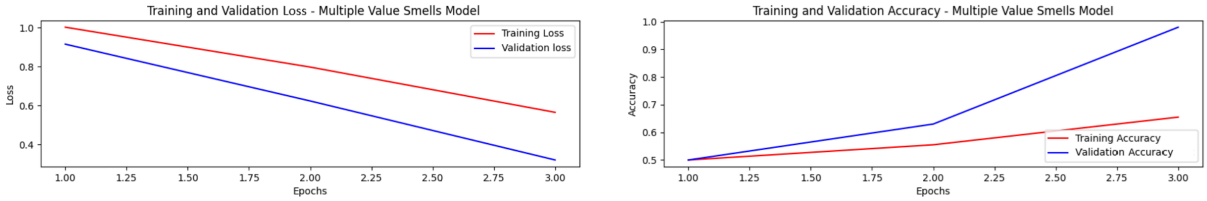

Figure 3: Training and validation curves for Multiple Value Smells.

Figure 3 shows the training and validation curves for loss and accuracy. As observed with the *Split Value* model, both loss functions decreased steadily, while accuracies converged to a high value, suggesting good generalization. The final validation accuracy plateaued at 0.98, with minimal divergence from the training accuracy, implying that the model avoids overfitting and maintains robustness.

The superior performance of the *Multiple Value* model can be attributed to the *more uniform representation* of the smells. Unlike *Split Value Smells*, which vary significantly in their structural form, *Multiple Value Smells* often exhibit consistent patterns that are more easily captured by the token-based architecture of BERT. These regularities provide stronger learning signals, enabling the model to distinguish between smelly instances and clean data with high reliability. From a practical perspective, the high recall achieved by the model ensures that very few smells are overlooked, making it suitable for autonomous data cleaning tools or alert systems in large-scale data pipelines. Furthermore, perfect precision allows its output to be used confidently without requiring downstream verification, minimizing the risk of incorrect flagging.

In conclusion, *BERT is not only capable of detecting Multiple Value Smells but does so with near-optimal performance.* The consistent structure of these smells aligns well with BERT's contextual encoding capabilities, enhancing the utility of the model in detecting semantic level anomalies in tabular data.

RQ_2 — Summary of the Results.

BERT achieves near-perfect detection of *Multiple Value Smells*, with 98% accuracy, perfect precision, and very high recall. The model is effectively generalized and benefits from the uniformity of the smell structure, making it suitable for reliable automated anomaly detection in tabular data.

5. Discussion and Implications

The results obtained in this study have important implications for both the development of data quality systems in practice and future research on semantic anomaly detection. By demonstrating that BERT can be fine-tuned to detect data anomalies with high performance, this work opens new perspectives on how language models can be leveraged to address long-standing issues in structured data validation.

Implications for Practitioners

From an applied standpoint, our findings suggest that transformer-based models are not only theoretically interesting but also practically helpful in improving data quality workflows. The model trained to detect *Multiple Value Smells* achieved near-perfect performance, with high recall and perfect precision, meaning that data engineers and analysts can rely on it to flag problematic entries with minimal risk of false positives automatically.

Equally important, the *Split Value Smells* detector, despite a lower recall, also showed perfect precision, indicating that its predictions are highly reliable when a smell is detected. The model could thus serve as a high-confidence filter embedded in human-in-the-loop workflows, where semantic anomalies are surfaced to data curators or engineers for review.

Furthermore, the text-based input representation—where each row is serialized into a "natural language-like" form—lowers the barrier to integration. It allows this kind of model to be embedded in tools without deep coupling to schema logic or data format assumptions, making deployment across diverse data platforms more feasible.

Implications for Researchers

For researchers, this study highlights the underexplored potential of using language models for structured data quality tasks. Although most prior work in NLP and tabular learning focuses on downstream prediction or representation learning, our results demonstrate that pre-trained models, such as BERT, can be adapted to detect subtle structural inconsistencies typically overlooked by both traditional rule-based systems and data-centric AI methods.

In addition, the successful use of synthetic smells, systematically injected into otherwise clean data, demonstrates a promising way to create large-scale training datasets when labeled examples are scarce. This result could be a useful methodological direction for further research on other types of semantic anomalies, including those related to believability, completeness, or cross-column dependencies.

Implications — Summary

- BERT can act as a high-precision semantic smell detector, enabling safe automation in data quality workflows.
- The approach fits well within human-in-the-loop pipelines due to its low false-positive rate.
- This work supports a novel line of research using language models to detect contextual inconsistencies in structured data.
- Synthetic smell injection is a scalable solution to train data in semantic validation tasks.

6. Threats to Validity

Although the results presented are encouraging, several threats to validity must be acknowledged.

Internal Validity. A potential threat to internal validity concerns the synthetic nature of the datasets used. Although we used data generation techniques that simulate realistic tabular structures and semantics, the smells were injected programmatically using deterministic rules. This injection may not fully capture the complexity and variability of smells that arise in organically grown data sets. Furthermore, the use of undersampling to achieve class balance may affect the natural distribution of "smelly" vs. "clean" instances.

External Validity. The external validity of our findings may be limited due to the reliance on synthetic and semi-synthetic datasets. Although the data sets were designed to reflect common patterns in real-world data, we did not perform experiments on large-scale industry data sets or diverse domains such as finance, healthcare, or e-commerce. As such, the generalizability of the trained models across different contexts remains to be evaluated.

Construct Validity. We defined *Split* and *Multiple Value Smells* based on logical criteria and formalized these into transformation rules. However, these definitions may not encompass all the nuances recognized by domain experts. Additionally, converting tabular rows into text sequences for BERT input involves abstraction choices that may omit structural cues present in multi-column formats.

Conclusion Validity. The conclusions drawn from this study are based on standard classification metrics, including accuracy, precision, recall, and F1 score. Although these metrics provide a reliable indication of model performance, a further evaluation of operational pipelines or user-facing tools is necessary to assess practical effectiveness. Moreover, the limited number of training epochs and the small amount of available GPU resources could constrain the models' full learning potential.

7. Conclusions

This paper investigated the applicability of transformer-based language models, specifically BERT, for the automated detection of semantic data smells in tabular datasets. Our goal was to assess the extent to which BERT can be fine-tuned to identify structural anomalies that require contextual understanding—namely, *Split Value Smells* and *Multiple Value Smells*.

Our results revealed that BERT performs well in both scenarios. The model trained for *Split Value Smells* achieved high precision and generalizability, despite the intrinsic variability of how such smells manifest in real-world schemas. In contrast, the model trained to detect *Multiple Value Smells* achieved near-perfect performance, supported by the regularity of the merged value patterns. These outcomes suggest that pre-trained language models can capture subtle semantic cues within structured data when appropriately adapted.

At the same time, our findings highlight that semantic variability and feature inconsistency can impact model recall, particularly with fragmented values, underscoring the need for more comprehensive data representations and possibly multi-column modeling strategies to leverage the potential of BERT in this domain fully.

Our future research agenda includes extending this work to additional categories of semantic smells and exploring cross-dataset generalization. We also plan to integrate metadata-aware modeling and assess the effectiveness of alternative transformer architectures optimized for tabular data. Ultimately, our goal is to support the development of intelligent data validation tools that can operate at the semantic level across diverse domains and formats.

Acknowledgments

This work has been partially supported by (1) the *QUAL-AI* national research projects funded by the EU - NGEU and the MUR under the PRIN 2022 program (Contracts 2022B3BP5S), and (2) the project *"FAIR"* (PE0000013). Additionally, the work is an extended version of Nicolò Gallotta's Bachelor's thesis, developed at the University of Salerno in 2024. We gratefully acknowledge his valuable contribution to the early stages of this research.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT-4o and Grammarly to check grammar and spelling.

References

[1] S. Martínez-Fernández, J. Bogner, X. Franch, M. Oriol, J. Siebert, A. Trendowicz, A. M. Vollmer, S. Wagner, Software engineering for ai-based systems: a survey, ACM Transactions on Software Engineering and Methodology (TOSEM) 31 (2022) 1–59.

[2] D. Kreuzberger, N. Kühl, S. Hirschl, Machine learning operations (mlops): Overview, definition, and architecture, IEEE access 11 (2023) 31866–31879.

[3] L. Baier, F. Jöhren, S. Seebacher, Challenges in the deployment and operation of machine learning in practice., in: ECIS, volume 1, 2019.

[4] H. Foidl, M. Felderer, R. Ramler, Data smells: Categories, causes and consequences, and detection of suspicious data in ai-based systems, arXiv preprint arXiv:2203.10384 (2022).

[5] G. Recupito, R. Rapacciuolo, D. Di Nucci, F. Palomba, Unmasking data secrets: An empirical investigation into data smells and their impact on data quality, in: Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering-Software Engineering for AI, 2024, pp. 53–63.

[6] G. F. et al., Faker: Python package for generating fake data, 2024. https://faker.readthedocs.io/.

[7] N. Patki, R. Wedge, K. Veeramachaneni, The synthetic data vault, in: IEEE International Conference on Data Science and Advanced Analytics (DSAA), IEEE, 2016, pp. 399–410.

[8] C. Renggli, L. Rimanic, N. M. Gürel, B. Karlaš, W. Wu, C. Zhang, A data quality-driven view of mlops, IEEE Data Engineering Bulletin 44 (2021) 11–23.

[9] S. Mohammed, L. Budach, M. Feuerpfeil, N. Ihde, A. Nathansen, N. Noack, H. Patzlaff, F. Naumann, H. Harmouch, The effects of data quality on machine learning performance on tabular data, Information Systems 132 (2025) 102549.

[10] G. Recupito, F. Pecorelli, G. Catolino, S. Moreschini, D. Di Nucci, F. Palomba, D. A. Tamburri, A multivocal literature review of mlops tools and features, in: 2022 48th Euromicro Conference on Software Engineering and Advanced Applications (SEAA), IEEE, 2022, pp. 84–91.

[11] P. Li, X. Rao, J. Blase, Y. Zhang, X. Chu, C. Zhang, Cleanml: A study for evaluating the impact of data cleaning on ml classification tasks, in: 2021 IEEE 37th International Conference on Data Engineering (ICDE), IEEE, 2021, pp. 13–24.

[12] P. Yin, G. Neubig, W.-t. Yih, S. Riedel, Tabert: Pretraining for joint understanding of textual and tabular data, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 8413–8426.

[13] J. Herzig, P. K. Nowak, T. Mueller, F. Piccinno, J. Eisenschlos, Tapas: Weakly supervised table parsing via pre-training, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 4320–4333.

[14] X. Deng, H. Sun, A. Lees, Y. Wu, C. Yu, Turl: Table understanding through representation learning, ACM SIGMOD Record 51 (2022) 33–40.

[15] Z. Wang, H. Dong, R. Jia, J. Li, Z. Fu, S. Han, D. Zhang, Tuta: Tree-based transformers for generally structured table pre-training, in: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, 2021, pp. 1780–1790.

[16] H. Iida, D. Thai, V. Manjunatha, M. Iyyer, Tabbie: Pretrained representations of tabular data, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021.

[17] A. Koleva, M. Ringsquandl, M. Buckley, R. Hasan, V. Tresp, Named entity recognition in industrial tables using tabular language models, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track, 2022, pp. 348–356.

[18] C. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell, A. Wesslén, Experimentation in software engineering, Springer Science & Business Media, 2012.

[19] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, in: International Conference on Learning Representations (ICLR), 2018.

[20] G. Recupito, G. Giordano, D. Di Nucci, F. Palomba, Detecting semantic data smells with bert: Replication package, https://doi.org/10.6084/m9.figshare.29328182.v1, 2025.