A Transparent Multiclass Classification Architecture for Credit Variable Categorization

André S. G. Moreira[1], Caio V. A. Justino[1], Igor W. C. Fernandes[1], Marcos Murguel[2], Luana Oliveira[2], Anna M. C. Araujo[3], Luana F. Alvarenga[1], Renan C. Moioli[1], Cesar Renno-Costa[1]

[1]Universidade Federal do Rio Grande do Norte (UFRN)
Natal, RN – Brazil

[2]Instituto de Ciência e Tecnologia Itaú
São Paulo, SP – Brazil

[3]Itaú-Unibanco S.A.
São Paulo, SP – Brazil

{andre.soares.moreira18, lferraza}@gmail.com, {caio.justino.128, igor.fernandes.705}@ufrn.edu.br, {marcos.murguel-farias, luana.moreira-oliveira, anna.cintra-araujo}@itau-unibanco.com.br, {renan.moioli, cesar}@imd.ufrn.br

Abstract. *Integrating Alternative data in credit scoring risks revealing sensitive consumer traits. To prevent privacy violations, this study proposes an automated screening tool that categorizes textual variable descriptions into an eleven category taxonomy. To overcome labeled data scarcity, the methodology generates a synthetic dataset using the Gemma4 31B model. Because regulations demand transparency, the architecture relies exclusively on explainable machine learning models. Validation on empirical data showed high clustering cohesion, and Logistic Regression achieved 73.02% accuracy. This workflow guarantees the interpretability required for safe financial governance.*

1. Introduction

Young populations and inhabitants of developing countries often lack a formal financial history. This absence limits the ability of classical evaluation models to estimate accurate risk scores [Guo et al. 2016, Björkegren and Grissen 2018]. To overcome this limitation, many financial institutions now integrate Alternative data sources into their training datasets. This approach aims to map the routine behavior and local credit culture of each individual [Mwesigwa et al. 2026b, Tounsi et al. 2017]. The combination of these new metrics with the technological advancements of Big Data has triggered a profound transformation in the field of risk assessment and default prediction [Hurley and Adebayo 2017].

The unrestricted use of alternative data presents a technical challenge. This information can act as proxy variables. These variables indirectly reveal sensitive consumer traits [Schwarcz and Prince]. This risk demands attention given the advancement of data protection laws like the LGPD in Brazil [Brasil 2018] and the GDPR in Europe [European Union 2016]. Current legislation imposes limits. It treats each type of information differently. A macroeconomic variable operates under flexible usage rules. A

health-related variable faces severe restrictions. Processing it improperly constitutes a privacy violation. Introducing a variable into the algorithm without identifying its discriminatory weight is risky. It can compromise the legal and technical validity of the entire prediction model.

To address this risk, this study proposes building a screening tool to act as a preventive governance filter. The system functions as an automated classifier, designed to identify the nature of a new variable before integrating it into credit models. The methodology uses the taxonomy of [Çallı and Coşkun 2021] to structure this initial classification. The original model divides financial risk variables into ten categories: Socioeconomic, Demographic, Values Attitudes and Behavioral, Institutional and Financial, Personality, Situational, Educational, Macroeconomic, Health-Related, and Alternative. Mapping variables into these groups beforehand allows for the scalable separation of freely usable information from data requiring legal justification. This separation prevents protected attributes from accidentally affecting the prediction model.

Despite its utility, the original taxonomy covers only the theoretical aspect of the problem. Real-world database implementation also requires metadata processing. The study introduces an eleventh category called Operational to meet this requirement. This new class exclusively organizes system attributes, such as identification keys and timestamps. This addition provides the structural dimension necessary to ensure the automatic auditing of the real data flow.

Moving beyond the theoretical dimension to build this automated architecture introduces technical and regulatory obstacles. For the filter to classify new information automatically, the training process requires real examples. The initial challenge, therefore, is the scarcity of labeled datasets. Because banks and financial institutions keep data dictionaries isolated to comply with regulatory requirements, the Alternative variables available in the literature present heterogeneous nomenclatures and highly diverse descriptions [Mwesigwa et al. 2026a]. Beyond this training data challenge, the growing complexity of Artificial Intelligence applications raises critical concerns regarding governance in the financial sector [Kurshan et al. 2021]. The literature warns against the risks of opaque models. Consequently, adopting algorithms with intrinsic explainability becomes indispensable to ensure auditing capabilities, human oversight, and compliance with new regulatory guidelines [Anomah 2026].

To resolve the barriers of data scarcity and the need for transparency, the methodology of this work proposes a three-stage workflow. First, the project addresses the lack of data by generating a synthetic training set using the Gemma4 31B language model [Team et al. 2024]. The process grounds the sampling directly in literature concepts through the few-shot prompting technique [Brown et al. 2020]. Second, the BAAI/bge-large-en-v1.5 model [Xiao et al. 2024] vectorizes these textual inputs to capture the required semantic context. A Batch Hard Triplet Loss function [Hermans et al. 2017] calibrates this model to maximize the separation between taxonomy categories and improve the internal cohesion of each class. Finally, the architecture employs three classical algorithms to perform the final classification of the variables: Logistic Regression, Random Forest, and Support Vector Machine [Fabian 2011]. Selecting these models over complex neural networks offers the intrinsic explainability and mathematical traceability strictly necessary for an implementation that complies with current data protection legislation.

2. Related Work

The current technology market features established privacy protection solutions. The industry widely adopts tools like Microsoft Presidio [Microsoft 2026], Google Cloud DLP [Google 2026], and Amazon Macie [Amazon Web Services 2026]. These systems locate personally identifiable information and execute anonymization routines. They function at the data instance level. This means they scan actual customer inputs, like identification numbers or physical addresses.

This characteristic dictates their intervention phase and foundational technology. These solutions secure populated databases or in-transit memory streams. They rely on named entity recognition and fixed patterns to detect data leaks. This corrective model fulfills its security role. However, it requires organizations to collect the data first.

The architecture proposed in this paper complements existing enterprise systems to bridge this technical gap. Rather than inspecting physical user records, the methodology focuses on preventive governance. Conventional platforms search for fixed formats within raw data. The proposed filter evaluates the variable description to extract its semantic meaning. This strategy establishes a defensive layer directly at the project design phase. It protects credit models from inappropriate attributes long before any data collection begins.

3. Methodology

The proposed architecture is a transparent multiclass classification model designed to categorize textual descriptions into eleven macro-categories. Ten of these categories are derived from the taxonomy established by [Çallı and Coşkun 2021], comprising the following classes: Socioeconomic, Demographic, Values Attitudes and Behavioral, Institutional and Financial, Personality, Situational, Educational, Macroeconomic, Health-Related, and Alternative. This theoretical framework is complemented by an eleventh class, designated Operational, which represents a methodological expansion introduced by this study to address real-world data structures.

To improve the understanding of this architecture, Figure 1 was created to illustrate the project's pipeline, which consists of three main modules: Synthetic Data Training, Test Data Consolidation, and Inference & Evaluation. In addition, to ensure full transparency and facilitate reproducible research, the complete source code and execution instructions for this model are publicly available in a GitHub repository: https://github.com/ThisAnonymousUser/feature-distribution.

3.1. Synthetic Data Generation

Training this classification architecture is inherently constrained by the scarcity of open, labeled, and balanced datasets across all variable macro-categories. Consequently, to enable this proof of concept, a synthetic dataset was generated. This step serves strictly as an instrumental mechanism to compensate for the lack of empirical data, thereby facilitating the robust training of the predictive model.

Variable synthesis was performed using the Gemma4 31B model with 8-bit quantization [Team et al. 2024]. The temperature parameter was explicitly set to 0.4 to ensure semantic stability and reduce hallucinations during generation. Context was established

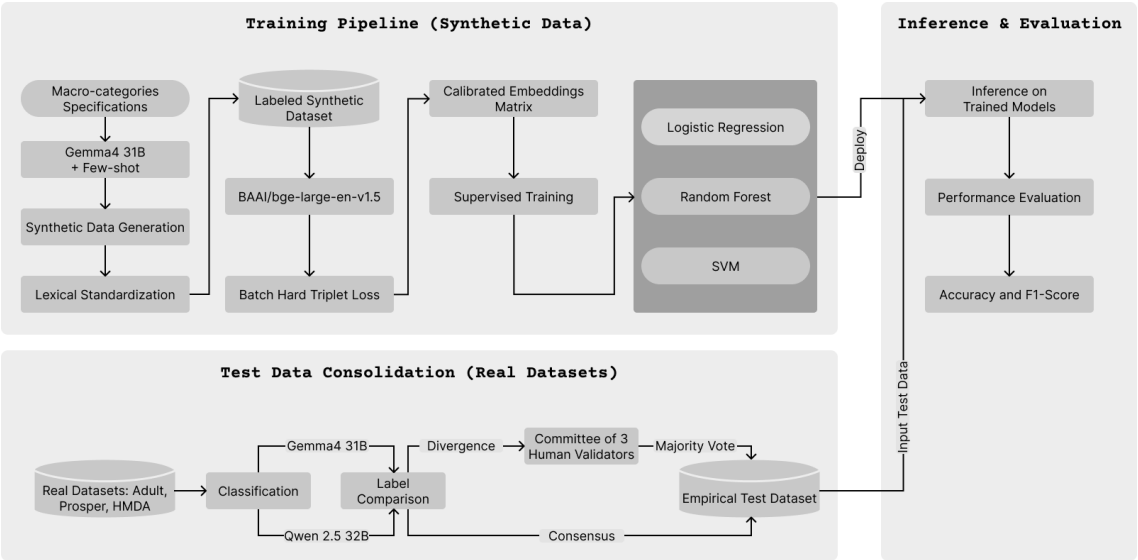


Figure 1. Diagram of the proposed transparent multiclass classification architecture, detailing the synthetic data training pipeline, the empirical test data consolidation process, and the final evaluation phase.

via few-shot prompting [Brown et al. 2020], grounding the sampling process directly in the theoretical concepts derived from [Çallı and Coşkun 2021].

The ten macro-categories proposed by [Çallı and Coşkun 2021] encompass only the theoretical dimension of the problem and fail to account for the technical structure of real-world databases. Consequently, to ensure the proper classification of all variables within production environments, it is necessary to introduce an eleventh macro-category, designated Operational. The scope of this category is strictly limited to grouping system metadata, such as primary identifiers, hashes, timestamps, and processing flags.

Subsequently, a text normalization routine was applied to the generated data. Specifically, all sentences were standardized by converting them to lowercase and removing special characters, thereby creating a clean, synthetically labeled training dataset for supervised learning.

The quality of the generated vector space and the cohesion across the eleven classes were evaluated using four clustering metrics: the Silhouette Score, the Davies-Bouldin (DB) Index, the Calinski-Harabasz (CH) Index, and Intra-class Cosine Similarity. Specifically, the Silhouette Score [Rousseeuw 1987] measured the proximity of elements within the same class relative to those in the nearest neighboring class, while the Davies-Bouldin (DB) Index [Davies and Bouldin 1979] evaluated the ratio between the internal dispersion of the clusters and their inter-cluster separation. In turn, the Calinski-Harabasz (CH) Index [Caliński and Harabasz 1974] was employed to assess the ratio of between-cluster to within-cluster variance, providing a measure of cluster separation. Furthermore, Intra-class Cosine Similarity [Salton et al. 1975] was used to verify the consistency of the semantic orientation among text vectors belonging to the exact same category.

3.2. Embedding Model

The architecture processes the textual descriptions of the variables rather than relying on their isolated labels, ensuring the model captures the semantic context required for accurate classification. As a practical example, if a variable appears in the original dataset under the label *age*, the model receives its full contextual description as input: *age of the individual in years*.

The vectorization of these inputs is performed by the foundational model BAAI/bge-large-en-v1.5 [Xiao et al. 2024]. The generated vector space is subsequently aligned with the eleven macro-categories through fine-tuning using the Batch Hard Triplet Loss function [Hermans et al. 2017]. By specifically penalizing the hardest triplets within the batch, this optimization process maximizes inter-class separation and enhances intra-class cohesion, ultimately yielding a highly calibrated embedding matrix.

3.3. Classifier Training

In the supervised modeling phase, training was conducted on the synthetic dataset, which had been previously labeled across the eleven macro-categories, using predictive features derived from the calibrated embedding matrix. This process employed three classical supervised classification algorithms: Logistic Regression, Random Forest, and Support Vector Machine (SVM) [Fabian 2011]. Using these established algorithms rather than complex opaque networks ensures the intrinsic explainability required for financial compliance.

3.4. Performance Evaluation

To evaluate the architecture's generalization capacity, the testing phase employed three real-world empirical datasets: Adult [Becker and Kohavi 1996], the Prosper Loan Dataset [Prosper Marketplace 2014], and Mortgage Data (HMDA 2017) [Federal Financial Institutions Examination Council 2025]. These specific datasets were chosen because they simultaneously provide both the target labels and detailed textual descriptions for their variables.

As these datasets lack native labels that map to the expanded taxonomy, a rigorous hybrid annotation protocol was adopted. Initially, variable classification was performed autonomously by both the Gemma4 31B [Team et al. 2024] and Qwen 2.5 32B [Yang et al. 2024] models. Instances of divergence between the predictions of these two models were referred to a committee comprising three human validators, with the final ground-truth label determined by majority vote.

4. Results

4.1. Synthetic Data Evaluation

The quantitative evaluation of the embedding space generated for the synthetic dataset demonstrated high internal cohesion and distinct separation across the evaluated macro-categories. As detailed in Table 1, the clustering metrics confirm that the generated groups are compact and well-separated. Specifically, the high Intra-class Cosine Similarity indicates a robust semantic alignment among vectors belonging to the same synthetic class, validating the quality of the data generation phase.

Table 1. Evaluation metrics for the embedding space generated from the synthetic dataset.

Metric	Value Obtained
Silhouette Score	0.768
Davies-Bouldin (DB) Index	0.344
Calinski-Harabasz (CH) Index	1742.9
Intra-class Cosine Similarity	0.946

4.2. Visual Analysis of Data Projection

To visually assess the calibrated embedding model's ability to align textual representations from different domains, the vectors from the synthetic (training) and empirical (testing) datasets were jointly projected into a two-dimensional space. Dimensionality reduction was performed using the UMAP algorithm [McInnes et al. 2020], as shown in Figure 2.

In both the local-focused and global-focused projections, eleven well-defined clusters emerge, corresponding directly to the theoretical macro-categories. Notably, the synthetic samples (circles) and the real samples (black-outlined diamonds) occupy the exact same regions within each respective cluster. This substantial overlap indicates that texts from both sources share similar semantic representations in the vector space, with no isolated subclusters emerging based on data provenance.

4.3. Predictive Performance of the Models

Table 2 presents the overall accuracy achieved by the classifiers across the evaluated datasets.

Table 2. Overall accuracy of the classification models across the test datasets.

Model	Adult	HMDA 2017	Prosper Loan Dataset	Complete Set
Logistic Regression	0.6667	0.6333	0.7778	**0.7302**
Random Forest	0.6667	0.5667	0.7778	0.7143
Linear SVM	0.6667	0.6000	0.7778	0.7222

Logistic Regression demonstrated the best overall performance and the highest predictive consistency across the individual evaluations. Interestingly, all three models achieved identical accuracy on both the Adult and Prosper Loan datasets. In contrast, the HMDA 2017 dataset presented the most complex generalization scenario, yielding the lowest accuracy across all tested algorithms.

The classification report and confusion matrix (Figure 3) further detail the performance of the best model on the complete test set in each category. It showed balanced results in the Macroeconomic class (precision and recall of 0.83 in 6 instances) and in the Demographic class (precision of 0.73 and recall of 0.89 in 9 instances). The Operational category showed more moderate results, registering a precision of 0.67, a recall of 0.59, and an F1 score of 0.62 in 17 instances.

For the majority class, Institutional and Financial with 64 instances, the algorithm achieved a high recall of 0.97 but a lower precision of 0.74. This discrepancy indicates

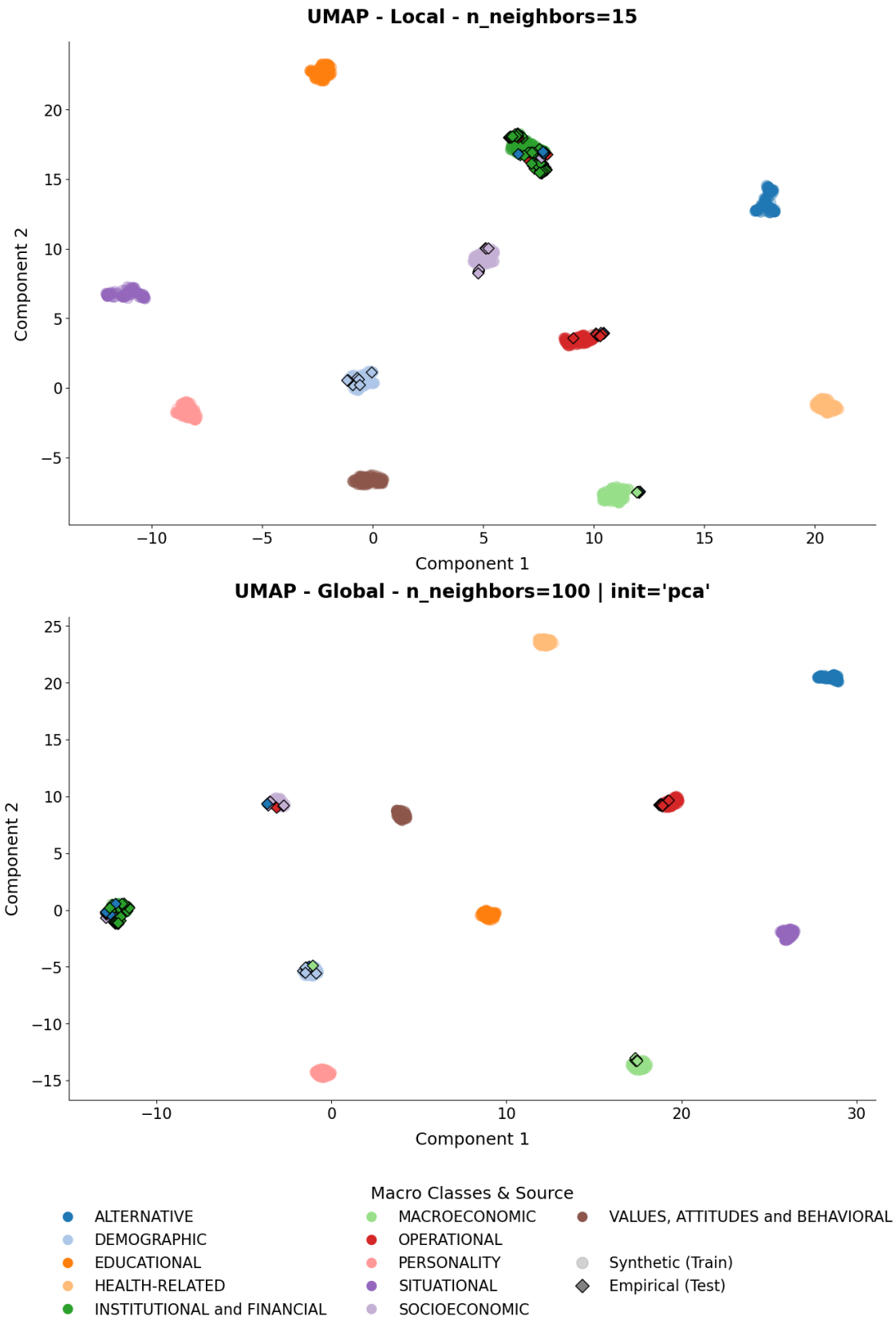


Figure 2. Two-dimensional UMAP visualization of the alignment between the synthetic (train) and empirical (test) domains. The panels contrast the preservation of local structure (`n_neighbors=15`, top) with the capture of global structure (`n_neighbors=100` and `init='pca'`, bottom) within the embedding space.

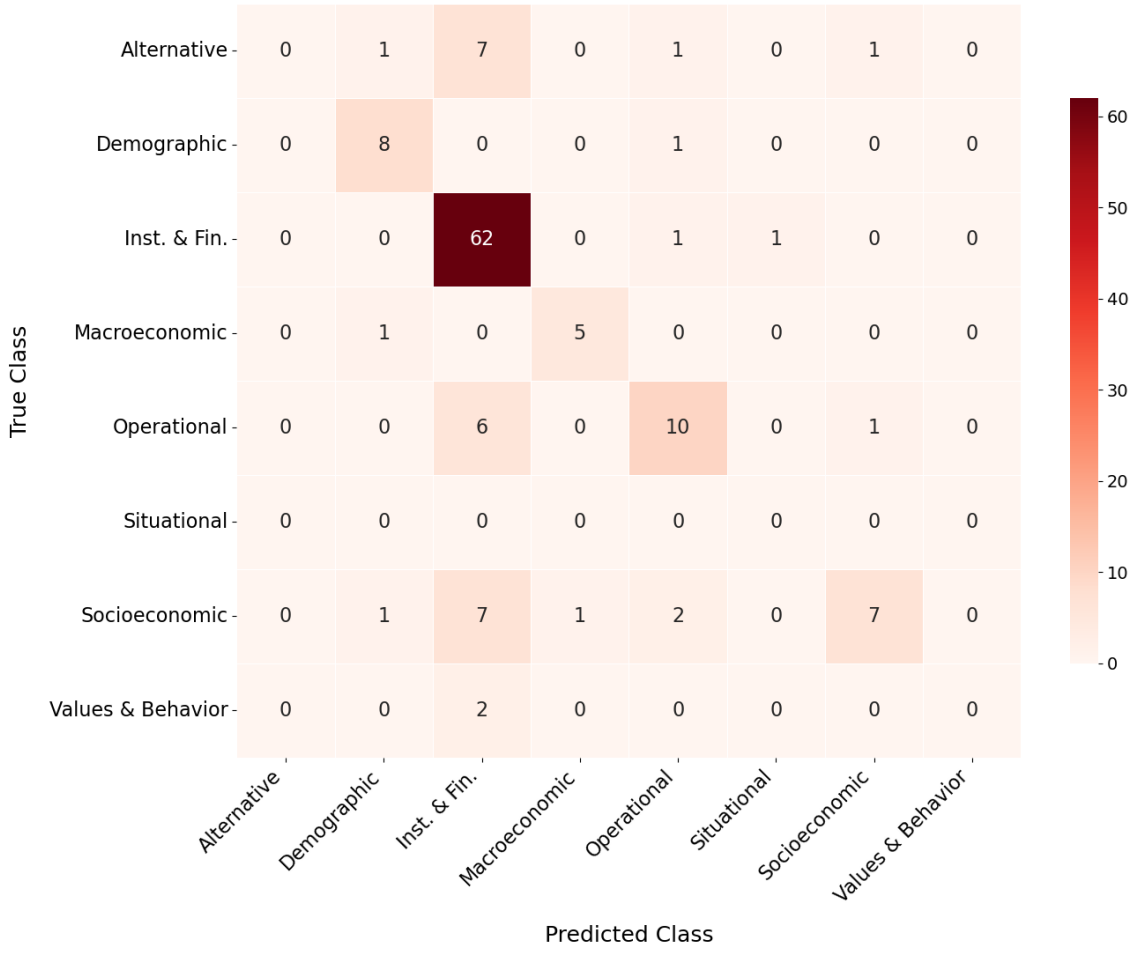


Figure 3. Heatmap detailing the absolute distribution of true positives, false negatives (omissions), and false positives (false alarms) within the complete test set.

a concentration of false positives. Conversely, the Socioeconomic category recorded a precision of 0.78, but its low recall of 0.39 across 18 instances highlights a substantial rate of false negatives.

Furthermore, the classifier failed to correctly predict any instances for the Alternative and Values Attitudes and Behavioral classes, which comprise 10 and 2 instances respectively. Both categories resulted in a precision and recall of 0.00, generating pure false negatives. Finally, the Situational, Personality, Educational, and Health-Related categories were completely absent from the test dataset. While the latter three received no predictions, the model incorrectly predicted one instance from the Institutional and Financial class as Situational, generating a single false positive.

5. Discussion

The results indicate that the embedding model properly clustered the eleven macrocategories. The Silhouette Score of 0.768 and the Davies-Bouldin Index of 0.344 show that the classes are well-separated with low internal dispersion. Furthermore, the Calinski-Harabasz Index of 1742.9 reinforces this geometric distance between the clusters.

On the other hand, the Intra-class Cosine Similarity reached 0.946. This high index indicates that the Gemma4 31B model generated samples with little vocabulary variation. The lack of natural noise found in real texts reduces semantic diversity during training, which hinders the architecture's generalization in specific contexts. This limitation explains the performance drop on the HMDA dataset, where accuracies ranged from 0.5667 to 0.6333 due to its strict governmental and regulatory pattern. This scenario differs from the broader Socioeconomic and financial descriptions of the Prosper Loan Dataset, which achieved the best performance in the evaluation (0.7778), while the Adult dataset also showed satisfactory generalization (0.6667).

The evaluation on the complete test set indicated a satisfactory overall performance for the architecture. The results showed a technical alignment among the classifiers, with Logistic Regression achieving a global accuracy of 0.7302, closely followed by the Linear SVM (0.7222) and Random Forest (0.7143). This equivalence demonstrates that the embedding matrix establishes stable decision boundaries, allowing simple linear algorithms to reach the same efficiency as more complex models.

Since Logistic Regression achieved the best performance in this stage, with an accuracy of 0.7302, the detailed error analysis will focus on its confusion matrix, shown in figure 3. The evaluation reveals an imbalance in the separation of certain categories, indicating that the linear model tends to allocate the most difficult and ambiguous cases to the dominant classes. The Institutional and Financial class attracted the highest volume of errors in the set, recording 22 false positives, which shows that its classification criterion became too broad in the vector space. In contrast, the Socioeconomic class presented an overly strict boundary, reflecting 11 false negatives against only 7 correct predictions, with a recall of 0.39. This imbalance directly impairs the segmentation of residual categories, as evidenced by the complete lack of correct predictions in the Alternative class, with 10 false negatives across 10 samples, and the Values Attitudes and Behavioral class, with 2 false negatives in 2 instances.

However, this difficulty with residual or highly abstract clusters does not invalidate the architecture, which demonstrated strong generalization and success in retrieving the other categories. The Macroeconomic and Demographic classes showed high predictive performance and well-defined boundaries; the former reached a precision of 0.83, correctly classifying 5 out of 6 available instances with only 1 omission, while the latter was retrieved highly efficiently, with a recall of 0.89. In addition to these results, the model's practical viability was reinforced by its performance on the newly introduced Operational class, which recorded 10 correct predictions and an F1-Score of 0.62.

Finally, the global evaluation demonstrates the method's viability. The main advantage of the proposed architecture is its modular structure, which allows for the isolated replacement of the synthetic data generator or the embedding model without compromising the rest of the pipeline. This flexibility facilitates future experiments with new encoders and the injection of targeted noise to increase diversity within higher-level semantic categories, aiming for accuracy levels above 0.9. Furthermore, by delegating the final prediction to classical algorithms, the trained model operates deterministically. Unlike direct classification via generative models, which lack interpretability and require complex *post hoc* linguistic and gradient analyses to explain their decisions [Ribeiro et al. 2024], this approach ensures the mathematical traceability of decision boundaries. This delivers

the exact level of auditability required for safe implementation in production environments, aligning directly with the legal and ethical considerations necessary for robust data governance, algorithmic accountability, and reliable AI systems in highly regulated domains [Alvarenga et al. 2026].

References

Alvarenga, L. F., Ferreira, N. F., Peres, A. S. C., Moioli, R. C., and Rennó-Costa, C. (2026). Legal and ethical considerations for translating federated learning into cross-border healthcare innovation. *IEEE Journal of Biomedical and Health Informatics*, (1).

Amazon Web Services (2026). Amazon macie documentation. AWS Official Documentation.

Anomah, S. (2026). Bridging the interpretability gap: integrating explainable ai into public sector audit systems in emerging economies. *Accounting Research Journal*, pages 1–27.

Becker, B. and Kohavi, R. (1996). Adult. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5XW20.

Björkegren, D. and Grissen, D. (2018). The Potential of Digital Credit to Bank the Poor. *AEA Papers and Proceedings*, 108:68–71.

Brasil (2018). Lei nº 13.709, de 14 de agosto de 2018. lei geral de proteção de dados pessoais (lgpd). Accessed: Jun 21, 2026.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Caliński, T. and Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27.

Çallı, B. A. and Coşkun, E. (2021). A longitudinal systematic review of credit risk assessment and credit default predictors. *Sage Open*, 11(4).

Davies, D. L. and Bouldin, D. W. (1979). A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2):224–227.

European Union (2016). Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (general data protection regulation). Official Journal of the European Union.

Fabian, P. (2011). Scikit-learn: Machine learning in python. *Journal of machine learning research 12*, page 2825.

Federal Financial Institutions Examination Council (2025). Home Mortgage Disclosure Act (HMDA) Data. FFIEC / CFPB.

Google (2026). Cloud data loss prevention documentation. Google Cloud Official Documentation.

Guo, G., Zhu, F., Chen, E., Liu, Q., Wu, L., and Guan, C. (2016). From Footprint to Evidence: An Exploratory Study of Mining Social Data for Credit Scoring. *ACM Trans. Web*, 10(4):22:1–22:38.

Hermans, A., Beyer, L., and Leibe, B. (2017). In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*.

Hurley, M. and Adebayo, J. (2017). Credit scoring in the era of big data. *Yale Journal of Law and Technology*.

Kurshan, E., Shen, H., and Chen, J. (2021). Towards self-regulating ai: challenges and opportunities of ai model governance in financial services. In *Proceedings of the First ACM International Conference on AI in Finance*, ICAIF '20, New York, NY, USA. Association for Computing Machinery.

McInnes, L., Healy, J., and Melville, J. (2020). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv:1802.03426 [stat.ML].

Microsoft (2026). Microsoft presidio: Data protection and anonymization sdk. Microsoft GitHub Repository.

Mwesigwa, D., J. Jackson, S., and Csikszentmihalyi, C. (2026a). Risk, data, alignment: Making credit scoring work in kenya. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI '26, New York, NY, USA. Association for Computing Machinery.

Mwesigwa, D., Jackson, S. J., and Csikszentmihalyi, C. (2026b). Risk, Data, Alignment: Making Credit Scoring Work in Kenya. Version Number: 1.

Prosper Marketplace (2014). Prosper Loan Data. Kaggle.

Ribeiro, M., Malcorra, B., Mota, N. B., Wilkens, R., Villavicencio, A., Hubner, L. C., and Rennó-Costa, C. (2024). A methodology for explainable large language models with integrated gradients and linguistic analysis in text classification. *arXiv*, (arXiv:2410.00250).

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

Salton, G., Wong, A., and Yang, C.-S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

Schwarcz, D. and Prince, A. E. R. Proxy Discrimination in the Age of Artificial Intelligence and Big Data.

Team, G., Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., Sifre, L., Rivière, M., Kale, M. S., Love, J., et al. (2024). Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Tounsi, Y., Hassouni, L., and Anoun, H. (2017). Credit scoring in the age of Big Data – A State-of-the-Art. 15(7).

Xiao, S., Liu, Z., Zhang, P., Muennighoff, N., Lian, D., and Nie, J.-Y. (2024). C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pages 641–649.

Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Dong, G., et al. (2024). Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.