The Impact of Data Imbalance on the Explainability of Text Classification Models: A Comparative Analysis with SVM and LSTM

Luirys do Nascimento Silva, Cleyton Mário de Oliveira Rodrigues

[1]Graduate Program in Computer Engineering – University of Pernambuco (UPE)
School of Engineering of Pernambuco (POLI) – 50.720-001 – Recife – PE – Brazil

luirys.nsilva@upe.br, cleyton.rodrigues@upe.br

Abstract. *This paper investigates the systematic impact of data imbalance and sampling strategies on the explainability of text classification models. By confronting a linear SVM with a LSTM network, we analyze the mathematical consistency of local post-hoc explanations generated by LIME and SHAP, alongside feature relevance and deep attention mechanisms across five datasets. Non-parametric statistical evaluations via Friedman's test reveal that while SHAP maintains high stability due to its axiomatic foundation, LIME and attention mechanisms exhibit severe structural vulnerabilities. Notably, oversampling text sequences for LSTM triggers localized redundancy, causing attention weights to oscillate erratically and undermining explanation reliability.*

1. Introduction

The dissemination of machine learning models and deep neural networks in high-criticality domains has consolidated Explainable Artificial Intelligence (XAI) as an indispensable pillar, mitigating black box behavior in favor of algorithmic transparency [ARAÚJO 2020]. However, current literature in XAI frequently assumes idealized scenarios of equitable distribution between classes. In practice, this premise collides with the nature of realistic unstructured data, such as texts from social media or journalistic portals, in which severe imbalance is inherent and compromises both predictive generalization capacity and the fidelity of the generated explanations [POCEBON 2023].

Despite the methodological implications of this phenomenon, the literature neglects the intersection between data asymmetry and explanation robustness [DABLAIN 2024]. Previous efforts focus on the validation of explanatory frameworks under controlled and artificially balanced environments, generating severe gaps regarding how such tools behave when faced with distortions of empirical distributions [INDRAHIMAWAN 2023]. This lack is critical in domains such as sentiment analysis in the financial market, marked by the massive predominance of neutral narratives, high semantic volatility, and a regulatory demand for auditable justifications in economic decision-making [PERES 2019].

To fill this gap, this article investigates the systematic impact of sampling strategies (*undersampling* and *oversampling*) on two inseparable dimensions: the predictive performance of classifiers and the stability of their associated explanations. The experimental architecture confronts a classic linear approach, based on *Support Vector Machine* (SVM), with a *Long Short-Term Memory* (LSTM) recurrent neural network. From

the perspective of interpretability, the consistency of post-training explanations is evaluated quantitatively and qualitatively, employing the LIME (*Local Interpretable Model-agnostic Explanations*) and SHAP (*SHapley Additive exPlanations*) frameworks, in addition to the comparative mapping of relevance via coefficient magnitude (SVM) and attention mechanism weights (LSTM).

By revealing the functional nuances between balancing techniques and the post-training behavior of XAI methods, this study provides practical and conceptual support for the engineering of more transparent and reliable AI systems when subjected to imperfect data. The remainder of this article is structured as follows: Section 2 revisits the theoretical foundations and related works; Section 3 details the methodology and the experimental pipeline; Section 4 reports the obtained results and the analytical discussion; finally, Section 5 synthesizes the main conclusions and outlines directions for future work.

2. Theoretical Foundations

2.1. Data Imbalance and Sampling Strategies

Class imbalance occurs when dataset frequencies show severe statistical disproportion, biasing inductive algorithms toward the majority class due to standard global loss functions [SILVA 2025]. Pre-processing sampling strategies reconfigure class densities in the feature space: *undersampling* removes majority instances (heuristically or probabilistically) [MENDES et al. 2019], whereas *oversampling* replicates or synthesizes minority instances. However, unrestricted artificial space expansion risks overfitting by memorizing sample noise [MOHAMMED 2020].

2.2. Predictive Models: SVM and LSTM

Support Vector Machines (SVM) implement Structural Risk Minimization (SRM) to compute a maximum-margin linear hyperplane in high-dimensional feature spaces, guaranteeing global convergence and robustness on sparse TF-IDF text matrices [CORTES 1995, CHANDRA 2021].

Long Short-Term Memory (LSTM) networks address vanishing/exploding gradients in recurrent architectures via a continuous cell state regulated by three non-linear gates [Hochreiter 1997]: the *forget gate* purges irrelevant historical state info, the *input gate* incorporates new textual stimuli, and the *output gate* filters cellular memory into the next hidden state [BADJATIYA 2017]. This dynamic preserves long-range semantic dependencies in complex textual sequences.

2.3. Explainable Artificial Intelligence (XAI)

XAI frameworks aim to restore transparency and mitigate inductive biases in complex black box models deployed in high-criticality domains [MINH 2022, VILONE 2021]. LIME (*Local Interpretable Model-agnostic Explanations*) fits a local linear surrogate model (e.g., ridge regression) on stochastically perturbed instances weighted by proximity, approximating local decision boundaries to measure feature impact [Zafar and Khan 2021].

SHAP (*SHapley Additive exPlanations*) relies on cooperative game theory to compute Shapley values, calculating average marginal feature contributions across all coali-

tion permutations [Roth 1988]. SHAP satisfies efficiency, symmetry, and additivity axioms, providing stable local and global attributions despite high computational complexity.

Finally, *Attention Mechanisms* assign intrinsic dynamic probability distributions over input tokens to highlight key semantic axes [WANG 2018]. While attention visualization offers direct insights into neural activation paths, its fidelity as a strict causal explanation remains debated [Zafar and Khan 2021].

2.4. Related Works

The intersection between class distribution asymmetry and the fidelity of interpretability frameworks constitutes a recent frontier, restricted to specific data domains. In the field of visual data, Dablain et al. [DABLAIN 2024] investigated the impact of imbalance on the internal representations of Convolutional Neural Networks (CNNs). Using saliency maps and neighborhood approaches to categorize latent instances, the study demonstrated that predictive degradation stems from class overlap in the projection spaces. The authors showed that traditional algorithmic interventions, such as stochastic sampling or cost-sensitive loss functions, partially mitigate global metrics but fail to reorganize the model's internal decision boundaries.

From the perspective of structured and tabular data, investigations on sample reconfiguration techniques (*Random Oversampling*, *Undersampling*, and SMOTE) associated with algorithms such as *Random Forest*, SVM, and XGBoost indicate consistent increments in predictive metrics [SILVA 2025]. However, these early efforts omitted how artificial modifications in sample density alter internal decision paths. Addressing this gap, Chen et al. [Chen et al. 2023] demonstrated on credit scoring data that progressive class imbalance significantly degrades the stability of post-hoc explanations generated by LIME and SHAP. Similarly, in healthcare, Rai et al. [Rai et al. 2025] evaluated LIME, SHAP, and Partial Dependence Plots (PDPs) on electronic health records, confirming that class skewness introduces severe inconsistency in feature attributions compared to balanced distributions.

To mitigate sampling bias while maintaining interpretability, architectural alternatives have also emerged. Fujiwara [Fujiwara 2024] combined Random Undersampling (RUS) with Knowledge Distillation (KD) to train student models that preserve SHAP explanation stability across cross-validation folds. While these recent studies validate the adverse effects of class asymmetry on visual, tabular, and clinical domains, they do not address unstructured textual environments.

Thus, the state of the art reveals a critical gap: the scarcity of integrated analyses measuring the covariance between population sampling strategies and the invariance of XAI explanations under textual data distributions. The present study positions itself to bridge this gap by conducting a joint evaluation of performance and post-sampling interpretability, contrasting agnostic (LIME/SHAP) and intrinsic (attention mechanisms) explanations across classic convex hyperplanes (SVM) and deep recurrent architectures (LSTM).

3. Methodology

The proposed methodological approach adopts a systematic flow structured into three operational layers (lanes) mapped according to the BPMN (*Business Process Model and Notation*) notation, as illustrated in Figure 1.

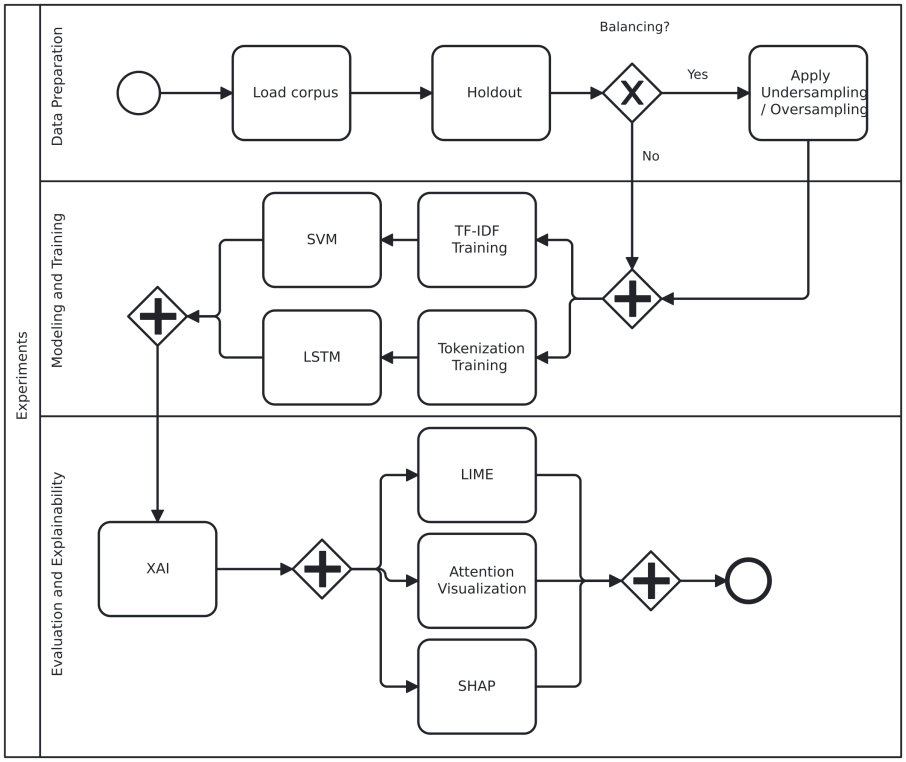


Figure 1. Figure showing the structured systematic flow following BPMN notation.

3.1. Data Preparation

The pipeline begins with the loading of the textual corpus (*Load corpus*), followed by data partitioning via the *Holdout* strategy to isolate the training and testing sets. To mitigate class imbalance, a conditional gateway (*Balancing?*) directs the flow through two controlled pathways: (i) **Scenario Without Balancing (No):** direct advancement to modeling with the original distribution; or (ii) **Scenario With Sampling (Yes):** application of undersampling or oversampling techniques (*Apply Undersampling/Oversampling*) on the training set. Both flows subsequently converge through a join gateway.

3.2. Modeling and Training

A divergent parallel gateway (+) simultaneously triggers two independent feature engineering and model training pipelines:

- **Classic Pipeline (SVM):** Feature extraction by statistical vectorization (*TF-IDF Training*) followed by the adjustment of the *SVM* classifier.
- **Deep Learning Pipeline (LSTM):** Sequential text processing via tokenization and padding (*Tokenization Training*) for feeding and training the *LSTM* neural network.

The conclusion of both fronts is consolidated by a convergent parallel gateway (+).

3.3. Evaluation and Explainability

After the unification of the flow, the trained models enter the Explainable Artificial Intelligence (*XAI*) layer. First, the overall predictive performance on the test base is quantified through traditional metrics (*Evaluate Metrics*). Next, a parallel gateway (+) divides the flow to conduct the interpretability analysis under three complementary perspectives:

1. **LIME:** Local agnostic explanations to evaluate the weight of terms in individual predictions.
2. **Attention Visualization:** Mapping of the internal attention weights of the LSTM architecture to identify contextual patterns captured by the network.
3. **SHAP:** Local and global feature importance attribution based on cooperative game theory.

The process ends with the joint export of the generated metrics and visual interpretations (*End Event*).

3.4. Explanation Consistency Metric

To quantify explanation stability under resampling, the y-axis in Figures 2, 3, and 4 represents the *Explanation Consistency Index*, calculated using Spearman's rank correlation coefficient (ρ) between the feature attribution vectors of the imbalanced baseline ($\mathbf{a}^{(u)}$) and the resampled model ($\mathbf{a}^{(s)}$):

$$C = \rho\left(\mathbf{a}^{(u)}, \mathbf{a}^{(s)}\right) = 1 - \frac{6\sum_{k=1}^{d} d_k^2}{d(d^2-1)} \tag{1}$$

where d_k is the rank difference for feature k ($d = 10$). Values near 1 indicate stable feature rankings, 0 indicates no correlation, and negative values ($C < 0$) represent rank inversions caused by resampling.

3.5. Class Distribution and Sample Space

The statistical characterization and descriptive properties of the datasets selected for the empirical validation of this study are summarized in Table 1. The data portfolio was strategically designed to cover different degrees of asymmetry, label cardinalities, average text lengths, and vocabulary sizes, emulating the distributive distortions inherent in real production environments.

Table 1. Summary and descriptive statistics of the evaluated textual datasets.

Dataset	Domain	Classes	Avg. Length	Vocab Size	Selection Criteria
Base 1	Financial Sentiment	neutral, negative, positive	24 words	10.1k	Moderate multi-class imbalance
Base 2	Spam Detection	spam, ham	16 words	8.7k	Severe binary imbalance
Base 3	Airline Tweets	neutral, negative, positive	18 words	15.0k	Social media lexical volatility
Base 4	Fake News Net	fake, real	11 words	17.5k	High imbalance in short headlines
Base 5	Clean Fake News	fake, real	12 words	20.8k	Balanced control baseline

The textual dataset of a financial nature (*Base_1*)[1] configures a ternary multiclass scenario, in which the majority class (typically neutral) displays a magnitude approximately five times greater than the minority class. This severe disproportion configuration is repeated in the corpus of messages from the social network *X/Twitter* (*Base_3*)[2], consolidating an evaluation pipeline for multiclass scenarios under strong sample density asymmetry.

In the scope of binary classification problems, the sets directed toward spam detection (*Base_2*)[3] and fake news mapping via *Base_4*[4] reproduce the classic topology of anomaly auditing tasks, in which the event of interest (the positive class) constitutes a sparsely occurring phenomenon relative to legitimate instances.

In contrast, the dataset *Base_5*[5] exhibits an approximately equitable frequency distribution among its labels, being introduced into the experimental design as an idealized control and reference group. The coexistence of textual matrices with varying gradients of imbalance enables a robust investigation into how the class distribution entropy modulates the mathematical stability of post-training explainability techniques.

3.6. Experimental Configuration

To ensure reproducibility, the architectural and XAI hyperparameter configurations used across all experiments are summarized in Table 2. The predictive models were trained using standard optimization settings, while post-hoc explanation frameworks utilized model-appropriate explainers to maintain evaluation consistency.

Table 2. Hyperparameter settings for predictive models and XAI frameworks.

Component	Hyperparameter	Value
LSTM	Max Vocabulary / Seq. Length	5,000 / 100
LSTM / Embedding	Hidden Units / Dim	128 / 128
Training	Epochs / Batch Size	10 / 32
Training	Optimizer / Loss	Adam (default) / Sparse Cat. Cross-Entropy
LIME	Features / Perturbations	10 / Default (5,000)
SHAP (SVM)	Explainer / Background	LinearExplainer / Up to 100 instances
SHAP (LSTM)	Explainer / Background	KernelExplainer / Up to 100 instances

4. Results and Discussion

To verify whether the observed consistency variations were statistically significant, a non-parametric Friedman test was conducted across the experimental configurations. The test

[1] https://www.kaggle.com/datasets/ankurzing/sentiment-analysis-for-financial-news

[2] https://www.kaggle.com/datasets/tango911/airline-sentiment-tweets

[3] https://www.kaggle.com/datasets/abdallahwagih/spam-emails

[4] https://www.kaggle.com/datasets/algord/fake-news

[5] https://www.kaggle.com/datasets/bornaetminan/cleaned-real-and-fake-news-dataset-english

yielded a chi-square statistic of $\chi^2 = 10.23$ with 5 degrees of freedom ($df = 5$) and a p-value of $p = 0.069$. Although this result indicates a marginal statistical trend ($p < 0.10$), it does not reject the null hypothesis at the standard significance threshold ($\alpha = 0.05$), suggesting that performance differences across configurations should be interpreted as directional patterns rather than strict statistically distinct rankings.

4.1. Attribution Analysis via LIME Values

The consistency indices of the post-training explanations generated by the LIME framework, mapped over the SVM and LSTM architectures under different sampling scenarios, were subjected to statistical validation. To verify the significance of the observed variations without assuming the premise of data normality, the non-parametric Friedman test was applied. Figure 2 illustrates the population dispersion and the ranking of these indices using boxplots.

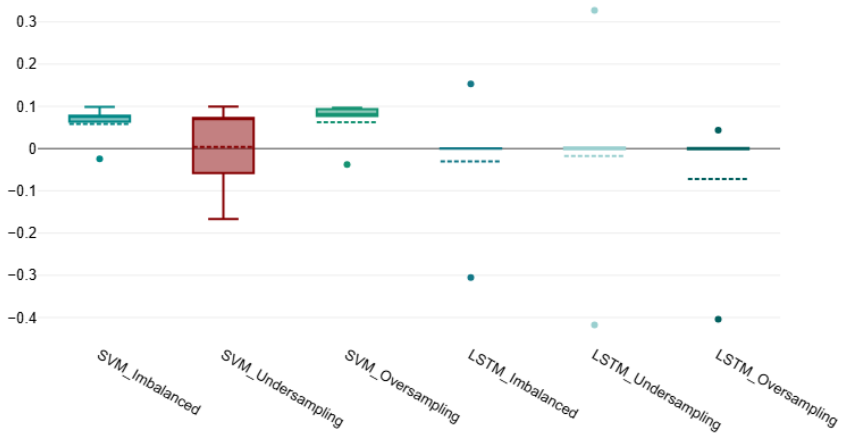


Figure 2. Non-parametric distribution and statistical validation of LIME consistency via Friedman test and boxplots.

The distributional analysis indicates that the mathematical stability of local explanations exhibits notable variation depending on the triadic combination of classifier topology, dataset nature, and the applied sampling strategy. As a primary pattern revealed by the non-parametric distribution, the linear SVM classifier demonstrated greater overall interpretative resilience, concentrating narrower interquartile dispersion boxes and medians close to zero. However, the scenario under *undersampling* broke this stability, drastically expanding the lower range of the box.

In the ternary multiclass domain under severe asymmetry (*Base_1* and *Base_3*), the introduction of sampling techniques triggered opposite behaviors between the classic and deep pipelines. For *Base_1*, *undersampling* reached local nominal maximums of 0.0996 in SVM and 0.0004 in LSTM. However, when evaluating *Base_3* (X/Twitter data), the application of *undersampling* resulted in severe consistency degradations, culminating in markedly negative valuations (-0.1667 in SVM and -0.4178 in LSTM). As evidenced in the non-parametric mapping of Figure 2, the LSTM-based models operating on modified data (LSTM_Undersampling and LSTM_Oversampling) generated explanations punctuated by critical outliers at the extremes of dispersion, extending to the threshold of -0.4.

This abrupt divergence, validated by the rank variance of the Friedman test, suggests a phenomenon of local instability in the LIME estimator. When the deep network (LSTM) is fitted over artificially modified sample spaces, especially via *undersampling* in bases with high lexical volatility, the stochastic perturbations generated by LIME in the local neighborhood fail to convexly approximate the highly non-linear decision boundary of the model. Finally, in the balanced control group (*Base_5*), the variations were statistically marginal, confirming that the impact of rebalancing on post-hoc fidelity is conditioned on the original distributive entropy of the corpus.

4.2. Analysis of Visualizations Based on Attention and Relevance Mechanisms

The consistency profiles linked to attributions based on attention mechanisms (and coefficient magnitude for the linear pipeline) were analytically mapped. The respective non-parametric distribution, validated via the Friedman test, is illustrated in the boxplots of Figure 3, detailing the behavioral variability of the predictive pipelines.

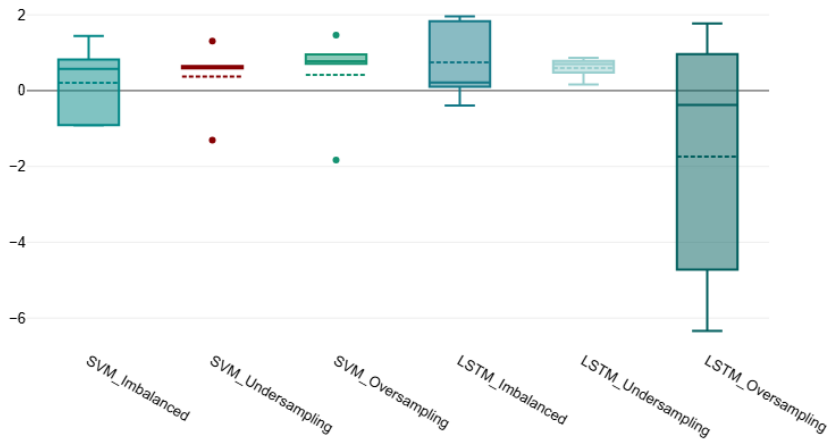


Figure 3. Non-parametric Friedman distribution and interquartile dispersion mapping for attention and linear magnitude explanations.

Contrary to the stability observed in the LIME framework, the explanatory profiles derived from attention and relevance revealed drastic mathematical volatility, with coefficients oscillating across wide scales. In controlled environments and under originally balanced or lower-complexity distributions (*Base_1* and *Base_2*), resampling strategies induced expressive increases in consistency. *Oversampling* maximized interpretability in *Base_1* for both SVM (0.9539) and LSTM (0.9629), indicating that the homogenization of class densities stabilizes the attribution of semantic weights in scenarios of high regulatory complexity.

In contrast, specific scenarios triggered structural collapses in explanation fidelity. The most critical phenomenon manifests itself in the *LSTM_Oversampling* configuration. As visually mapped by the elongated boxplot in Figure 3, this scenario introduced a massive and asymmetrical interquartile dispersion, extending its lower limit beyond -6, a behavior driven by the severe drops observed in *Base_4* (-4.7209) and *Base_5* (-6.3414), both directed at the fake news detection domain.

This instability points to a side effect of text instance replication via *oversampling* on deep networks. By artificially inflating the minority class through strict duplication of sequential vectors, the LSTM loss function converges by creating excessively narrow and redundant optimization valleys. During post-hoc mapping, minor lexical variations or perturbations in the hidden state flow cause the attention probability distribution to jump abruptly between competing tokens. Thus, the mechanism loses the capacity to weigh the real causal relevance of terms, generating mathematically unstable explanatory patterns despite eventual improvements in aggregated predictive metrics.

4.3. Attribution Analysis via Shapley Values (SHAP)

The explanatory consistency computed from the SHAP framework was subjected to the same statistical scrutiny. The non-parametric distribution associated with these results, evaluated under the rank metric of the Friedman test, is presented through the boxplots in Figure 4.

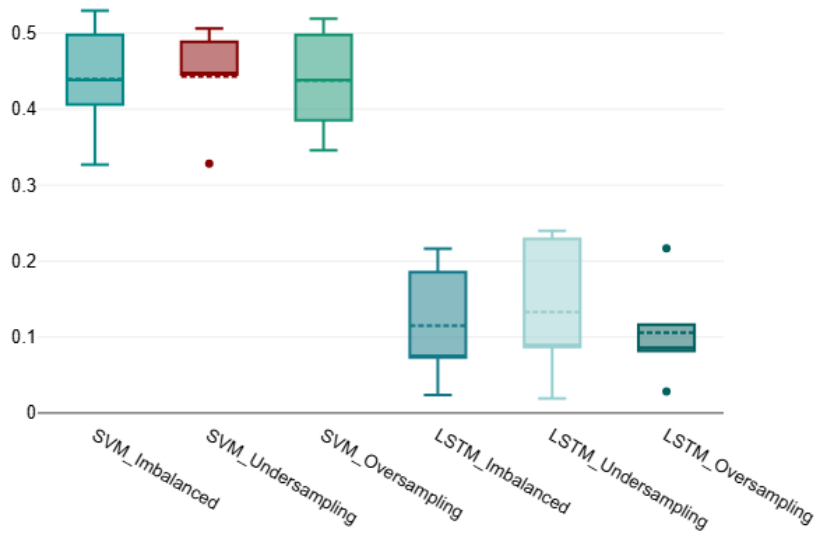


Figure 4. Non-parametric Friedman distribution and dispersion analysis for SHAP local attributions.

Empirical inspection of the dispersion data reveals a distinctly different behavior compared to the LIME and attention frameworks. The method based on Shapley values demonstrated markedly higher absolute consistency levels and greater robustness against extreme fluctuations. As evidenced in Figure 4, the linear SVM classifier operated within a high stability range, concentrating its non-parametric rank medians consistently above the 0.4 threshold.

The oscillations in SVM fidelity indices between the original imbalanced scenarios and the sampling pipelines proved to be statistically marginal. In *Base_4*, for instance, the consistency index remained strictly invariant at the 0.4980 level in both the original scenario and under *oversampling*. This trend indicates that the convex and linear formulation of SVM, when combined with the axiomatic foundations of additivity and efficiency

of SHAP, shields the model's interpretation against perturbations in the population density of classes.

On the other hand, the deep pipeline based on LSTM manifested greater sensitivity and volatility in relation to resampling strategies, although maintaining well-behaved interquartile dispersions and without the structural collapses observed in the attention mechanisms. The application of *undersampling* acted beneficially on distributions with strong textual asymmetry (*Base_3* and *Base_4*), raising the consistency of LSTM to its local peaks of 0.2399 and 0.2295, respectively. As geometrically illustrated by the purple block in Figure 4 (*LSTM_Undersampling*), the artificial balancing of classes positively expanded consistency dispersion into upper quadrants, approaching the behavior of the linear model.

This dynamic signals that SHAP, by computing combinatoric marginal contributions across all permutations of feature subsets, manages to capture more faithfully the alterations imposed by sampling techniques on the topology of weights in deep networks. Although the absolute consistency values of LSTM remain lower than those of SVM due to the stochastic and highly non-linear nature of its activation functions, the SHAP framework consolidates itself as the most stable and reliable XAI tool to audit the balance between predictive performance and algorithmic transparency in scenarios with severe data imbalance.

5. Final Considerations

This work systematically investigated the impact of statistical sampling strategies (*undersampling* and *oversampling*) on the stability and fidelity of post-training explanations in text classification scenarios with severe data imbalance. By confronting a classic linear approach (SVM) with a deep recurrent architecture (LSTM), the research revealed that interventions in the sample space, though frequently necessary to optimize raw predictive metrics, introduce critical perturbations to the internal semantics and reliability of XAI frameworks.

The experimental results, backed by the non-parametric validation of the Friedman test, revealed distinct behaviors among the evaluated explanatory methods. The SHAP framework consolidated itself as the most robust and resilient approach when faced with population distribution changes. Due to its mathematical foundations based on cooperative game theory and the combinatoric computation of marginal contributions, SHAP shielded local interpretations, especially when coupled with the linear SVM classifier, maintaining high consistency indices and stable interquartile dispersions.

On the other hand, LIME and visualizations based on attention mechanisms manifested severe structural vulnerabilities under resampling regimes. LIME exhibited chronic local instability when attempting to map the stochastic neighborhood of highly non-linear decision boundaries in the LSTM when the latter was fitted via *undersampling*. However, the most alarming phenomenon manifested itself in the *oversampling* configuration applied to the LSTM under attention mechanisms. The strict replication of textual sequences induced the deep network to converge toward redundant optimization valleys, causing the attention probability distributions to oscillate erratically between competing tokens, culminating in severe collapses of mathematical consistency in the textual anomaly detection bases.

As a methodological implication, this study evidences that the engineering of transparent AI systems cannot evaluate predictive performance in isolation from explainability metrics. Naive interventions in data density tend to mask interpretative biases and compromise the auditability of critical domains, such as financial sentiment analysis and disinformation monitoring.

As future work, we plan to evaluate whether predictive model performance (e.g., accuracy, F1-score) directly modulates explanation stability, specifically investigating if misclassifications or low predictive confidence drive the observed representation collapses. Additionally, we intend to expand this evaluation pipeline to Transformer-based architectures (such as BERT variants) to analyze whether multi-head attention mechanisms provide greater robustness compared to recurrent networks. Finally, we aim to explore text-specific synthetic sampling techniques, such as synonym substitution and back-translation, to determine if artificial semantic variability can reconcile predictive accuracy gains with post-hoc explanatory fidelity.

To access the code and dataset used in this study, visit: `https://github.com/LuirysN/SVM-x-LSTM-com-dados-desbalanceados.git`

References

ARAÚJO, J. P. B. d. (2020). Interpretabilidade de modelos de machine learning: aplicação no mercado de crédito. UFC.

BADJATIYA, P. e. a. (2017). Deep learning for hate speech detection in tweets. In *INTERNATIONAL WORLD WIDE WEB CONFERENCES STEERING COMMITTEE*, page 759–760. Proceedings of the 26th International Conference on World Wide Web Companion.

CHANDRA, Mayank Arya; BEDI, S. S. (2021). Survey on svm and their application in image classification. In *International Journal of Information Technology*, pages 1–11.

Chen, X. et al. (2023). Interpretable machine learning for imbalanced credit scoring datasets. *European Journal of Operational Research*, 308(1):356–370.

CORTES, Corinna; VAPNIK, V. (1995). Support-vector networks. In *Machine learning*, pages 273–297.

DABLAIN, D. e. a. (2024). Understanding imbalanced data: Xai& interpretable ml framework. In *Machine learning*, pages 3751–3769. DBLP.

Fujiwara, T. (2024). Knowledge distillation with resampling for imbalanced data classification: Enhancing predictive performance and explainability stability. *Results in Engineering*, 24:103406.

Hochreiter, S. (1997). Long short-term memory. *Neural Computation MIT-Press*.

INDRAHIMAWAN, M. R. e. a. (2023). Handling data imbalance using text augmentation for classifying public complaints. In *2023 International Conference on Computer, Control, Informatics and its Applications (IC3INA)*, pages 284–289. IEEE.

MENDES, T. M., Ferreira, D. D., SILVA, L. R. M., and DUQUE, C. A. (2019). Aplicação de subamostragem para análise de distorções supraharmônicas. In *Congresso Brasileiro de Automática-CBA*, volume 1.

MINH, D. e. a. (2022). Explainable artificial intelligence: a comprehensive review. In *Artificial Intelligence Review*, pages 1–66. SPRINGER NATURE.

MOHAMMED, R. e. a. (2020). Machine learning with oversampling and undersampling techniques: overview study and experimental results. In *2020 11th international conference on information and communication systems (ICICS)*, pages 243–248. IEEE.

PERES, Vitor, e. a. (2019). Análises de sentimentos: abordagem lexical de classificação de opinião no contexto mercado financeiro brasileiro. In *Workshop of Artificial Intelligence Applied to Finance*. ITA.

POCEBON, M. (2023). Ética na ia: Desafios e considerações éticas em aplicativos de ia e machine learning. In *Jornada Acadêmica*, pages 2–2. FAQI.

Rai, A. et al. (2025). Xai techniques under class imbalance. *Frontiers in Artificial Intelligence*, 8:1682919.

Roth, A. E. (1988). *The Shapley value: essays in honor of Lloyd S. Shapley*. Cambridge University Press.

SILVA, A. B. e. a. (2025). Evaluating how different balancing data techniques impact on prediction of premature birth using machine learning models. In *PloS one*, page e0316574.

VILONE, Giulia; LONGO, L. (2021). Notions of explainability and evaluation approaches for explainable artificial intelligence. In *Information Fusion*, pages 89–106. ELSEVIER.

WANG, S. e. a. (2018). Densely connected cnn with multi-scale feature attention for text classification. In *IJCAI*, pages 4468–4474. COAI.

Zafar, M. R. and Khan, N. (2021). Deterministic local interpretable model-agnostic explanations for stable explainability. *Machine Learning and Knowledge Extraction*, 3(3):525–541.