Techniques for Dealing with Imbalanced Data: A Systematic Literature Review

Leandro O. da Silva[1], **Daniela L. Freire**[2],
Márcio P. Basgalupp[1], **André C.P.L.F. de Carvalho**[2]

[1]Instituto de Ciência e Tecnologia
Universidade Federal de São Paulo
São José dos Campos – SP – Brasil

[2]Instituto de Ciências de Computação e Matemática Computacional
Universidade de São Paulo
São Carlos – SP – Brasil

{oliveira.leandro,basgalupp}@unifesp.br

{danielalfrere,andre}@icmc.usp.br

Abstract. *This systematic review of the literature addresses techniques employed to address the problem of data imbalance. A variety of articles were analyzed, exploring strategies such as under-sampling, oversampling, and their combinations to address asymmetry in class distributions. Sensitive metrics, including recall, precision, and F1 score, emerge as crucial in imbalanced contexts. The studies reveal the challenges in selecting appropriate strategies and underscore the importance of adaptive approaches. Innovative solutions, such as adaptive combinations of techniques and integration with specific algorithms, are discussed. The ongoing need for research to address the specific challenges of data imbalance is highlighted.*

1. Introduction

Data imbalance is a common problem in machine learning, occurring when one class of data is significantly more frequent than the others. This issue can negatively affect the performance of machine learning algorithms, as they may learn to easily recognize the majority class while ignoring or underestimating the minority class.

Data imbalance can hinder the performance of machine learning algorithms in several ways. First, it can lead to an increase in false-positive errors. For example, if a machine learning algorithm is trained to detect fraud in financial transactions, and the fraud class is much less frequent than the legitimate transactions class, the algorithm may be more likely to classify a legitimate transaction as fraud.

Second, data imbalance can lead to an increase in false negative errors. For instance, if a machine learning algorithm is being trained to diagnose cancer, and the class of cancer patients is much less frequent than the class of non-cancer patients, the algorithm may be more likely to fail to diagnose a patient with cancer.

Several techniques can be used to address data imbalance. The two main approaches are:

- **Artificial data balancing:** This technique manipulates the dataset to balance the classes. There are various artificial data balancing methods, such as oversampling, undersampling, and SMOTE.
- **Algorithm modification:** This technique involves adjusting machine learning algorithms to make them more tolerant to data imbalance.

Data imbalance is a common issue in machine learning that can compromise algorithm performance. Several techniques can be applied to address this problem, including artificial data balancing and algorithm modification. The choice of the most suitable technique depends on the specific problem being addressed.

2. Methodology

This research followed a methodology known as Systematic Literature Review (SLR). According to Kitchenham and Charters 2007, an SLR "is a means of identifying, evaluating, and interpreting all available research relevant to a particular research question, topic area, or phenomenon of interest".

There are three main phases in an SLR: planning, conducting, and reporting [Kitchenham and Charters 2007]. The first phase is related to the pre-review stage. Its purpose is to address the need for the SLR by defining the basic and essential procedures of the review. In the conducting phase, the selection of studies is carried out based on the criteria established in the previous phase. Data extraction and synthesis are performed during this stage to answer the research questions. The final phase involves reporting the review results, which are presented according to the objectives of the SLR.

2.1. Planning

In this stage, the objectives of the Systematic Literature Review (SLR), the research questions, search strategies, repositories, and the inclusion and exclusion criteria for the articles are defined.

The objective of this research is to investigate the most commonly used techniques for addressing data imbalance and to compare the different methods applied and the results obtained. According to the goals of this study, the following research questions were defined:

- ✓ RQ1: Which imbalance techniques are most effective in improving model performance in text classification tasks?
- ✓ RQ2: How do imbalance techniques affect the generalization ability of models across different application domains?
- ✓ RQ3: Which performance evaluation metrics are most sensitive to applying imbalance techniques?
- ✓ RQ4: Which studies identify the challenges and limitations of imbalance techniques and propose solutions or improvements?

The searches were conducted between November and December 2023. The selected repositories were: ACM Digital Library, Google Scholar, IEEE Xplore, and ScienceDirect. The search string used to find results in the abstracts was:

(“imbalanced data” OR “data imbalanced” OR “class imbalanced” OR “imbalanced datasets”) AND (“rebalancing techniques” OR “performance evaluation in imbalanced data” OR “strategies for handling imbalanced data” OR “methods for handling imbalanced data”)

The inclusion and exclusion criteria considered for evaluating whether the articles are aligned with the objectives of the Systematic Literature Review and the research questions are described as follows.

Inclusion Criteria:

1. Studies that apply Natural Language Processing;
2. Research that compares different balancing techniques;
3. Studies that propose new approaches to data balancing.

Exclusion Criteria:

1. Studies not related to Natural Language Processing;
2. Studies already selected from another database;
3. Previous versions of selected studies;
4. Studies not written in English or Portuguese.

2.2. Conducting the Study

After defining the research plan with objectives and methodology, the selection of articles was carried out using the specified repositories and a search string covering key terms.

Predefined inclusion and exclusion criteria ensured alignment with the research objectives and filtered out irrelevant studies.

The selection occurred in two stages: a first screening based on titles and abstracts, followed by a detailed review of introductions and conclusions of the shortlisted articles. The criteria were consistently applied in both stages to guarantee the relevance and quality of the final set.

Table 1 presents the number of articles selected at each research stage, for each of the repositories considered. In the first stage, selection was based on search results using titles and abstracts. In the second stage, the articles that passed the first selection were analyzed more thoroughly, focusing on the main aspects of the studies, and the inclusion and exclusion criteria were applied again to ensure that only the most relevant articles were retained.

Table 1. Number of Articles Identified, Screened, and Included in the Review Process.

Repository	Search String	1st Selection	2nd Selection
ACM Digital Library	39	8	6
Google Scholar	88	8	6
IEEE Xplore	43	6	4
ScienceDirect	40	6	4
Total	210	28	20

By applying the defined search string in the selected repositories, 210 articles were initially retrieved. From this initial set of articles, the selection process was conducted in two stages to refine the search and ensure that only the most relevant studies for the research were retained.

In the first stage, articles were selected based on a preliminary reading of their titles and abstracts. This initial filtering helped reduce the number of articles by eliminating those that were not directly related to the topic of interest or did not meet the basic criteria for inclusion. After this analysis, 28 articles were selected for a more in-depth review.

In the second stage, the articles selected in the first phase were analyzed more thoroughly, emphasizing the details presented in the introductions and conclusions of each study. This process involved a critical evaluation of the objectives, methodologies, and results presented to ensure that the articles were fully aligned with the research questions and the previously defined inclusion criteria. As a result of this more detailed analysis, 20 articles were finally selected to form the final dataset of this Systematic Literature Review. Table 2 shows the studies selected for analysis and discussion in this systematic review of the literature.

Table 2. Distribution of Selected Articles by Digital Library or Repository.

Repository	Articles
ACM Library	[Yang et al. 2023a] [Zheng 2023] [Tashkandi and Wiese 2020] [Sowah et al. 2021] [Ren et al. 2023] [Moniruzzaman et al. 2020]
Google Scholar	[Maldonado et al. 2019] [Jonathan et al. 2020] [Rathpisey and Adji 2019] [Rupapara et al. 2021] [Pal and Patel 2020] [Verdikha et al. 2023]
IEEE	[Nhita et al. 2023] [Wang et al. 2022] [Suhana and Kumar 2022] [Kiran and Kumar 2023]
Science Direct	[Guo et al. 2023] [Xue et al. 2023] [Sun et al. 2023] [Yang et al. 2023b]

3. Results and Discussion

This section will address and answer the previously established research questions based on the studies selected and analyzed during the Systematic Literature Review.

Which balancing techniques are most effective in improving model performance in text classification tasks? After analyzing the selected articles, it was possible to categorize them into distinct groups based on the techniques addressed in each study. This classification facilitates the understanding of the approaches used in each work, providing a clearer view of the specific techniques employed in each research context. The techniques that achieved the best performance in the analyzed studies were selected.

Based on the results, it is possible to observe the type of technique used in each study and verify that these were the ones that achieved the best performance, meaning they were the most effective techniques. It can also be seen that the most frequently used techniques were the traditional ones and the advanced or proposed techniques. Traditional

Table 3. Overview of Traditional Techniques Used for Data Imbalance Handling.

Article	Techniques
[Yang et al. 2023a]	Undersampling
[Maldonado et al. 2019]	SMOTE
[Rathpisey and Adji 2019]	Oversampling and Logistic Regression
[Pal and Patel 2020]	SVM, Naive Bayes, and Random Forest
[Verdikha et al. 2023]	Support Vector Regression and SMOTE
[Nhita et al. 2023]	RandomUnderSampler and RandomOverSampler
[Wang et al. 2022]	Reference-point based k Neighbors algorithm
[Suhana and Kumar 2022]	Decision Tree

Table 4. Use of Ensemble-Based Approaches for Addressing Class Imbalance.

Article	Techniques
[Zheng 2023]	Random Forest Classifier and Multi-Layer Perceptron
[Tashkandi and Wiese 2020]	Gradient Boosting Decision Tree
[Rupapara et al. 2021]	Regression Vector Voting Classifier

techniques have a history of efficiency in various applications across different contexts. In contrast, the proposed techniques were adjusted according to the specific problems addressed in each study to achieve better results.

How do imbalance techniques affect the generalization ability of models across different application domains? Imbalance handling techniques have a significant impact on the generalization ability of models in diverse application domains. The reviewed studies explore strategies such as undersampling, oversampling, and hybrid approaches, highlighting the challenges faced when dealing with asymmetric class distributions. For example, [Sowah et al. 2021] reports that combining undersampling and oversampling can mitigate imbalance and improve generalization. Similarly, [Maldonado et al. 2019] and [Jonathan et al. 2020] investigate SMOTE and its combination with the Tomek method, showing benefits from generating synthetic samples and removing noisy instances.

A more diversified approach is proposed by [Nhita et al. 2023], who combines RandomUnderSampler, InstanceHardnessThreshold, and RandomOverSampler to address imbalance from multiple perspectives. However, these strategies may also introduce new challenges, particularly in complex environments. In such cases, adaptive methods and ensemble techniques, as suggested by [Xue et al. 2023], can help enhance generalization across varied domains.

Table 5. Sampling Strategies Applied to Deal with Data Imbalance.

Article	Techniques
[Sowah et al. 2021]	Combination of Undersampling and Oversampling
[Moniruzzaman et al. 2020]	Undersampling
[Maldonado et al. 2019]	SMOTE
[Jonathan et al. 2020]	G-Mean Combination SMOTE-Tomek with SVM
[Rathpisey and Adji 2019]	Oversampling, Logistic Regression
[Nhita et al. 2023]	RandomUnderSampler, InstanceHardnessThreshold, RandomOverSampler

Table 6. Advanced and Proposed Techniques Developed to Handle Data Imbalance.

Article	Techniques
[Yang et al. 2023a]	Optimal G-Mean
[Ren et al. 2023]	Slack-Factor-Based Fuzzy Support Vector Machine
[Jonathan et al. 2020]	G-Mean combination SMOTE-Tomek with SVM
[Kiran and Kumar 2023]	Generative Adversarial Network
[Guo et al. 2023]	Adaptive SV-Borderline SMOTE-SVM
[Xue et al. 2023]	Multi-feature fusion and convolutional neural network
[Sun et al. 2023]	Adaptive fuzzy multi-neighborhood feature selection with hybrid sampling
[Yang et al. 2023a]	Multi-view feature fusion and SMOTE

Overall, the literature indicates that selecting and combining imbalance techniques appropriately is crucial to improving model robustness and performance in real-world scenarios.

Which performance evaluation metrics are most sensitive to the application of imbalance techniques? Applying imbalance techniques in machine learning models can influence various performance evaluation metrics, with some being more sensitive to changes in class distribution. Notably, metrics such as Recall, Precision, and F1-Score are particularly sensitive in imbalanced contexts.

Recall is crucial because it measures the model's ability to correctly identify instances of the minority class, being especially relevant when the goal is to detect rare events or patterns. Precision, in turn, highlights the proportion of correctly identified positive instances in relation to all instances predicted as positive, making it sensitive to changes in the number of false positives, which may occur in imbalanced scenarios.

F1-Score, which combines Recall and Precision, provides a balanced metric that accounts for both false-positives and false-negatives, making it sensitive to performance across both classes. On the other hand, metrics such as Accuracy may be less sensitive in imbalanced scenarios, as the dominance of the majority class can skew them.

Therefore, when evaluating the performance of models under imbalanced conditions, it is essential to pay special attention to metrics that capture the model's ability to effectively handle minority class instances, offering a more accurate and comprehensive performance assessment.

Which studies identify the challenges and limitations of imbalance techniques and propose solutions or improvements? Several studies address the challenges and limitations associated with imbalance techniques in machine learning models, particularly in the context of Natural Language Processing (NLP). Some of these studies not only identify key obstacles but also propose solutions or improvements to tackle these issues. Among them, [Sowah et al. 2021] explores the combination of undersampling and oversampling, highlighting the difficulty in selecting appropriate strategies and sampling ratios.

Additionally, [Xue et al. 2023] acknowledges the complexity of integrating imbalance techniques into neural network architectures and suggests specific adaptations to address these challenges. [Nhita et al. 2023] points out the limitations of individual approaches and proposes combining multiple techniques to mitigate such constraints.

Understanding these challenges and the proposed solutions is essential for advancing the effectiveness of imbalance handling methods. The literature continues to evolve, reflecting an ongoing need to address both practical and theoretical issues related to class imbalance in NLP tasks and machine learning more broadly.

4. Conclusion

This Systematic Literature Review provides an overview of the challenges and strategies associated with the class imbalance problem. The analysis of the selected studies reveals a variety of approaches, including undersampling, oversampling, combinations of these techniques, and integration with specific algorithms such as Random Forest and Neural Networks.

The reviewed studies consistently highlight the sensitivity of evaluation metrics such as Recall, Precision, and F1-Score in imbalanced scenarios. These metrics emerge as critical model performance indicators, particularly when addressing the minority class.

While some proposals, such as the adaptive combination of techniques, show promise in overcoming specific challenges, the studies also emphasize the ongoing need for innovative approaches to deal with the complexity of imbalanced data. The integration of ensemble methods, the consideration of multiple sampling strategies, and the adaptation to specific environments are promising areas for future research.

This review underscores that improving the effectiveness of imbalance techniques requires a deep understanding of the specific challenges involved and reinforces the ongoing importance of research in promoting innovative and practical solutions. A

multidimensional approach to these challenges is crucial for advancing the development of robust and generalizable models in imbalanced data scenarios.

Acknowledgment

This work is supported by *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior* (CAPES), *Conselho Nacional de Desenvolvimento Científico e Tecnológico* (CNPq), grants 406417/2022-9, 102475/2024-5, and 312209/2022-3, and *Fundação de Amparo à Pesquisa do Estado de São Paulo* (FAPESP), grant 2020/09835-1 (CPA IARA).

References

[Guo et al. 2023] Guo, J., Wu, H., Chen, X., and Lin, W. (2023). Adaptive sv-borderline smote-svm algorithm for imbalanced data classification. *SSRN Electronic Journal*.

[Jonathan et al. 2020] Jonathan, B., Putra, P. O. H., and Ruldeviyani, Y. (2020). Observation imbalanced data text to predict users selling products on female daily with smote, tomek, and smote-tomek. pages 81–85.

[Kiran and Kumar 2023] Kiran, A. and Kumar, S. S. (2023). A comparative analysis of gan and vae based synthetic data generators for high dimensional, imbalanced tabular data. In *2023 2nd International Conference for Innovation in Technology (INOCON)*, pages 1–6.

[Kitchenham and Charters 2007] Kitchenham, B. and Charters, S. (2007). Guidelines for performing systematic literature reviews in software engineering. Technical report, EBSE Technical Report, Keele University and University of Durham.

[Maldonado et al. 2019] Maldonado, S., López, J., and Vairetti, C. (2019). An alternative smote oversampling strategy for high-dimensional datasets. *Applied Soft Computing*, 76:380–389.

[Moniruzzaman et al. 2020] Moniruzzaman, M., Bagirov, A., and Gondal, I. (2020). Partial undersampling of imbalanced data for cyber threats detection. In *Proceedings of the Australasian Computer Science Week Multiconference*, ACSW '20, New York, NY, USA. Association for Computing Machinery.

[Nhita et al. 2023] Nhita, F., Adiwijaya, K., and Kurniawan, I. (2023). Performance and statistical evaluation of three sampling approaches in handling binary imbalanced data sets. pages 420–425.

[Pal and Patel 2020] Pal, K. and Patel, B. V. (2020). Data classification with k-fold cross validation and holdout accuracy estimation methods with 5 different machine learning techniques. In *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, pages 83–87.

[Rathpisey and Adji 2019] Rathpisey, H. and Adji, T. B. (2019). Handling imbalance issue in hate speech classification using sampling-based methods. In *2019 5th International Conference on Science in Information Technology (ICSITech)*, pages 193–198.

[Ren et al. 2023] Ren, J., Wang, Y., and Deng, X. (2023). Slack-factor-based fuzzy support vector machine for class imbalance problems. *ACM Trans. Knowl. Discov. Data*, 17(6).

[Rupapara et al. 2021] Rupapara, V., Rustam, F., Shahzad, H. F., Mehmood, A., Ashraf, I., and Choi, G. S. (2021). Impact of smote on imbalanced text features for toxic comments classification using rvvc model. *IEEE Access*, 9:78621–78634.

[Sowah et al. 2021] Sowah, R. A., Kuditchar, B., Mills, G. A., Acakpovi, A., Twum, R. A., Buah, G., and Agboyi, R. (2021). Hcbst: An efficient hybrid sampling technique for class imbalance problems. *ACM Trans. Knowl. Discov. Data*, 16(3).

[Suhana and Kumar 2022] Suhana, S. S. and Kumar, S. A. (2022). An novel adaptive solution in machine learning approaches for mining serendipitous drug usage to handle imbalanced data from social media comparing with adaboost algorithm. In *2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS)*, pages 311–314.

[Sun et al. 2023] Sun, L., Li, M., Ding, W., and Xu, J. (2023). Adaptive fuzzy multi-neighborhood feature selection with hybrid sampling and its application for class-imbalanced data. *Applied Soft Computing*, 149:110968.

[Tashkandi and Wiese 2020] Tashkandi, A. and Wiese, L. (2020). A hybrid machine learning approach for improving mortality risk prediction on imbalanced data. In *Proceedings of the 21st International Conference on Information Integration and Web-Based Applications & Services*, iiWAS2019, page 83–92, New York, NY, USA. Association for Computing Machinery.

[Verdikha et al. 2023] Verdikha, N. A., Thamrin, H., Triyono, A., Abdillah, M. F., and Suryawan, S. H. (2023). Regression and oversampling method for indonesian language automated essay scoring. *AIP Conference Proceedings*, 2727(1):040020.

[Wang et al. 2022] Wang, J., Wu, Y., Qi, J., and Chen, Z. (2022). An efficient reference-point based k neighbors algorithm for imbalanced data. In *2022 7th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA)*, pages 513–517.

[Xue et al. 2023] Xue, L., Wu, H., Zheng, H., and He, Z. (2023). Control chart pattern recognition for imbalanced data based on multi-feature fusion using convolutional neural network. *Comput. Ind. Eng.*, 182(C).

[Yang et al. 2023a] Yang, C., Dong, Y., Lu, J., and Peng, Z. (2023a). Solving imbalanced data in credit risk prediction: A comparison of resampling strategies for different machine learning classification algorithms, taking threshold tuning into account. In *Proceedings of the 2022 5th International Conference on Machine Learning and Machine Intelligence*, MLMI '22, page 30–40, New York, NY, USA. Association for Computing Machinery.

[Yang et al. 2023b] Yang, R., Liu, J., Zhang, Q., and Zhang, L. (2023b). Multi-view feature fusion and density-based minority over-sampling technique for amyloid protein prediction under imbalanced data. *Applied Soft Computing*, 150:111100.

[Zheng 2023] Zheng, K. (2023). Identifying churning employees: Machine learning algorithms from an unbalanced data perspective. In *Proceedings of the 2022 5th International Conference on Machine Learning and Machine Intelligence*, MLMI '22, page 14–22, New York, NY, USA. Association for Computing Machinery.