Estudo do Impacto de Dados Sintéticos e Paráfrases na Mitigação do Desbalanceamento em Tarefas de Classificação de Textos em Português com Baixa Amostragem

Claudio M. V. de Andrade[1], **Gestefane Rabbi Magalhães**[1], **Raiane Asevedo**[1]
Julia Paes[1], **Isaias José Ramos Oliveira**[1], **Adriana Pagano**[1]
Zilma Reis[1], **Marcos A. Gonçalves**[1]

[1]Unversidade Federal de Minas Gerais (UFMG)

{claudio.valiense,gestefane,juliapaes,mgoncalv}@dcc.ufmg.br

{apagano,raiasevedo,zilma}@ufmg.br, isaias@medicina.ufmg.br

Resumo. *O desbalanceamento de classes é um desafio relevante na classificação automática de textos, especialmente em contextos de dados anotados escassos e línguas ainda sub-representadas, como é caso do português. Este estudo investiga a classificação de um conjunto de dados escassos desbalanceado constituído por solicitações de suporte técnico registradas por profissionais de saúde relativas aos sistemas do e-SUS APS. Foram avaliadas seis estratégias de reamostragem — duas de subamostragem e quatro de sobreamostragem, incluindo geração de paráfrases com Large Language Models. A combinação de sobreamostragem via paráfrases com rotulação seletiva elevou a Macro-F1 do BERTimbau em 18%, alcançando desempenho estatisticamente equivalente ao da Regressão Logística (RL) aplicada à junção de dados originais, random oversampling e rotulação seletiva, que atingiu 70% de melhorias em relação ao método original. A RL é contudo cerca de 3690x mais eficiente que o BERTimbau considerando a versão mais efetiva de ambos os métodos.*

Abstract. *Class imbalance poses a major challenge in text classification, particularly in cenarios of scarce annotated data and low-resource languages like Portuguese. This study addresses the classification of technical support requests from healthcare professionals on the use of the Electronic Health Record in Primary Health Care of the Brazilian Unified Health System (e-SUS APS), based on a small and highly imbalanced dataset. Six resampling strategies were evaluated, including paraphrase-based data augmentation using Large Language Models. Combining paraphrasing with selective labeling improved BERTimbau's Macro-F1 by 18%, matching the performance of Logistic Regression (LR) with random oversampling and selective labeling, which yielded a 70 improvement over the original method. However, LR is approximately 3690 times more efficient than BERTimbau when considering the most effective version of each method.*

1. Introdução

Este estudo aborda a classificação automática de solicitações de suporte técnico, conhecidas no jargão como "tickets", geradas a partir de dúvidas de usuários dos softwares do e-SUS Atenção Primária a Saúde (APS)[1]. Essa classificação pode ajudar o atendente do

[1]https://sisaps.saude.gov.br/sistemas/esusaps/

eSus a melhor responder ao *ticket*, indicando as melhores fontes para obter as respostas, além de auxiliar na construção de sistemas automáticos de perguntas e repostas sendo desenvolvidos. Cada solicitação contém campos textuais de assunto e descrição e deve ser classificada com uma entre 11 categorias predefinidas, as quais permitem direcionar a consulta para um tópico relevante (ex.: relatórios ou cadastro de pacientes). A distribuição de classes nos dados rotulados sendo trabalhados, que reflete a distribuição real dos *tickets*, é desigual, dificultando a classificação. Por exemplo, há apenas 33 solicitações na categoria vacina, enquanto a categoria manutenção de software e versões possui 155.

O desbalanceamento induz modelos de aprendizado a favorecer classes majoritárias, dificultando a identificação de categorias minoritárias — especialmente em corpora em português com poucas instâncias rotuladas. Essa combinação torna a classificação automática de solicitações uma tarefa desafiadora. Técnicas de reamostragem, divididas em sub e sobre-amostragem, podem mitigar esse problema. A sub-amostragem reduz exemplos da classe majoritária para equilibrar a distribuição, mas pode acarretar perda de informação, o que é particularmente crítico em conjuntos pequenos, como o do e-SUS APS (Tabela 1, coluna Original). Para minimizar esse impacto, métodos como *NearMiss* [He & Garcia 2009], *Neighborhood Cleaning Rule* [Tabar et al. 2018], E2SC [Cunha et al. 2023] e bioIS [Cunha et al. 2025] propõem remoções seletivas que preservam a representatividade dos dados e promovem modelos mais robustos.

A sobre-amostragem busca aumentar a representação de classes minoritárias por replicação ou geração de exemplos sintéticos. O SMOTE [Chawla et al. 2002] é o método mais difundido, gerando amostras por interpolação entre instâncias e seus vizinhos. Extensões como o *Borderline-SMOTE* [Han et al. 2005] e o *SVM-SMOTE* [Nguyen et al. 2011] tornam as amostras mais informativas ao focar em regiões próximas às fronteiras de decisão. Apesar de eficazes, tais técnicas podem induzir *overfitting* se as amostras não refletirem bem a distribuição dos dados. Recentemente, estratégias baseadas em LLMs [Vaswani et al. 2017] têm sido exploradas para *oversampling* textual por meio da geração de paráfrases [Yadav et al. 2024]. Essa abordagem é especialmente útil em contextos de baixa amostragem, onde a rotulagem especializada é limitada. Neste trabalho, empregamos *LLMs* para ampliar as classes sub-representadas de um conjunto de solicitações de suporte técnico registradas por profissionais de saúde relativas aos sistemas do e-SUS APS, avaliando seu impacto na classificação.

O desbalanceamento torna-se mais crítico em contextos com poucos dados rotulados e textos em idiomas distintos do inglês, devido ao viés de treinamento dos modelos, que tendem a apresentar melhor desempenho em inglês [Hu et al. 2020]. Este trabalho busca avaliar o impacto de técnicas de reamostragem (sub e sobre-amostragem) em cenários de alto desbalanceamento e escassez de dados rotulados em português - uma realidade frequente em aplicações no Brasil.

Neste estudo, avaliamos dois métodos de *undersampling*, que removem documentos da classe majoritária, e quatro técnicas de *oversampling*, baseadas na replicação ou geração sintética de exemplos minoritários. Essas abordagens foram aplicadas a dois modelos de classificação textual: regressão logística e *BERTimbau* - modelo com *embeddings* pré-treinados baseado em BERT, adaptado ao português [Souza et al. 2020]. Adicionalmente, incorporamos dados anotados via *amostragem seletiva inversamente proporcional à distribuição de classes* e avaliamos seu impacto no desempenho. Em

síntese, este trabalho busca responder as seguintes perguntas de pesquisa:

- **RQ1:** Qual é o impacto da aplicação de técnicas de reamostragem de dados para a tarefa de classificação de textos em cenários com escassez de dados e em língua portuguesa?
- **RQ2:** A anotação adicional de dados por meio de *amostragem seletiva inversamente proporcional à distribuição de classes* pode ajudar a mitigar o desbalanceamento e quais são suas implicações no desempenho da classificação?

Na RQ1, observamos desempenho superior das abordagens de *oversampling* em relação ao *undersampling*. Na Regressão Logística (RL), os ganhos em *macro-F1* com *SMOTE* ou paráfrases alcançaram 57%; no BERTimbau, 16% com paráfrases. Na RQ2, aplicamos amostragem baseada na confiança do modelo para selecionar exemplos de classes minoritárias para anotação. A inclusão desses dados reduziu o desbalanceamento e ampliou os ganhos para 70% com RL (Original vs. Original + Novos Dados + *Random Oversample*, Tabela 2) e 18% com BERTimbau — ambos atingindo seus melhores resultados experimentais. Em termos de eficiência, porém, a RL é aproximadamente 3690 vezes mais eficiente que o BERTimbau, quando comparando as suas respectivas versões com a maior efetividade.

2. Trabalhos Relacionados

Técnicas de *undersampling* buscam mitigar a dominância da classe majoritária preservando a representatividade dos dados. O *NearMiss*[He & Garcia 2009] seleciona amostras majoritárias próximas às minoritárias; o *Tomek Links*[Batista et al. 2004] remove pares ambíguos entre classes opostas; e o *Neighborhood Cleaning Rule* (NCL)[Tabar et al. 2018] combina *undersampling* com aprendizado por vizinhança para excluir amostras rodeadas por outras classes. No *oversampling*, o *SMOTE* (Synthetic Minority Over-sampling Technique) gera amostras sintéticas por interpolação entre instâncias minoritárias e seus vizinhos[Chawla et al. 2002], mas pode gerar ruído em regiões de sobreposição. Para contornar esse problema, o *KMeans-SMOTE* [Last et al. 2017] agrupa instâncias minoritárias antes da interpolação, respeitando melhor a estrutura dos dados, enquanto o *SVM-SMOTE* [Nguyen et al. 2011] utiliza margens de SVM para gerar exemplos em regiões decisivas. Essas variantes visam preservar a distribuição e reduzir ruídos. Modelos generativos também oferecem alternativas promissoras para enriquecimento de dados. O PAG-LLM [Yadav et al. 2024] emprega LLMs para gerar paráfrases de intenções, ampliando o treino e reduzindo erros em cenários com poucos dados. Adaptamos essa técnica para parafrasear documentos de classes minoritárias e reequilibrar o conjunto. Em paralelo, [McClure et al. 2024] utiliza otimização de prompt e in-context learning para lidar com o desbalanceamento, enquanto [Taskiran et al. 2025] aplica Transformers combinados a técnicas de oversampling como o SMOTE. Diferentemente desses trabalhos, nossa proposta emprega LLMs para geração de paráfrases e contempla a língua portuguesa, ampliando sua aplicabilidade em cenários com poucos dados.

3. Proposta

Apresentamos a proposta principal do artigo, focada na classificação textual de um conjunto pequeno e desbalanceado de documentos em português. A Figura 1 ilustra o fluxo: os dados são divididos em treino, validação e teste por validação cruzada estratificada com 5 dobras, preservando a distribuição original. As técnicas de reamostragem são aplicadas apenas ao conjunto de treinamento, garantindo a integridade da distribuição no teste.

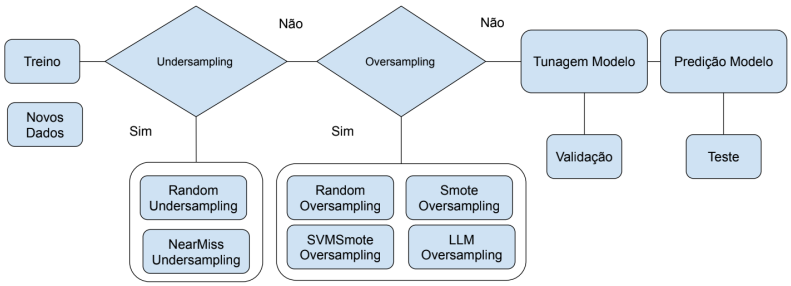

Figura 1. Fluxograma da proposta.

No conjunto de treinamento, a reamostragem é feita por *undersampling* ou *oversampling*. Para *undersampling*, avaliamos o *Random Undersample*, que remove aleatoriamente instâncias das classes majoritárias até que haja balanceamento com a classe minoritária. Também testamos o *NearMiss*, que elimina instâncias majoritárias menos similares à classe minoritária, buscando balancear as classes.

Para *oversampling*, avaliamos quatro técnicas que equilibram as classes minoritárias em relação à majoritária: (1) *Random Oversampling*, com replicação aleatória; (2) *SMOTE*, com interpolação de amostras; (3) *SVMSMOTE*, centrado na fronteira de decisão; e (4) geração de paráfrases via LLM (LLaMA 3.1). Após aplicar *under-* ou *oversampling* nos dados de treinamento, realizamos ajuste de parâmetros por validação e avaliamos no teste. Utilizamos dois classificadores: regressão logística (superior a *Random Forest* e *XGBoost* em testes preliminares) e BERTimbau, modelo baseado em BERT adaptado ao português. Os textos foram representados por TF-IDF na regressão logística; no BERTimbau, tal transformação não é necessária. Os resultados apresentados referem-se ao desempenho desses modelos no conjunto de teste.

Além do conjunto inicial, novos dados foram anotados estrategicamente para mitigar o desbalanceamento, considerando a distribuição das classes e seu impacto no treinamento. Coletamos cerca de 1300 instâncias não rotuladas e aplicamos um modelo de Regressão Logística com oversample dos dados originais para prever automaticamente as classes desses dados sem rótulo. A escolha da RL deve-se à sua maior *calibragem* em relação a modelos Transformers como o BERTimbau, proporcionando maior *confiabilidade* nas probabilidades preditas. Selecionamos 400 documentos para anotação manual, priorizando aqueles associados a classes minoritárias, com base na inversa da distribuição original e nas probabilidades do classificador. Quanto mais rara a classe prevista, maior a chance de seleção. Cada documento foi avaliado por 3 especialistas, que escolheram uma entre 11 categorias. Rotularam-se 349 documentos (87%) com concordância mínima entre dois avaliadores, denominados “Novos Dados”, cuja inclusão no treinamento foi analisada.

4. Metodologia experimental

Conjunto de Dados A Tabela 1 mostra a distribuição solicitações de suporte técnico (tickets) do dataset APS nas 11 categorias - a coluna “Original” representa a base original e “Novos Dados” os documentos rotulados pelo processo de anotação descrito na Seção 3. Observa-se forte desbalanceamento na base original, enquanto a nova rotulação aumentou a representatividade das classes minoritárias, como, por exemplo, a classe “Vacina”, devido à seleção baseada na inversa da distribuição original.

Tabela 1. Estatística do E-SUS APS Dataset

Classe	Original	Novos Dados
Agente Comunitário de Saúde (ACS) e de Combate às Endemias (ACE)	66	29
Aplicativos e-SUS	36	51
Cadastro de Pessoas (Cidadão)	43	40
Cadastro de Profissional	31	38
Demais Sistemas Relacionados à APS	128	10
Gestão de Recursos, Serviços ou Relatórios	57	32
Instalação e Implantação do Prontuário Eletrônico e-SUS APS	91	24
Manutenção do Software e Versões	155	20
Preenchimento da Coleta de Dados Simplificada (CDS)	33	37
Processo de Atendimento	107	17
Vacina	33	51

Prompt para Paráfrases O Quadro 1 exibe o prompt usado para gerar paráfrases, consistindo de uma instrução seguida do texto original a ser parafraseado. As paráfrases geradas são incorporadas ao conjunto de treinamento para mitigar a escassez de dados nas classes minoritárias.

Prompt

Parafraseie o texto a seguir usando palavras e estruturas de frase diferentes, mantendo o significado original:
Texto: Atualizamos a versão do e-SUS para a 4.39, porém ao finalizar o atendimento ele aparece o erro de funcionalidade e não deixa salvar o atendimento. Anexo segue foto.
Paráfrase: [atualizamos a versão do e-sus para a 4.39, mas ao finalizar o atendimento, o erro de funcionalidade aparece e não permite salvar o atendimento. segue a foto.]

Quadro 1. Modelo de prompt submetido ao LLaMA 3.1 e exemplo de resultado.

5. Resultados Experimentais

Os resultados foram obtidos via validação cruzada estratificada com 5 dobras, preservando a distribuição original das classes. Em cada iteração, uma partição é usada para teste e as demais para treino e validação. A média dos cinco desempenhos é reportada com intervalo de confiança de 95%.

RQ1: Qual é o impacto da aplicação de técnicas de reamostragem de dados para a tarefa de classificação em cenários com escassez de dados e em língua portuguesa? A Tabela 2 apresenta as médias de acurácia, tempo total gasto para obter a predição final – incluindo tempo de treinamento do modelo, tempo de inferência (predição), e tempo para geração de paráfrases, oversampling ou undersampling – e Macro-F1 — esta última atribuindo peso igual a todas as classes — para os métodos avaliados. Resultados indicados com '-' refletem a incompatibilidade de algumas técnicas com o BERTimbau, que não utiliza representações TF-IDF necessárias para a geração sintética de dados (e.g., *SMOTE*)[2]. Comparando-se os classificadores com dados originais, o BERTimbau supera a regressão logística em acurácia e macro-F1. A classificação sem reamostragem (Original) apresenta desempenho inferior em relação ao uso de *oversampling* ou *undersampling*, sendo o *oversampling* superior, possivelmente pela menor perda de informação.

Entre os métodos de *oversampling*, tanto a regressão logística quanto o BERTimbau apresentaram melhorias em relação ao uso exclusivo dos dados originais, com ganho de até 57% na macro-F1 (Original vs. *SMOTE*, coluna LR). No BERTimbau, a geração

[2]Embora seja teoricamente possível aplicar técnicas de *under*/*oversampling* sobre *embeddings* densos gerados pelo BERTimbau, tal abordagem exige adaptações significativas fora do escopo original dessas técnicas, que serão alvo de trabalhos futuros.

de paráfrases via LLM teve maior impacto, alcançando macro-F1 de 62%, um aumento de 16% (Original vs. *LLM Paraphrase Oversample*, coluna BERTimbau). Assim, em resposta à questão de pesquisa, técnicas de reamostragem — especialmente *oversampling* com *SMOTE* e paráfrases — demonstraram impacto positivo no desempenho. Um fator importante a se observar, contudo, é o custo computacional necessário para executar cada solução. Os maiores custos são relacionados à geração de paráfrases com o LLM e ao *fine-tuning* do modelo BERTimbau, como podemos observar na Tabela 2.

Tabela 2. Resultado de média de acurácia e macro-f1 para E-SUS APS dataset, seguido de intervalo de confiança com 95% e tempo total médio gasto para obter a predição final em segundos.

Método	Regressão Logística		BERTimbau		Tempo Total (segundos)	
	Acurácia	Macro-F1	Acurácia	Macro-F1	LR	BERTimbau
Original	55.7 ± 2.8	38.6 ± 2.3	60.6 ± 1.3	53.3 ± 2.8	1.7	264
Random Undersample	57.1 ± 4.5	56.7 ± 4.0	45.1 ± 3.6	41.0 ± 2.7	0.6	128
NearMiss Undersample	52.8 ± 4.6	51.7 ± 4.7	-	-	1.2	-
Random Oversample	63.2 ± 5.2	60.1 ± 4.4	61.4 ± 1.8	55.6 ± 3.9	2.1	545
Smote Oversample	63.9 ± 5.2	60.9 ± 5.7	-	-	3.3	-
SVMSmote Oversample	61.4 ± 4.4	56.1 ± 5.9	-	-	12.4	-
LLM Paraphrase Oversample	63.6 ± 4.7	60.6 ± 4.8	64.5 ± 5.1	62.0 ± 6.6	10362	10908
Original + Novos Dados	66.9 ± 2.8	64.1 ± 5.0	63.4 ± 2.5	61.1 ± 0.6	2	429
Original + Novos Dados + Random Oversample	66.9 ± 4.3	65.7 ± 4.6	64.3 ± 3.1	61.7 ± 1.3	3	648
Original + Novos Dados + Paraphrase	65.6 ± 4.6	64.3 ± 5.1	65.1 ± 4.0	63.1 ± 5.7	10363	11072

RQ2: A anotação adicional de dados por meio de *amostragem seletiva inversamente proporcional à distribuição de classes* pode mitigar o desbalanceamento? Esta questão avalia o impacto da inclusão de dados rotulados por especialistas. A Tabela 2 indica um ganho de 70% ao comparar Original + Novos Dados + Random Oversample" com "Original". Em relação ao melhor *oversampling* anterior (SMOTE), os novos dados proporcionam um acréscimo de 8% em *Macro-F1*. Para o BERTimbau, o ganho é menor (1.75%), mas há redução da variância (intervalo de confiança mais estreito). Os resultados evidenciam o benefício da anotação adicional, especialmente pela priorização de classes raras. Comparando RL (65.7) e BERTimbau (63.1), ambos são estatisticamente equivalentes (teste t), mas a RL se destaca pelo menor custo computacional — cerca de 3690x mais eficiente, conforme a Tabela 2 (3 segundos para RL versus 11072 para BERTimbau).

6. Conclusões e Trabalhos Futuros

Este trabalho avaliou o impacto de técnicas de reamostragem para mitigar desbalanceamento em dados textuais. Os resultados mostram que estratégias de *oversampling* — especialmente *Random*, *SMOTE* e paráfrases — superam *undersampling*, com ganhos significativos em *Macro-F1*. A combinação de paráfrases com dados anotados por especialistas aumentou o desempenho em 18% sobre o BERTimbau original e 70% sobre a Regressão Logística (LR) original. A melhor efetividade foi obtida com o BERTimbau a partir da junção de *oversampling* por paráfrases e rotulação seletiva, estatisticamente equivalente à RL com *random oversampling* e rotulação seletiva. Contudo, o LR é cerca de 3690x mais eficiente que o BERTimbau comparando ambas as versões mais efetivas de cada método. Como trabalho futuro, empregaremos geração sintética com LLMs para lidar com classes raras, além de *active learning* e testes em outros domínios, como na classificação de sentimento e tópicos de documentos.

Agradecimentos

Este trabalho foi apoiado por CNPq, Capes, Fapemig, Fapesp, AWS, NVIDIA, CIIA-Saúde, INCT-TILD-IAR (processo 408490/2024-1) e Secretaria de Atenção Primária do Ministério da Saúde (TED 31/2022).

Referências

Batista, G. E. A. P. A., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *SIGKDD Explor. Newsl.*, 6(1):20–29.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: synthetic minority oversampling technique. *J. Artif. Int. Res.*, 16(1):321–357.

Cunha, W., França, C., Fonseca, G., Rocha, L., & Gonçalves, M. A. (2023). An effective, efficient, and scalable confidence-based instance selection framework for transformer-based text classification. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 665–674, New York, NY, USA. Association for Computing Machinery.

Cunha, W., Moreo Fernández, A., Esuli, A., Sebastiani, F., Rocha, L., & Gonçalves, M. A. (2025). A noise-oriented and redundancy-aware instance selection framework. *ACM Trans. Inf. Syst.*, 43(2).

Han, H., Wang, W.-Y., & Mao, B.-H. (2005). Borderline-smote: A new over-sampling method in imbalanced data sets learning. In Huang, D.-S., Zhang, X.-P., & Huang, G.-B., editors, *Advances in Intelligent Computing*, pages 878–887, Berlin, Heidelberg. Springer Berlin Heidelberg.

He, H. & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284.

Hu, J., Ruder, S., Siddhant, A., Neubig, G., Firat, O., & Johnson, M. (2020). XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proc. of Machine Learning Research*, pages 4411–4421. PMLR.

Last, F., Douzas, G., & Bação, F. (2017). Oversampling for imbalanced learning based on k-means and SMOTE. *CoRR*, abs/1711.00837.

McClure, J., Shimmei, M., Matsuda, N., & Jiang, S. (2024). Leveraging prompts in llms to overcome imbalances in complex educational text data.

Nguyen, H. M., Cooper, E. W., & Kamei, K. (2011). Borderline over-sampling for imbalanced data classification. *Int. J. Knowl. Eng. Soft Data Paradigm.*, 3(1):4–21.

Souza, F., Nogueira, R., & Lotufo, R. (2020). BERTimbau: pretrained BERT models for Brazilian Portuguese. In *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*.

Tabar, V. R., Eskandari, F., Salimi, S., & Zareifard, H. (2018). Finding a set of candidate parents using dependency criterion for the k2 algorithm. *Pattern Recognition Letters*, 111:23–29.

Taskiran, S. F., Turkoglu, B., Kaya, E., & Asuroglu, T. (2025). A comprehensive evaluation of oversampling techniques for enhancing text classification performance. *Scientific Reports*, 15:21631.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30.

Yadav, V., Tang, Z., & Srinivasan, V. (2024). Pag-llm: Paraphrase and aggregate with large language models for minimizing intent classification errors. In *Proc. of the International ACM SIGIR Conference*, SIGIR '24, page 2569–2573.