Avaliando o Desempenho de Modelos Generativos de Dados para Classificação de Notícias Falsas

William Teles de Andrade Júnior[1], **João Gabriel Rocha Silva**[2], **Rodrigo Cesar Lira**[1], **Antônio Correia de Sá Barreto Neto**[1]

[1]Instituto Federal de Educação, Ciências e Tecnologia de Pernambuco (IFPE)
Campus Paulista – PE – Brasil

[2]Instituto Federal de Educação, Ciências e Tecnologia de Brasília (IFB)
Campus São Sebastião – DF – Brasil

wtaj@discente.ifpe.edu.br, joao.gabriel@ifb.edu.br,

{rodrigo.lira, antonio.neto}@paulista.ifpe.edu.br

Abstract. *This paper aimed to investigate the potential of models to generate synthetic data to improve fake news detection. The research compares the results obtained from a real dataset, containing news information, with those obtained from four synthetic datasets generated using GAN, VAE, DDPM and SMOTE. The study results indicate that classification performance improved when using artificial data, with an accuracy score of approximately 87%. These results suggest that synthetic data can be a valuable tool for improving fake news classification performance.*

Resumo. *Este artigo teve como objetivo investigar o potencial dos modelos generativos de dados sintéticos para a abordagem de detecção de notícias falsas. A pesquisa compara os resultados obtidos de um conjunto de dados real, contendo informações obtidas de notícias da internet, com aqueles obtidos de quatro conjuntos de dados sintéticos gerados usando GAN, VAE, DDPM e SMOTE. Os resultados do estudo indicam que o desempenho da classificação obteve uma melhora quando usado os dados sintéticos, com uma pontuação de acurácia de, aproximadamente, 87%. Esses resultados sugerem que dados sintéticos podem servir como ferramentas valiosas para melhorar o desempenho classificação de notícias falsas.*

1. Introdução

As notícias falsas (do inglês, *fake news*) são um problema significativo na era digital. Com o surgimento das mídias sociais e outras plataformas *online*, tornou-se fácil disseminar informações falsas, uma ação com potencial de causar danos a indivíduos, comunidades e até nações [Vosoughi et al. 2018]. Para combater esse problema, existe um interesse crescente no desenvolvimento de métodos e ferramentas para detectar e classificar notícias falsas, visando promover fontes de informação mais precisas e confiáveis [Zhou and Zafarani 2020].

Os métodos tradicionais de classificação baseiam-se em conjuntos de dados reais, que podem ser limitados em tamanho e podem não capturar toda a gama de artigos de notícias falsas. Conforme apontado por [Horne and Adali 2017], a falta de conjuntos de

dados confiáveis e abrangentes de notícias falsas é um grande obstáculo no desenvolvimento de algoritmos eficazes para detectar notícias falsas. Além disso, os conjuntos de dados de notícias falsas existentes sofrem geralmente de vieses e limitações, como serem limitados a um período ou região específica. Um caminho para superar esse problema é a utilização de conjuntos de dados sintéticos gerados por diferentes modelos como uma ferramenta para melhorar o tamanho do conjunto de dados real e o desempenho da classificação [Salimans et al. 2016].

Os dados sintéticos são conjuntos de dados criados artificialmente que imitam as características de dados reais, eles podem ser gerados usando técnicas como Redes Adversárias Generativas (GANs) [Frid-Adar et al. 2018], modelos generativos profundos [Suroso et al. 2023] e modelos probabilísticos de difusão de redução de ruído [Carrillo-Perez et al. 2023]. Ao gerar dados sintéticos que se assemelham a dados reais, esses métodos podem ajudar a superar algumas das limitações das abordagens tradicionais de Aprendizado de Máquina.

Neste estudo é explorado o uso de modelagem generativa de dados para criação de um conjunto de dados para detecção de notícias falsas, usando dados sintéticos criados a partir de um conjunto de dados real com informações de características das notícias obtidas da internet. O objetivo é avaliar a capacidade dos modelos gerar dados sintéticos para esta abordagem de detecção de notícias falsas. Para isso, são aplicadas técnicas de modelagem generativa para criar um conjunto de dados sintéticos que imita as propriedades estatísticas do conjunto de dados reais. Em seguida, o modelo de classificação é treinado no conjunto de dados originais e sintéticos e seus desempenhos são comparados com métricas de classificação.

Este trabalho está dividido da seguinte maneira: na Seção 2 é apresentada a fundamentação teórica do trabalho, onde é brevemente discutido a área de geração sintética de dados. Na Seção 3 é explicado como foi desenvolvida a metodologia utilizada para atingir os objetivos. Na Seção 4 são mostrados os resultados obtidos. Por fim, a Seção 5 trata sobre a conclusão e trabalhos futuros.

2. Geração de dados sintéticos

A geração de dados sintéticos é um processo de criação de novos conjuntos de dados que imitam as propriedades estatísticas dos dados do mundo real [Assefa et al. 2020]. Uma das abordagens mais comuns para geração de dados sintéticos é por meio do uso de técnicas de modelagem generativa, como redes adversárias generativas (GANs) [Goodfellow et al. 2020], codificadores automáticos variacionais [Kingma and Welling 2022], modelos probabilísticos de difusão de redução de ruído (DDPMs) [Nichol and Dhariwal 2021] e técnica de sobre-amostragem minoritária sintética (SMOTE) [Mukherjee and Khushi 2021].

Embora a geração de dados sintéticos tenha muitas vantagens, como reduzir a necessidade de abundantes dados rotulados, ela também apresenta algumas limitações. Sendo uma delas que os dados sintéticos podem não capturar totalmente a complexidade e a variabilidade dos dados reais, levando a vieses e imprecisões em modelos de Aprendizado de Máquina treinados com eles [Paszke et al. 2019]. Outra limitação é que os dados

sintéticos podem não ser representativos da população em estudo, levando a uma fraca capacidade de generalização de modelos de Aprendizado de Máquina treinados com esses mesmos dados [Lu et al. 2023].

3. Metodologia

Visando alcançar o objetivo deste estudo, os dados sintéticos foram gerados a partir de um conjunto de dados reais composto por 7 critérios (*features*) e 100 notícias no total (instâncias), sendo 50 falsas e 50 verdadeiras utilizados por Ferreira et al. [Ferreira et al. 2020]. A base gerada artificialmente possui novas instâncias sintéricas que serão analisadas pelo modelo de classificação proposto pelo mesmo trabalho.

Foram utilizados os seguintes modelos generativos para geração de dados: Redes adversárias generativas (GAN); Codificadores Automáticos Variacionais (VAE); modelos probabilísticos de difusão de redução de ruído (DDPM) e técnica de sobre-amostragem minoritária sintética (SMOTE), cada modelo gerou 100 novas instâncias a partir do conjunto de dados real. Os modelos generativos foram selecionados para este estudo por possuírem estratégias fundamentalmente distintas no processo de geração dos dados, além de apresentarem bons resultados em outros estudos [Kotelnikov et al. 2023].

Para o modelo GAN, a rede do gerador consiste em várias camadas densamente conectadas com funções de ativação Leaky ReLU, normalização de lote e uma camada final com ativação sigmoide. A entrada para o gerador foi um vetor de ruído aleatório de dimensionalidade 50. A rede discriminadora foi construída com camadas densamente conectadas, utilizando ativações *Leaky ReLU*, e concluída com um único neurônio com função de ativação sigmoide para classificação binária (falso ou verdadeiro). O modelo GAN foi treinado de forma adversária com otimizador *Adam* com taxa de aprendizado de 0,002 e o treinamento foi realizado por um total de 1000 épocas. Em cada iteração de treinamento, um lote de amostras de dados reais e um lote de dados sintéticos de tamanho igual gerado pelo gerador atual foram combinados, o discriminador foi treinado nesse lote combinado com rótulos apropriados. A perda foi calculada usando entropia cruzada binária, o gerador foi então treinado para minimizar a capacidade do discriminador de distinguir entre dados reais e sintéticos.

Para o modelo VAE, dado o tamanho limitado do conjunto de dados, a dimensionalidade do espaço latente foi definida como 3. O modelo VAE foi treinado por um total de 1000 épocas, usando uma função de perda personalizada que combina perda de reconstrução (entropia cruzada binária) e um termo relacionado à divergência de *Kullback-Leibler* [Seghouane and Amari 2007], essa função de perda encorajou o espaço latente aprendido a se aproximar de uma distribuição gaussiana unitária. O modelo também foi treinado usando retro propagação e gradiente descendente estocástico. Os hiper-parâmetros como taxa de aprendizagem de 0,001, tamanho de lote de 64 e a perda de reconstrução de 0,5 foram ajustados para otimizar o desempenho.

No DDPM o codificador consistiu em uma série de camadas totalmente conectadas com ativações de unidades lineares retificadas (*ReLU*). Já o decodificador foi composto por camadas totalmente conectadas com ativações ReLU, responsáveis por reconstruir os dados do espaço latente. O DDPM durante seu processo de treinamento envolveu

a conversão do conjunto de dados reais para uma matriz multidimensional contendo elementos de um único tipo de dados, e também envolveu a minimização da perda do erro quadrático médio (MSE) entre os dados reconstruídos e a entrada original. O otimizador Adam foi utilizado com taxa de aprendizado de 0,001 e o treinamento foi realizado por um total de 1000 épocas. Durante cada época, o modelo foi treinado em mini lotes, de tamanho 8, para melhorar a eficiência computacional.

A implementação do SMOTE teve seu hiper-parâmetro de quantidade de vizinhos definida como 5 sendo implementada utilizando a biblioteca *imbalanced-learn*. Nele, a sobre-amostragem foi feita em interações, desbalanceando a base original até alcançar 100 elementos sintéticos, 50 verdadeiros e 50 falsos.

O processo de geração de dados foi realizado duas vezes com o subconjunto de cada classe no DDPM, GAN e VAE. Para a técnica de geração de dados SMOTE, o processo foi realizado iterativamente desbalanceando as classes do conjunto de dados reais até que o nível desejado de sobre-amostragem fosse alcançado.

Para fazer a classificação de uma notícia como verdadeira ou falsa e verificar o desempenho dos dados artificiais criados, foi utilizado o método de classificação proposto por [Ferreira et al. 2020], no qual, a avaliação de uma notícia é realizada por meio da multiplicação do peso de cada critério (P_i) pelo valor deste critério estrutural na notícia (C_i). Quando o valor da avaliação (A_i) é maior que $0,6$ (60%) a notícia é classificada pelo modelo como verdadeira e caso contrario, falsa.

Para encontrar os pesos ideais que ponderam cada critério do modelo matemático, foi realizada uma etapa de busca utilizando um algoritmo genético (GA), como proposto por [Almeida et al. 2021]. No processo de treinamento para os dados reais foi necessário dividi-los pela metade, sendo 50 notícias para treino e 50 notícias de teste, ambos com 25 notícias verdadeiras e 25 notícias falsas. O algoritmo GA foi executado 30 vezes para cada conjunto de dados sintético gerado. Os parâmetros de operadores genéticos do GA utilizados foram 100 gerações, 100 indivíduos, 5% de taxa de mutação e 50% de taxa de cruzamento.

A metodologia utilizada neste trabalho pode ser dividido em três etapas, conforme demostrado na Figura 1. Na Etapa 1 é realizada a geração dos dados sintéticos, onde o conjunto de dados real é usado em um método de geração de dados (GAN, VAE, DDDPM ou SMOTE) e esse método cria dois conjuntos de dados separados. Um conjunto de dados possui as notícias da classe falsa e a outra a classe verdadeira. Esses conjuntos de dados são agrupados para formar o conjunto de dados sintético final. Na Etapa 2 é iniciado o processo de classificação, onde os dados sintéticos gerados são utilizados no algoritmo genético, para obter os coeficientes que ponderam os critérios que serão utilizados no modelo matemático, como desenvolvido em trabalhos anteriores [Ferreira et al. 2020]. E por fim, na última etapa (Etapa 3), são calculadas as métricas de avaliação da classificação que neste trabalho foram acurácia e F1.

Nesta pesquisa os experimentos foram realizados utilizando a plataforma do Google Colab com Python 3 e CPU como acelerador de hardware e também em um notebook com CPU AMD Ryzen™ 7 2700U a 2,20 GHz, com GPU AMD Radeon™ 540 com 2 GB de memória e memória RAM de 12 GB no sistema operacional Windows 10. Todos os códigos-fontes das simulações foram desenvolvidos *Python*, com as bibliotecas

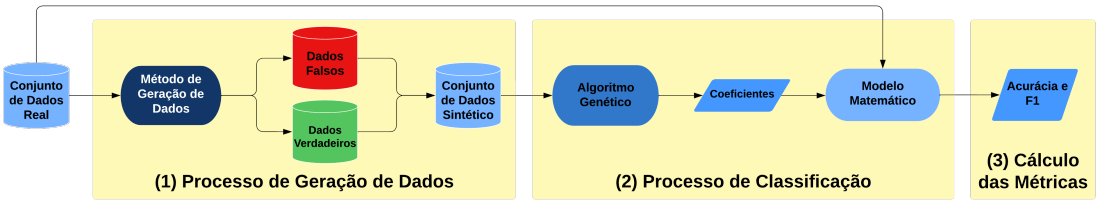

Figura 1. Fluxograma do processo de geração de dados sintéticos e avaliação de desempenho.

tensorflow, keras, pytorch e imbalanced-learn, scikit-learn, geneticalgorithm, pandas e *numpy*.

4. Resultados

Na Tabela 1 são apresentados os resultados de avaliação de desempenho do modelo de classificação de notícias falsas, com os quatro diferentes conjuntos de dados sintéticos juntamente com o custo de tempo para gerar cada conjunto de dados. Além dessas informações, também são apresentados os resultados utilizando o base de dados real - base de dados criado a partir de informações reais colhidas da internet, que servirá como um valor de referência (*baseline*) dos resultados obtidos com os dados sintéticos.

Tabela 1. Resultados da avaliação de desempenho da classificação.

Conjunto de Dados	Acurácia *Média* (σ)	F1 *Média* (σ)	Custo Computacional *Média* (σ)
Real (*baseline*)	$0,8020$ $(0,0059)$	$0,8347$ $(0,0042)$	Não Aplicável
GAN	$0,8736$ $(0,0048)$	$0,8878$ $(0,0038)$	$150s$ $(4,16s)$
VAE	$0,8806$ $(0,0057)$	$0,8934$ $(0,0046)$	$33,8s$ $(6,28s)$
DDPM	$0,8543$ $(0,0513)$	$0,8746$ $(0,0396)$	$1002s$ $(128s)$
SMOTE	$0,8786$ $(0,0033)$	$0,8918$ $(0,0026)$	$2528\text{e}{-6}\text{s}$ $(552\text{e}{-6}\text{s})$

Os resultados do modelo de classificação no conjunto de dados real alcançaram 80% de acurácia. A pontuação F1 para esse o conjunto de dados também foi calculada e considerada consistente com a medida de acurácia. Em contraste, os conjuntos de dados sintéticos produziram melhores resultados. A acurácia média nos quatro conjuntos de dados sintéticos foi de aproximadamente 87%. Isso indica que o modelo alcançou um maior nível classificação de artigos de notícias falsas nos conjuntos de dados sintéticos em comparação com o conjunto de dados real.

A pontuação F1 para os conjuntos de dados sintéticos também foram calculados e considerados consistentemente superiores aos obtidos no *baseline*. É importante destacar que o modelo utilizando apenas o conjunto de dados real foi utilizado e avaliado apenas para efeito de referência, pois o seu cenário de treinamento e teste difere dos modelos utilizando os dados sintéticos.

Os resultados deste estudo sugerem que o desempenho do modelo de classificação de notícias falsas foi melhor nos conjuntos de dados sintéticos em comparação com o con-

junto de dados real, isso pode ser dar devido à base de treinamento dos dados sintéticos possuir mais instâncias. Esta descoberta é consistente com pesquisas anteriores que mostraram a eficácia dos conjuntos de dados sintéticos na melhoria do desempenho dos modelos de detecção de notícias falsas [Shu et al. 2017, Wang 2017, Horne and Adali 2017].

Uma possível explicação para esta observação é que os conjuntos de dados sintéticos foram gerados especificamente para imitar as características dos dados originais. Esses conjuntos de dados podem conter exemplos mais diversos e representativos de notícias falsas e verdadeiras, permitindo que o modelo de classificação desempenhasse de forma mais eficaz e alcançasse maior precisão. Além disso, os conjuntos de dados sintéticos podem ter sido menos tendenciosos ou ruidosos em comparação com o conjunto de dados real, levando a um melhor desempenho.

Outro ponto que deve ser notado é o custo computacional de cada método de geração de dados sintéticos. O SMOTE, por ser o mais simples em questão de complexidade, se mostrou ter um tempo muito menor em relação aos outros, enquanto o DDPM, por ser mais complexo e devido a sua lenta velocidade de inferência, foi o que mais demorou para ter seu processo finalizado. É importante mencionar que mesmo o SMOTE sendo o mais rápido, ele possui uma complexidade maior de ser utilizado quando a base de dados não está desbalanceada, tornando-se necessário o desbalanceamento iterativo do conjunto de dados para realizar a sobre-amostragem na quantidade desejada.

Usando uma taxa de confiança de 99%, foi aplicado o teste de *Wilcoxon* para comparar os resultados da métrica F1 entre os algoritmos de geração de dados sintéticos. Conforme mostrado na Tabela 2. O símbolo '-' indica não haver diferença estatística entre as soluções, o símbolo '▲' indica que a abordagem obteve resultados melhores que o outro algoritmo e o símbolo '▽' representa que a proposta alcançou resultados piores que o método comparado.

Tabela 2. Resultado do Teste Wilcoxon com 99,9% de confiança.

Modelo de Geração	GAN	VAE	DDPM	SMOTE
GAN	-	-	▲	▽
VAE	-	-	▲	▽
DDPM	▽	▽	-	▽
SMOTE	▲	▲	▲	-

Observa-se que os resultados alcançados pelo SMOTE superam, em geral, o GAN, VAE e DDPM. SMOTE pode não ser o mais eficiente algoritmo para todos os cenários, mas por seus resultados da pontuação F1 durante a classificação terem um desvio padrão menor que os outros, conseguiu ter um resultado estatístico melhor que as outras abordagens. Isso pode ser devido ao fato que os dados SMOTE devem ter tido uma distribuição mais parecida com as dos dados reais, pela forma que o SMOTE cria os dados sintéticos, indo em busca de vizinhos no espaço e também devido à baixa complexidade e quantidade de dados reais usados. Outra possibilidade se dá pelo fato de que os outros métodos utilizados poderiam ser melhor parametrizados, visto que em comparação ao SMOTE, os outros possuem um conjunto maior de hiper-parâmetros e configurações que não foram investigados profundamente.

5. Conclusões e Trabalhos Futuros

Este trabalho comparou quatro diferentes métodos de geração de dados sintéticos para um problema de classificação de notícias falsas. A base de dados real foi desenvolvida a partir de um conjunto de notícias reais obtidas na web. As ferramentas utilizadas para geração desses dados foi a linguagem de programação *Python*.

Os resultados indicaram que a classificação foi superior com os conjuntos de dados sintéticos. Esta descoberta sugere que os conjuntos de dados sintéticos podem ser eficazes na melhoria do desempenho dos modelos de detecção de notícias falsas. No entanto, são necessárias mais estudos para compreender os fatores subjacentes que contribuem para esta diferença de desempenho e para explorar outras abordagens potenciais para melhorar a precisão dos modelos de classificação de notícias falsas.

A técnica SMOTE demonstrou uma vantagem em relação aos outros métodos ao ser avaliada pelo teste estatístico Wilcoxon na pontuação F1. Esse resultado pode ser atribuído ao fato de o algoritmo SMOTE ter conseguido criar um conjunto de dados com características mais próximas às do conjunto de dados reais. Dessa forma, ao encontrar os coeficientes com o algoritmo genético e aplicá-los ao modelo matemático, o algoritmo genético variou menos que os outros nos coeficientes, resultando numa pontuação F1 com um desvio padrão menor em comparação com os testes dos outros métodos.

Nos trabalhos futuros, espera-se analisar a parametrização dos modelos já usados como também outros métodos de criação de dados tabulares sintéticos desenvolvidos, bem como classificar os dados com diferentes técnicas e algoritmos de Aprendizagem de Máquinas. Também é esperado analisar a existência de um viés na geração dos dados sintéticos com a mesma base de validação.

Referências

Almeida, A. L., Carrara, G., Prates, I., Nascimento, L. C., Souza, P. H., Almeida, T., Cani, R., and Silva, J. G. (2021). Modelo matemático apoiado por um algoritmo genético para classificação de fake news na web. In *Anais do VIII Encontro Nacional de Computação dos Institutos Federais*, pages 17–20, Porto Alegre, RS, Brasil. SBC.

Assefa, S. A., Dervovic, D., Mahfouz, M., Tillman, R. E., Reddy, P., and Veloso, M. (2020). Generating synthetic data in finance: opportunities, challenges and pitfalls. In *Proceedings of the First ACM International Conference on AI in Finance*, pages 1–8.

Carrillo-Perez, F., Pizurica, M., Zheng, Y., Nandi, T. N., Madduri, R., Shen, J., and Gevaert, O. (2023). Rna-to-image multi-cancer synthesis using cascaded diffusion models. *bioRxiv*.

Ferreira, A. L. N., Nascimento, D. G., Basílio, S. C. A., and Silva, J. G. R. (2020). Um modelo matemático para classificação de fake news na web. In *Anais do Simpósio Brasileiro de Pesquisa Operacional*.

Frid-Adar, M., Diamant, I., Klang, E., Amitai, M., Goldberger, J., and Greenspan, H. (2018). Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification. *Neurocomputing*, 321:321–331.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2020). Generative adversarial networks. *Commun. ACM*, 63(11):139–144.

Horne, B. and Adali, S. (2017). This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1):759–766.

Kingma, D. P. and Welling, M. (2022). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kotelnikov, A., Baranchuk, D., Rubachev, I., and Babenko, A. (2023). Tabddpm: Modelling tabular data with diffusion models. In *International Conference on Machine Learning*, pages 17564–17579. PMLR.

Lu, Y., Wang, H., and Wei, W. (2023). Machine learning for synthetic data generation: a review. *arXiv preprint arXiv:2302.04062*.

Mukherjee, M. and Khushi, M. (2021). Smote-enc: A novel smote-based method to generate synthetic data for nominal and continuous features. *Applied System Innovation*, 4(1):18.

Nichol, A. Q. and Dhariwal, P. (2021). Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X., and Chen, X. (2016). Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Seghouane, A.-K. and Amari, S.-I. (2007). The aic criterion and symmetrizing the kullback–leibler divergence. *IEEE Transactions on Neural Networks*, 18(1):97–106.

Shu, K., Sliva, A., Wang, S., Tang, J., and Liu, H. (2017). Fake news detection on social media: A data mining perspective. *SIGKDD Explor. Newsl.*, 19(1):22–36.

Suroso, D., Cherntanomwong, P., and Sooraksa, P. (2023). Synthesis of a small fingerprint database through a deep generative model for indoor localisation. *Elektronika Ir Elektrotechnika*, 29:69–75.

Vosoughi, S., Roy, D., and Aral, S. (2018). The spread of true and false news online. *Science*, 359(6380):1146–1151.

Wang, W. Y. (2017). "liar, liar pants on fire": A new benchmark dataset for fake news detection. In Barzilay, R. and Kan, M.-Y., editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada. Association for Computational Linguistics.

Zhou, X. and Zafarani, R. (2020). A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Comput. Surv.*, 53(5).